A new link-based method to ensemble clustering and cancer microarray data analysis

Natthakan Iam-On*

School of Information Technology,
Mae Fah Luang University,
333 Moo1, Tasud, Muang Chiang Rai 57100, Thailand
E-mail: nt.iamon@gmail.com
*Corresponding author

Tossapon Boongoen and
Nattawut Kongkotchawan

Department of Mathematics and Computer Science,
Royal Thai Air Force Academy,
171/1 Klongtanhon, Saimai, Bangkok 10220, Thailand
E-mail: tossapon b@rtaf.mi.th
E-mail: nattawut.kongkotchawan@gmail.com

Abstract: Ensemble clustering or cluster ensembles have been shown to be better than any standard clustering algorithm at improving accuracy. This meta-learning formalism helps users to overcome the dilemma of selecting an appropriate technique and the parameters for that technique, given a set of data. It has proven effective for many problem domains, especially microarray data analysis. Among different state-of-the-art methods, the link-based approach (LCE) recently introduced by Iam-On et al. (2011) provides a highly accurate clustering. This paper presents the improvement of LCE with a new link-based metric being developed and engaged. Additional information that is already available in a network is included in the similarity assessment. As such, this refinement can increase the quality of the measures, hence the resulting cluster decision. The performance of this improved LCE is evaluated on synthetic and UCI benchmark datasets, in comparison with the original and several well-known cluster ensemble techniques. In addition, the application of improved LCE to microarray data analysis is also empirically assessed. The findings suggest that the new model can improve the accuracy of LCE and performs better than the others investigated in this study.

Keywords: data clustering; ensemble; link-based similarity; microarray.

Reference to this paper should be made as follows: Iam-On, N., Boongoen, T. and Kongkotchawan, N. (2014) ‘A new link-based method to ensemble clustering and cancer microarray data analysis’, *Int. J. Collaborative Intelligence*, Vol. 1, No. 1, pp.45–67.

Biographical notes: Natthakan Iam-On received her PhD in Computer Science from Aberystwyth University, UK. Currently, she is a Lecturer at School of Information Technology, Mae Fah Luang University, Thailand. Her research focuses on data clustering, cluster ensembles and applications to biomedical data analysis.

Tossapon Boongoen is a Lecturer with the Department of Mathematics and Computer Science, Royal Thai Air Force Academy, Thailand. Prior to this appointment, he obtained his PhD in Artificial Intelligence from Cranfield University and worked as a Postdoctoral Research Associate at Aberystwyth University. His research interests include data mining, link analysis, data clustering, fuzzy aggregation and classification system.

Nattawut Kongkotchawan received his BSc in Computer Science from Royal Thai Air Force Academy. He is presently a Research Assistant with the Department of Mathematics and Computer Science, Royal Thai Air Force Academy, Thailand. His research is in line with intelligent system, social network and big data analysis.

1 Introduction

Cluster analysis is usually employed in the initial stage of understanding a raw data, especially for new problems where prior knowledge is minimal. Also, in the pre-processing stage of supervised learning, it is exploited to identify outliers and possible object classes for the following expert-directed labelling process. This is crucial when the complexity of modern-age information is generally overwhelming for a human investigation. The need to acquire knowledge or learn from the excessive amount of data is hence a major driving force for making clustering a highly active research subject. Data clustering is applied to a variety of problem domains such as biology (Jiang et al., 2004), customer relationship management (Wu et al., 2005), information retrieval (Bhatia and Deogun, 1998; Zhang et al., 2002), image processing (Costa and de Andrade Netto, 1999), and recommender systems (Kim and Ahn, 2008). In addition, the recent development of clustering cancer gene expression data has attracted a lot of interests amongst computer scientists, biological and clinical researchers (Iam-On et al., 2010; Kim et al., 2009). Given its potential, a large number of research studies focus on several aspects of cluster analysis: for instance, clustering algorithms and extensions for particular data type (Ahmad and Dey, 2007), relevance of data attributes per cluster or subspace clustering (Boongoen et al., 2011), evaluation of clustering results (Rand, 1971), and cluster ensembles (Iam-On et al., 2010).

Specific to cluster ensembles, this practice is motivated by the fact that the performance of most clustering techniques are highly data dependent. A clustering model may produce an acceptable result for one dataset, but possibly become ineffective for others. Generally, there are two major challenges 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 optimise a specific criterion. Second, a single clustering algorithm with different parameter settings can also reveal various structures on the same dataset. One setting may be good for a few, but not all datasets. These consequently make the selection of a proper clustering technique very difficult.

In order to resolve this problem, researchers attempt to combine 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; Topchy et al., 2005). Over the past decade, many techniques have been developed along this line of research. They can be categorised into direct approach (Fischer and Buhmann, 2003; Gionis et al., 2007), feature-based approach (Boulis and Ostendorf, 2004; Nguyen and Caruana, 2007; Topchy et al., 2005), pairwise-similarity approach (Ayad and Kamel, 2003; Fred and Jain, 2005; Monti et al., 2003), and graph-based algorithms (Fern and Brodley, 2004; Strehl and Ghosh, 2002). 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 meta-level information matrix) which is simply created by stacking up ensemble members' decisions. The relations between these decisions (or data partitions) have been unfortunately overlooked.

Inspired by such an observation, a link-based method (LCE) is introduced by Iam-On et al. (2010, 2011) to address and use those associations to their true potential. It models base clustering results as a link network from which the relations between and within these decisions can be systematically obtained. This is accomplished through the link-based similarity measure called 'weighted connected triple (WCT)'. Disclosed relations are then exploited to refine the conventional meta-level matrix that has been the centre of several benchmark techniques. It is reported that the resulting technique performs consistently better than several state-of-the-art alternatives on both UCI benchmark and gene expression datasets. These findings have encouraged the recent improvement of LCE, which is presented in this paper. A new link-based similarity measure, weighted triple uniqueness (WTU), is brought into the underlying similarity assessment. Whilst being as efficient as WCT, WTU makes use of more information already available in a network to estimating a similarity measure. As such, the quality of information matrix, hence the final clustering, can be improved.

The rest of this paper is organised as follows. Section 2 introduces the basis of cluster ensembles, including formal definition, framework, ensemble generation strategies, and consensus functions. Then, the improved link-based approach and the underlying similarity measure is presented in Section 3. Based on benchmark datasets, Section 4 includes the evaluation of the proposed model as compared to the original LCE and other well-known methods. In addition, the application of the improved LCE method to cancer microarray data is reported in Section 5. This paper is concluded in Section 6 with the perspective of future research.

2 Basis of cluster ensembles

The aim of cluster ensembles is to combine different decisions of various clusterings such that the resulting accuracy superior to those of individual clusterings is obtained and robust across different datasets. Studies on developing cluster ensemble methods have shown that cluster ensembles achieve the benefits beyond what a standard algorithm 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 microarray 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 later, this section presents the basic concepts of cluster ensembles.

2.1 Problem formulation

Let $X = \{x_{1,} \ldots, x_N\}$ be a set of N data points, where each $x_i \in X$ is represented by a vector of D attribute values, i.e., $x_i = (x_{i,1}, \ldots, x_{i,D})$. Also, let $\Pi = \{\pi_1, \ldots, \pi_M\}$ 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 $\pi_g = \{C_1^g, C_2^g, ..., C_{k_g}^g\}$, 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 $\pi^* = C_1^*, ..., C_K^*$, where K denotes the number of clusters in the final clustering result, of a data set X that summarises the information from the cluster ensemble Π. The general framework of cluster ensembles is shown in Figure 1. In essence, solutions achieved from different base clusterings are aggregated to form a final partition. This meta-level method involves two major tasks of:

1 generating a cluster ensemble

2 producing the final partition (normally referred to as 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, several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, hence the diversity within a cluster ensemble.

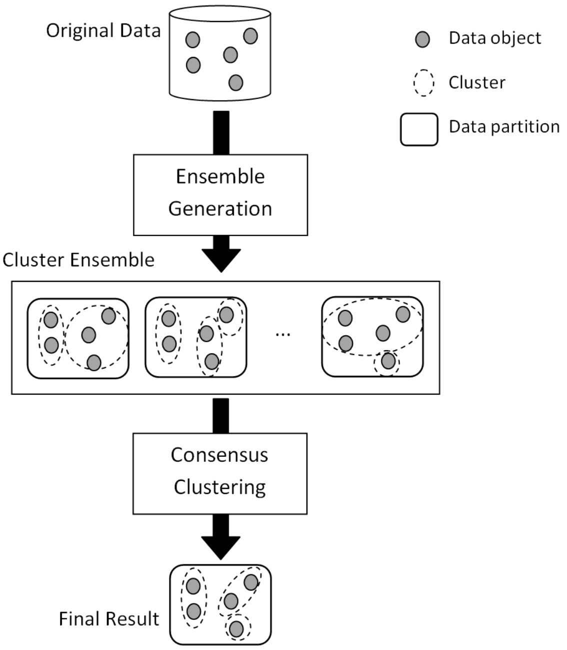
- Homogeneous ensembles: Base clusterings are created using repeated runs of a single clustering algorithm, with several sets of input parameters. For instance, *k*-means has often been employed with a random initialisation of cluster centres (Fred and Jain, 2005; Gionis et al., 2007; Iam-On et al., 2008).

- 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 *k* or randomly selected *k* from a pre-specified interval.

- Random subspacing/sampling: An ensemble can also be achieved by applying manifold subsets of initial data to base clusterings. In practice, theses can be obtained by projecting data onto different subspaces (Fern and Brodley, 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).

- 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).
- Mixed strategy: In addition to the aforementioned strategies, any combination of them can be applied as well. An example can be found in the study of Strehl and Ghosh (2002), where several clustering algorithms are used with multiple subspaces of data.

Figure 1 The basic process of cluster ensembles



Notes: It first applies multiple base clusterings to a dataset *X* to obtain diverse clustering decisions ($\pi_1, \ldots, \pi_M$). Then, these solutions are combined to establish the final clustering result (π^*) using a consensus function.

2.3 Consensus functions

Having obtained an ensemble, a variety of consensus functions have been developed to create the ultimate data partition. Each consensus function utilises a specific form of information matrix, which summarises the base clustering results. From the cluster ensemble shown in Figure 2(a), three general types of such ensemble-information matrix can be constructed. Firstly, the 'label-assignment' matrix [e.g., Figure 2(b)], of size $N \times M$, represents cluster labels that are assigned to each data point by different base clusterings. Secondly, the 'pairwise similarity' matrix [e.g., Figure 2(c)], of size $N \times N$, summarises co-occurrence statistics amongst data points. Furthermore, the 'binary cluster-association (BA)' matrix [e.g., Figure 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 categorisations.

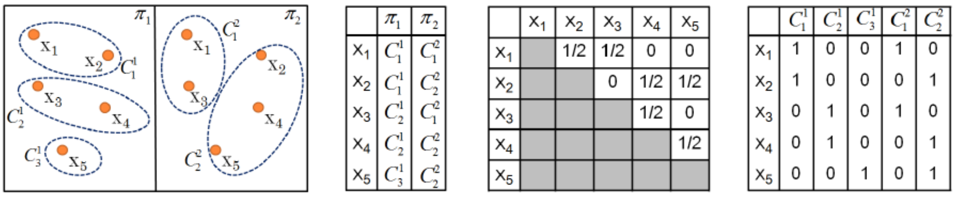
- *Feature-based approach*: It transforms the problem of cluster ensembles to the clustering of categorical data. Each base clustering provides a cluster label as a new feature describing each data point [Figure 2(b)], which is utilised to formulate the final solution (Boulis and Ostendorf, 2004; Nguyen 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., 2004, 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 Figure 2(b)].

- *Pairwise similarity approach*: It creates a matrix, containing the pairwise similarity among data points [see Figure 2(c)], to which any similarity-based clustering algorithm (e.g., hierarchical clustering) can be applied (Fred and Jain, 2005).

- *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 Figure 2(c). To achieve the final clustering result, this graph is divided into a definite number of approximately equal-sized partitions, using METIS (Karypis and Kumar, 1998). In addition, the BA matrix shown in Figure 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).

Figure 2 Examples of (a) cluster ensemble, (b) label-assignment matrix, (c) pairwise similarity matrix and (d) BA matrix (see online version for colours)



(a)

	π_1	π_2
x_1	C_1^1	C_1^2
x_2	C_1^1	C_2^2
x_3	C_2^1	C_1^2
x_4	C_2^1	C_2^2
x_5	C_3^1	C_2^2

(b)

	x_1	x_2	x_3	x_4	x_5
x_1	▨	1/2	1/2	0	0
x_2	▨	▨	0	1/2	1/2
x_3	▨	▨	▨	1/2	0
x_4	▨	▨	▨	▨	1/2
x_5	▨	▨	▨	▨	▨

(c)

	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	1	0	0	1	0
x_2	1	0	0	0	1
x_3	0	1	0	1	0
x_4	0	1	0	0	1
x_5	0	0	1	0	1

(d)

Notes: $X = \{x_1, \dots, x_5\}$, $\Pi = \{\pi_1, \pi_2\}$, $\pi_1 = \{C_1^1, C_2^1, C_3^1\}$ and $\pi_2 = \{C_1^2, C_2^2\}$.

3 New link-based method to ensemble clustering

The improved framework of link-based cluster ensembles (LCE) consists of three major steps:

1. creating a cluster ensemble Π
2. aggregating base clustering results, $\pi_g \in \Pi$, $g = 1, \ldots, M$, into a meta-level data matrix Θ
3. generating the final data partition π^* using the SPEC algorithm.

3.1 Creating cluster ensembles

The proposed approach is generic such that it can be coupled with several different ensemble generation methods. As for the present study, the following two types of cluster ensembles are investigated. Following the original work (Iam-On et al., 2010), the classical *k*-means is used to form base clusterings, each of which is initialised with a random set of cluster prototypes.

- *Fixed-k*: Each clustering $\pi_g \in \Pi$, is created using the data set $X \in \mathbb{R}^{N \times D}$ with all D attributes. The number of clusters in each base clustering is fixed to $k = \lceil \sqrt{N} \rceil$. Intuitively, to obtain a meaningful partition, k becomes 50 if $\lceil \sqrt{N} \rceil > 50$.
- *Random-k*: Each π_g is created using the data set with all attributes, and the number of clusters is randomly selected between $\{2, \ldots, \lceil \sqrt{N} \rceil\}$. Note that both 'fixed-k' and 'random-k' generation strategies are initially introduced in the primary work (Iam-On et al., 2008).

3.2 Aggregating base clustering results

Having obtained the cluster ensemble Π, the corresponding base clustering results are aggregated into an information matrix $\Theta \in [0, 1]^{N \times P}$, from which the final data partition π^* is generated. Note that P denotes the total number clusters in the ensemble under examination. For each clustering $\pi_g \in \Pi$ and their corresponding clusters $C_1^g, \ldots, C_{k_g}^g$, a matrix entry $\Theta(x_i, cl)$ which represents the association degree that the sample $x_i \in X$ has with each cluster $cl \in \{C_1^g, \ldots, C_{k_g}^g\}$, is estimated as follows:

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

where $C_*^g(x_i)$ is a cluster label to which sample 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 algorithm presented next.

WTU algorithm: has been developed to evaluate the similarity between any pair of clusters C_x, $C_y \in \Pi$. Note that WTU is based on the uniqueness measure developed as part of the algorithm called ‘connected-path’, which has been introduced by Boongoen et al. (2010) for the task of alias detection. Given a graph $G(V, E)$ in which objects and their relations are represented with members of the sets of vertices V and edges E, respectively, a uniqueness measure UQ_{ij}^{k} of any two objects i and j (denoted by vertices v_i, $v_j \in V$) can be approximated from each joint neighbour k (denoted by the vertex $v_k \in V$) as follows:

$$UQ_{ij}^{k} = \frac{f_{ik} + f_{jk}}{\sum_{m} f_{mk}} \tag{2}$$

where f_{ik} is the frequency of the link between objects i and k occurring in data, f_{jk} is the frequency of the link between objects j and k, and f_{mk} is the frequency of the link between object k and any object m.

WTU is considered as an extension to the WCT initially proposed with the LCE model. Whilst maintaining the efficiency, it makes use of additional information that is already available within a network. As such, the quality of similarity measure derived by WTU can be higher then that generated by WCT. At the outset of WTU evaluation, the ensemble Π is represented as a weighted graph $G = (V, W)$, where V is the set of vertices each representing a cluster in Π and W is a set of weighted edges between clusters. The weight $|w_{xy}| \in [0, 1]$ assigned to the edge $w_{xy} \in W$ between C_x, $C_y \in V$, is estimated as

$$\left|w_{xy}\right| = \frac{\left|L_x \cap L_y\right|}{\left|L_x \cup L_y\right|}, \tag{3}$$

where $L_z \subset X$ denotes the set of samples belonging to cluster $C_z \in \Pi$. Note that G is an undirected graph such that $|w_{xy}|$ is equivalent to $|w_{yx}|$, $\forall C_x, C_y \in V$. The WTU algorithm is summarised below.

```
ALGORITHM: WTU($G$, $C_x$, $C_y$)
$G = (V, W)$, a weighted graph, where $C_x, C_y \in V$;
$N_k \subset V$, a set of adjacent neighbors of $C_k \in V$; $C_z \in N_k$
    when $|w_{kz}| > 0$;
$WTU_{xy}$, the WTU measure of $C_x$ and $C_y$;
    $WTU_{xy} \leftarrow 0$
    For each $c \leftarrow N_x$
        If $c \leftarrow \mathrm{N}y$
            $WTU_{xy} \leftarrow WTU_{xy} + \frac{|w_{xc}| + |w_{yc}|}{\sum_{\forall C_t \in \Pi} |w_{tc}|}$
    Return $WTU_{xy}$
```

It is noteworthy that the size of neighbours N_k is not a user-defined parameter. This can be different from one cluster to another. Following the estimation of WTU measure, the similarity between clusters C_x and C_y can be estimated by

$$sim\left(C_x, C_y\right) = \frac{WTU_{xy}}{WTU_{\max}} \times DC, \tag{4}$$

where $WTU_{\max}$ is the maximum $WTU_{x'y'}$ value of any two clusters $C_{x'}$, $C_{y'} \in V$ and $DC \in [0, 1]$ is a constant decay factor (DC) (i.e., confidence level of accepting two non-identical clusters as being similar). With this link-based similarity metric, $sim(C_x, C_y) \in [0, 1]$ with $sim(C_x, C_x) = 1$, $C_x, C_y \in V$. It is also reflexive such that $sim(C_x, C_y) = sim(C_y, C_x)$.

3.3 Generating final data partition

Once the matrix Θ is created, the SPEC algorithm (Ng et al., 2001) is used to generate the final data partition. This technique was first introduced by Fern and Brodley (2004) as part of the hybrid bipartite graph formation (HBGF) framework. In particular, SPEC is exploited to divide a bipartite graph, which is transformed from the matrix $\Theta' \in \{0, 1\}^{N \times P}$ (a crisp variation of Θ), into K clusters. Given this insight, HBGF can be considered as the baseline model of LCE, where a more refined information matrix is exploited to improve the solution accuracy. The process of generating the final data partition π^* from Θ is summarised as follows.

Firstly, a weighted bipartite graph $G' = (V', W')$ is constructed from the matrix Θ, where $V' = V^X \cup V^C$ is a set of vertices representing both samples V^X and clusters V^C, and W' denotes a set of weighted edges. The weight $| w'_{ij} |$ of edge w'_{ij} connecting vertices v_i, $v_j \in V$, can be defined by

- $| w'_{ij} | = 0$ when $v_i, v_j \in V^X$ or $v_i, v_j \in V^C$.
- Otherwise, $|w_{ij}| = \Theta(v_i, v_j)$ when $v_i \in V^X$ and $v_j \in V^C$. Note that G is bidirectional such that $|w_{ij}| = |w_{ji}|$. In other words, $W' \in [0, 1]^{(N+P)\times(N+P)}$ can also be specified as

$$W' = \begin{bmatrix} 0 & \Theta \\ \Theta^T & 0 \end{bmatrix} \tag{5}$$

After that, the K largest eigenvectors $u_1, u_2, \ldots, u_K$ of W' are used to produce the matrix $U = [u_1 u_2, \ldots, u_K]$, in which the eigenvectors are stacked in columns. Then, another matrix $U^* \in [0, 1]^{(N+P)\times K}$ is formed by normalising each of U's row to have unit length.

$$U^*_{ss'} = \frac{U_{ss'}}{\sqrt{\sum_{s'=1}^{K} U^2_{ss'}}}, \tag{6}$$

where $s = 1, \ldots, (N + P)$. By considering each row of U^* as K-dimensional embedding of a graph vertex or a sample in $[0, 1]^K$), k-means is finally used to generate the final partition $\pi^* = \{C^*_1, \ldots, C^*_K\}$ of K clusters.

4 Performance evaluation

This section presents the performance evaluation of the new link-based approach, using a number of benchmark validity criteria, datasets and compared methods.

4.1 Investigated datasets and compared methods

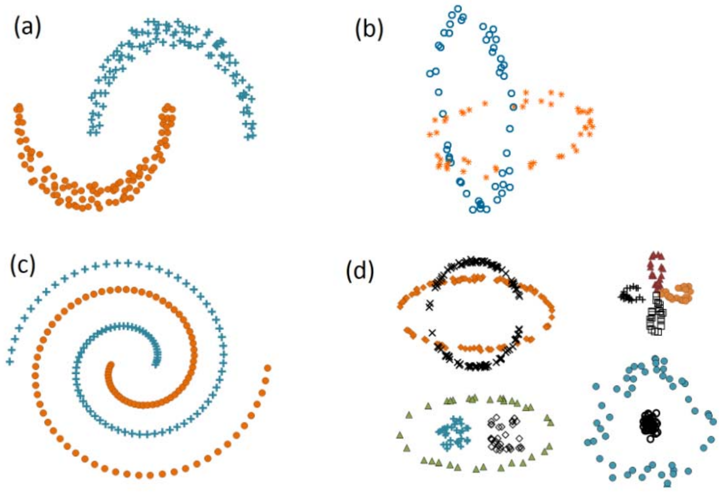
The experimental evaluation is conducted over eight datasets. Table 1 summarises the details of these datasets that are grouped into synthetic and real categories. In addition to the synthetic data collection obtained from the previous studies of cluster ensembles and shown in Figure 3, four real datasets obtained from the benchmark UCI repository (Asuncion and Newman, 2007) are also employed.

A collection of compared methods includes several state-of-the-art cluster ensemble categories: pairwise-similarity based [EAC-SL and EAC-AL of Fred and Jain (2005)], graph-based [HBGF of Fern and Brodley (2004), CSPA, HGPA and MCLA of Strehl and Ghosh (2002)], and feature-based [IVC of Nguyen and Caruana (2007)]. Details of these algorithms are not provided owing to the limited space. Note that the original link-based method and the improved model will be referred to as LCE and LCE* hereafter.

Table 1 Description of datasets: number of data points (N), number of attributes (D), number of classes (K) and source

Dataset	*N*	*D*	*K*	*Source*
Synthetic dataset:				
2-banana	200	2	2	Iam-On et al. (2008)
2-doughnut	200	2	2	Iam-On et al. (2008)
2-spiral	190	2	2	Iam-On et al. (2008)
Complex image	500	2	11	Iam-On et al. (2011)
Real dataset:				
Glass	214	9	6	UCI; Asuncion and Newman (2007)
Iris	150	4	3	UCI; Asuncion and Newman (2007)
Ionosphere	351	34	2	UCI; Asuncion and Newman (2007)
Breast-cancer	683	9	2	UCI; Asuncion and Newman (2007)

Figure 3 Synthetic datasets: (a) 2-banana, (b) 2-doughnut, (c) 2-spiral, and (d) complex image (see online version for colours)



4.2 Experimental design

For comparison, as suggested by Fern and Brodley (2004), Fred and Jain (2005), and Iam-On et al. (2011), each clustering 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 using the evaluation indices of: adjusted rand (AR) index (Rand, 1971) and classification accuracy (CA; Nguyen and Caruana, 2007). Note that, true classes are known for all datasets 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 initialisation of cluster centres.
- Two schemes for selecting the number of clusters (k) in each base clustering are:

  1 Fixed-k where k is fixed to $\lceil \sqrt{N} \rceil$

  2 Random-k where k is a random number in the range of $\{2, \lceil \sqrt{N} \rceil\}$.

  These strategies aim to generate diversity in the ensemble by following the intuition introduced by Fred and Jain (2005), Hadjitodorov et al. (2006), and Kuncheva and Vetrov (2006). It is suggested that k should be greater than the expected number of clusters and the common rule-of-thumb is $k = \sqrt{N}$.
- Ensemble size (M) of 10 is experimented.
- The DC of 0.9 is used with the link-based similarity algorithms.
- The quality of each cluster ensemble method with respect to a specific ensemble setting is generalised as the average of 50 trials.

4.3 Experimental results

Based on the AR measure, Table 2 compares the performance of different cluster ensemble methods over synthetic and real datasets, respectively. These results suggest that LCE* can generally improve the accuracy of the original model, i.e., LCE. Also, its performance is usually better than other cluster ensemble methods examined in this experiment, including HBGF that is its baseline model. Note that EAC-SL is highly accurate for synthetic data which is typically a connected pattern. However, it is not quite effective for the real datasets. This also applies to the graph-based methods of CSPA, HGPA and MCLA. As an example of the feature-based approach, IVC appears to be less accurate than those belonging to the other categories. Similar experimental results with these methods are observed using CA evaluation index.

In order to further evaluate the quality of identified techniques, the number of times that one method is significantly *better* and *worse* (of 95% confidence level) than the others are assessed across experimented datasets. Let $\overline{X}_C(i, \beta)$ be the average value of validity index $C \in \{AR, CA\}$ across n runs (n = 50 in this evaluation) for a clustering method $i \in CM$ (CM is a set of 9 experimented clustering methods), on a specific experiment setting $\beta \in ST$ (ST is a set of 16 unique combination of two ensemble types and eight datasets). The 95% 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 the followings.

$$L_{\overline{X}_C(i,\beta)} = \overline{X}_C(i, \beta) - 1.96 \frac{S_C(i, \beta)}{\sqrt{n}} \tag{7}$$

$$U_{\overline{X}_C(i,\beta)} = \overline{X}_C(i, \beta) + 1.96 \frac{S_C(i, \beta)}{\sqrt{n}} \tag{8}$$

Note that $SC(i, \beta)$ is the standard deviation of the validity index C across n runs for a clustering method i and an experiment setting β, and 1.96 is the critical z-score value for a 95% confidence interval (Gosling, 1995). 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

$$B_C(i) = \sum_{\forall \beta \in ST} \sum_{\forall i^* \in CM, i^* \neq i} better_C^{\beta}(i, i^*) \tag{9}$$

$$better_C^{\beta}(i, i^*) = \begin{cases} 1 & \text{if } L_{\overline{X}_C(i,\beta)} > U_{\overline{X}_C(i,\beta)} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

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

$$W_C(i) = \sum_{\forall \beta \in ST} \sum_{\forall i^* \in CM, i \neq i} worse_C^{\beta}(i, i^*) \tag{11}$$

$$worse_C^{\beta}(i, i^*) = \begin{cases} 1 & \text{if } U_{\overline{X}_C(i,\beta)} < L_{\overline{X}_C(i,\beta)} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Figure 4 The statistics of better and worse performance, summarised across all experiment settings, based on AR validity measures (see online version for colours)

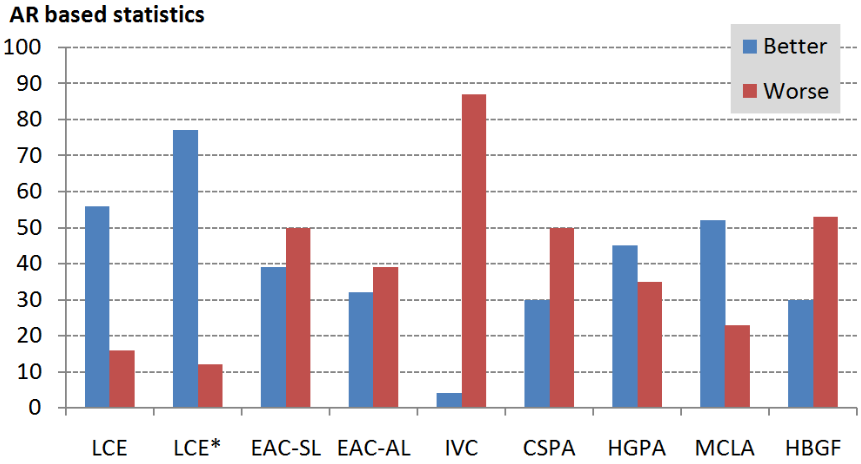


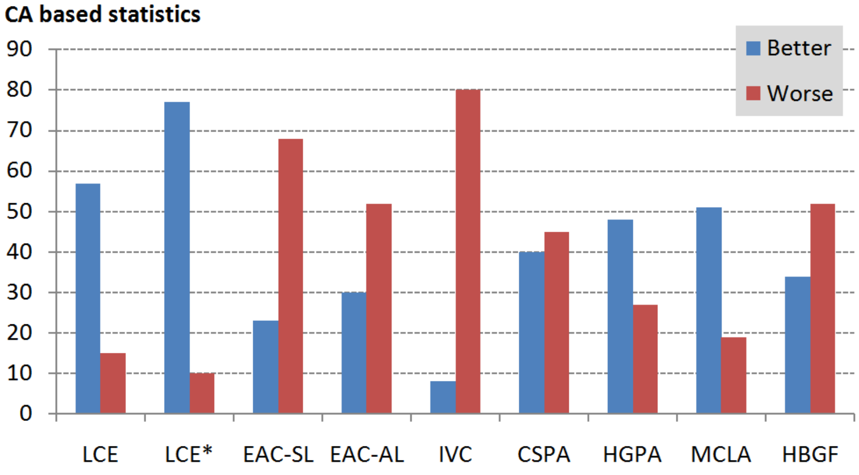
Table 2 AR scores of different cluster ensemble methods

Dataset	*Type*	*Methods*								
		LCE	*LCE**	*EAC-SL*	*EAC-AL*	*IVC*	*CSPA*	*HGPA*	*MCLA*	*HBGF*
2-banana	F	*1.000*	*1.000*	*1.000*	*1.000*	0.043	*1.000*	*1.000*	*1.000*	0.769
		(0.000)	(0.000)	(0.000)	(0.000)	(0.020)	(0.000)	(0.000)	(0.000)	(0.372)
	R	0.827	0.948	*0.963*	0.665	0.371	0.650	0.877	*0.967*	0.583
		(0.200)	(0.090)	(0.182)	(0.262)	(0.190)	(0.264)	(0.218)	(0.120)	(0.210)
2-doughnut	F	0.900	*0.943*	0.871	0.758	0.116	0.756	*1.000*	0.934	0.542
		(0.191)	(0.157)	(0.303)	(0.367)	(0.075)	(0.394)	(0.000)	(0.225)	(0.391)
	R	0.164	*0.476*	0.157	0.109	0.146	0.047	*0.318*	0.214	0.128
		(0.174)	(0.257)	(0.227)	(0.092)	(0.110)	(0.044)	(0.379)	(0.296)	(0.105)
2-spiral	F	0.003	0.005	*0.027*	0.012	0.007	0.026	0.022	0.033	0.007
		(0.006)	(0.008)	(0.036)	(0.032)	(0.019)	(0.023)	(0.032)	(0.033)	(0.030)
	R	0.058	*0.070*	0.015	0.047	0.028	*0.063*	0.059	0.062	0.055
		(0.022)	(0.025)	(0.028)	(0.031)	(0.028)	(0.032)	(0.038)	(0.032)	(0.025)
Complex image	F	*0.463*	0.462	*0.508*	0.406	0.344	0.409	0.443	0.430	0.304
		(0.029)	(0.024)	(0.042)	(0.037)	(0.068)	(0.017)	(0.030)	(0.026)	(0.044)
	R	0.442	*0.450*	*0.524*	0.396	0.408	0.405	0.389	0.421	0.390
		(0.032)	(0.030)	(0.072)	(0.040)	(0.048)	(0.021)	(0.037)	(0.058)	(0.050)
Glass	F	0.196	*0.229*	*0.237*	0.194	0.131	0.178	0.139	0.186	0.190
		(0.038)	(0.024)	(0.056)	(0.020)	(0.059)	(0.023)	(0.029)	(0.035)	(0.020)
	R	0.180	*0.205*	0.190	0.175	0.168	0.165	0.157	0.185	*0.195*
		(0.023)	(0.033)	(0.085)	(0.028)	(0.031)	(0.026)	(0.033)	(0.034)	(0.036)
Iris	F	0.864	0.867	0.592	0.686	0.192	*0.878*	*0.868*	0.804	0.574
		(0.038)	(0.043)	(0.136)	(0.134)	(0.053)	(0.045)	(0.061)	(0.133)	(0.203)
	R	0.797	*0.817*	0.614	0.672	0.462	*0.842*	0.809	0.810	0.653
		(0.088)	(0.075)	(0.152)	(0.084)	(0.149)	(0.110)	(0.122)	(0.135)	(0.169)
Ionosphere	F	0.127	*0.175*	-0.018	0.143	0.033	0.104	0.125	*0.161*	0.116
		(0.079)	(0.021)	(0.032)	(0.123)	(0.099)	(0.039)	(0.045)	(0.065)	(0.087)
	R	*0.173*	*0.171*	-0.021	0.122	0.115	0.128	0.081	0.146	0.131
		(0.020)	(0.016)	(0.024)	(0.108)	(0.166)	(0.006)	(0.059)	(0.046)	(0.089)
Breast-cancer	F	*0.881*	0.865	0.003	0.862	0.062	0.177	0.513	0.524	*0.876*
		(0.011)	(0.334)	(0.018)	(0.035)	(0.058)	(0.187)	(0.088)	(0.076)	(0.017)
	R	0.818	*0.882*	0.189	0.847	0.443	0.365	0.452	0.625	*0.878*
		(0.227)	(0.010)	(0.337)	(0.046)	(0.352)	(0.158)	(0.076)	(0.164)	(0.014)

Notes: The two highest scores of each ensemble type are highlighted in italic and corresponding standard deviation values are given in parentheses. Note that ‘type’ is the ensemble type of fixed-k or random-k, which is denoted shortly by ‘F’ and ‘R’, respectively.

Using the aforementioned assessment formalism, Figure 4 summarises for each method $i \in CM$ both better $B_C(i)$ and worse $W_C(i)$ statistics, based on the validity index AR. The results shown in this figure indicate the superior effectiveness of the link-based methods, as compared to other cluster ensemble techniques. This reinforces the aforementioned findings that LCE* can improve the quality of clusterings of the previous LCE counterpart. In addition, MCLA and IVC appear to be the most and the least accurate amongst the compared methods, respectively. Similar trends are observed with the statistics collected from CA index. This is presented in Figure 5.

Figure 5 The statistics of better and worse performance, summarised across all experiment settings, based on CA validity measures (see online version for colours)



4.4 Time complexity analysis

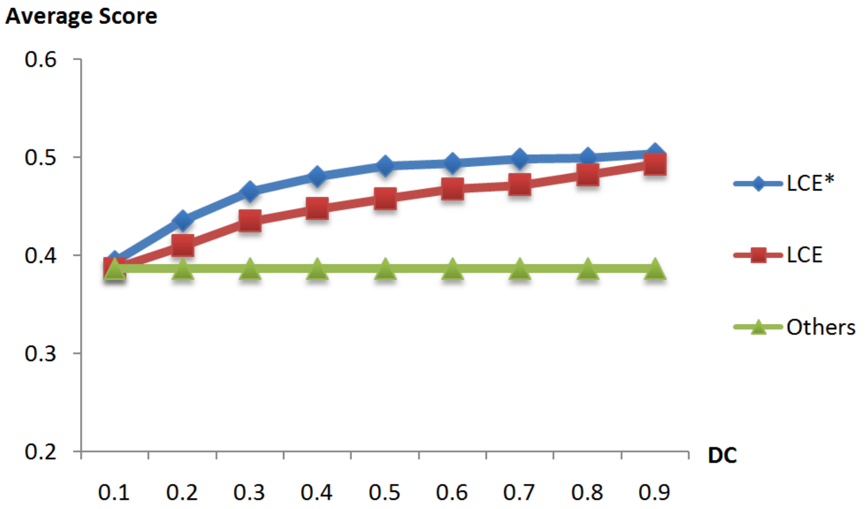
Besides the previous quality assessments, computational time requirements of the link-based methods are discussed here. As reported by Iam-On et al. (2010) for the WCT algorithm, the time complexity of creating the matrix Θ is $O(P^2l + NP)$, where N is the number of data points, P denotes the number of all clusters in an ensemble Π and l represents the average number of neighbours 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 neighbours of C_x (or C_y) to identify triples. Following this, the Θ matrix of size $N \times P$ is created using the aforementioned similarity matrix. As the extension of WCT, WTU continues searching through l neighbours of each potential triple identified earlier. Hence, the time complexity of WTU is $O(P^2l^2 + NP)$, which converges to that of WCT.

4.5 Parameter analysis

In addition, the parameter analysis is also conducted with the aim to provide a practical means by which users can make the best use of the proposed link-based method. As with LCE, the performance of LCE* is also dependent on the *DC* value, which is used in estimating the similarity amongst clusters and refining the original BA matrix. Specific to this issue, Figure 6 illustrates such a relationship, based on the average of two validity measures (AR and CA) across all experiment settings. For both link-based cluster ensemble models, 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 Figure 6). 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 LCE* are consistently higher than those of LCE, even when the value of *DC* is low.

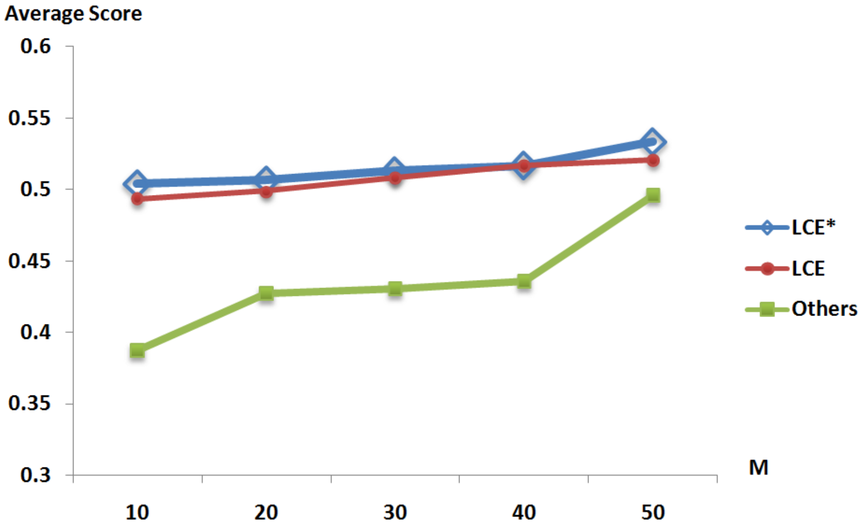
Another important parameter that may determine the quality of a cluster ensemble technique, is the ensemble size (*M*). Intuitively, the larger an ensemble is, the better the performance becomes. According to Figure 7 in which *DC* = 0.9, this heuristic is applicable to LCE*, where its validity measures (averages of AR and CA across all experimental settings) gradually incline to the increasing value of $M \in \{10, 20, \ldots, 50\}$. Furthermore, LCE* performs better than LCE and other competitors (indicated by 'others') with all different ensemble sizes investigated here.

Figure 6 The relations between $DC \in \{0.1, 0.2, \ldots, 0.9\}$ and the performance of link-based algorithms (the average of AR and CA measures), whose values are presented in X-axis and Y-axis, respectively (see online version for colours)



Note: Others denotes the average evaluation measure across all compared cluster ensemble methods.

Figure 7 Performance of different cluster ensemble methods in accordance with ensemble size ($M \in \{10, 20, \ldots, 50\}$), as the averages of validity measures (AR and CA) across all experiment settings (see online version for colours)



5 Application to cancer microarray data analysis

Cancer gene expression data obtained from microarray experiments has inspired several applications, including the identification of differentially expressed genes for molecular studies or drug therapy response and the creation of classification systems for improved cancer diagnosis (Spang, 2003). Cluster analysis has proven useful for identifying biologically relevant groups of tissue samples and genes. The present research focuses on the former where samples with similar profiles of gene-specific expression values are grouped together. Clinical researchers commonly use simple clustering methods, such as agglomerative hierarchical and k-means to cluster cancer microarray samples, despite the advent of several new techniques that capitalise on the inherent characteristics of gene expression data (noise, high dimensionality) to improve clustering quality (McLachlan et al., 2002). This is because the use of such methods is difficult for non-expert users (de Souto et al., 2008).

Cluster ensembles have recently become an attractive alternative for microarray data analysis (Kim et al., 2009; Monti et al., 2003; Yu et al., 2007). In particular, a link-based method (LCE) introduced by Iam-On et al. (2010) has shown to be more effective than the previous cluster ensemble methods adopted for microarray data analysis. This motivates the use of the new link-based model (LCE*) for such an analytic task. In the following sections, the empirical study of LCE* and a set of gene expression data is presented.

5.1 Experimental design

The experimental assessment is conducted over six published microarray datasets. Table 3 summarises their details, which can be further consulted in the study of de Souto et al. (2008). A collection of compared methods includes several state-of-the-art cluster ensemble categories: pairwise-similarity based [MULTI-K of Kim et al. (2009) and CC_{HC} of Monti et al. (2003)], graph-based [HBGF of Fern and Brodley (2004), CSPA, HGPA and MCLA of Strehl and Ghosh (2002) and GCC of Yu et al. (2007)], and feature-based [AGG of Gionis et al. (2007), IVC of Nguyen and Caruana (2007), MM of Topchy et al. (2005), and QMI of Topchy et al. (2005)]. Details of these algorithms are not provided due to the available space. Note that the original link-based method and the improved model will be referred to as LCE and LCE* hereafter. In addition to the aforementioned ensemble techniques, *k*-means (KM) and three basic variations of hierarchical clusterings (i.e., SL, CL and AL) are also examined in this empirical study.

Table 3 Description of investigated gene expression datasets: numbers of samples (*N*), genes (*D*) and known classes (*K*)

Dataset	*Chip*	*Tissue*	*N*	*D*	*K*
Brain Tumor	Affy	Brain	22	1,152	2
CNS	Affy	Brain	42	1,379	5
Breast Tumor	Affy	Breast	49	1,198	2
HCC	cDNA	Liver	180	85	2
SRBCTs	cDNA	Multi-tissue	83	1,069	4
Prostate Cancer	cDNA	Prostate	104	2,315	5

For comparison, each clustering 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 using the evaluation indices of: AR index (Rand, 1971) and CA (Nguyen and Caruana, 2007). Note that, true classes are known for all datasets 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 identical listed in Section 4.

5.2 Experimental results

Based on the AR measure, Table 4 compares the performance of different cluster ensemble methods and basic clustering algorithms over the investigated datasets. Note that standard deviation values of SL, CL and AL are not presented as they are deterministic technique whose results are indifferent for multiple runs. These results suggest that LCE* can generally improve the accuracy of the original model, i.e., LCE. Also, its performance is usually better than other cluster ensemble methods examined in this experiment. Another important observation is that LCE* always produces a data partition of higher quality than base clusterings, i.e., KM. This does not hold true for some techniques, such as MULTI-K, MM and IVC. Similar experimental results with these methods are observed using CA evaluation index.

Table 4 AR measures of examined clustering methods

Method	*Dataset*											
	Brain tumour		*CNS*		*Breast tumour*		*HCC*		*SRBCTs*		*Prostate cancer*	
	F	*R*	*F*	*R*	*F*	*R*	*F*	*R*	*F*	*R*	*F*	*R*
LCE*	0.643	*1.000*	*0.389*	*0.312*	0.461	0.229	*0.471*	*0.550*	*0.140*	*0.119*	*0.261*	*0.250*
	(0.412)	(0.000)	(0.103)	(0.047)	(0.125)	(0.244)	(0.122)	(0.174)	(0.047)	(0.065)	(0.047)	(0.049)
LCE	0.471	0.959	*0.352*	*0.299*	*0.457*	0.308	*0.478*	*0.551*	0.115	0.067	0.240	*0.251*
	(0.413)	(0.166)	(0.075)	(0.061)	(0.145)	(0.250)	(0.087)	(0.219)	(0.037)	(0.052)	(0.050)	(0.058)
MULTI-K	0.561	0.874	0.185	0.187	–0.002	–0.002	–0.005	–0.006	0.108	*0.098*	0.183	0.184
	(0.429)	(0.312)	(0.091)	(0.086)	(0.001)	(0.001)	(0.004)	(0.000)	(0.043)	(0.061)	(0.023)	(0.014)
CCHC	0.617	0.988	0.186	0.250	–0.002	–0.002	–0.006	0.005	0.077	0.043	0.205	0.244
	(0.416)	(0.087)	(0.087)	(0.085)	(0.001)	(0.001)	(0.004)	(0.078)	(0.048)	(0.046)	(0.039)	(0.047)
GCC	*0.742*	0.975	0.263	0.296	0.440	0.025	0.005	0.050	0.072	0.045	0.221	0.224
	(0.393)	(0.122)	(0.077)	(0.053)	(0.148)	(0.102)	(0.069)	(0.192)	(0.047)	(0.056)	(0.058)	(0.035)
CSPA	0.167	0.465	0.287	0.247	0.442	*0.441*	0.416	0.453	0.074	0.079	0.226	0.215
	(0.120)	(0.085)	(0.074)	(0.064)	(0.000)	(0.031)	(0.052)	(0.087)	(0.026)	(0.019)	(0.035)	(0.022)
HGPA	0.144	0.220	0.272	0.249	0.416	0.237	0.459	0.451	0.107	0.091	*0.244*	0.238
	(0.040)	(0.089)	(0.068)	(0.083)	(0.144)	(0.229)	(0.069)	(0.108)	(0.051)	(0.041)	(0.025)	(0.038)
MCLA	0.174	0.826	0.257	0.174	0.387	0.186	0.347	0.422	*0.131*	0.081	0.227	0.216
	(0.146)	(0.297)	(0.102)	(0.055)	(0.103)	(0.206)	(0.109)	(0.124)	(0.055)	(0.060)	(0.043)	(0.060)

Notes: The highest two measures for each dataset are highlighted in BOLDFACE, and the corresponding standard deviation values are given in parentheses. Note that the ensemble type of Fixed-k and Random-k are denoted shortly by ‘F’ and ‘R’, respectively.

Table 4 AR measures of examined clustering methods (continued)

Method	*Dataset*											
	Brain tumour		*CNS*		*Breast tumour*		*HCC*		*SRBCTs*		*Prostate cancer*	
	F	*R*	*F*	*R*	*F*	*R*	*F*	*R*	*F*	*R*	*F*	*R*
HBGF	0.740	0.957	0.303	0.293	0.151	0.018	0.237	0.435	0.101	0.052	0.229	0.231
	(0.384)	(0.172)	(0.077)	(0.051)	(0.224)	(0.098)	(0.245)	(0.317)	(0.039)	(0.046)	(0.058)	(0.041)
AGG	*0.886*	*1.000*	0.287	0.251	0.387	0.014	0.001	0.108	0.070	0.051	0.105	0.246
	(0.163)	(0.000)	(0.069)	(0.073)	(0.099)	(0.076)	(0.005)	(0.198)	(0.037)	(0.064)	(0.042)	(0.054)
IVC	0.449	0.714	0.314	0.236	0.426	0.368	0.173	0.348	0.078	0.054	0.185	0.208
	(0.441)	(0.423)	(0.102)	(0.096)	(0.179)	(0.182)	(0.206)	(0.304)	(0.072)	(0.054)	(0.068)	(0.067)
MM	0.268	0.467	0.283	0.259	0.382	*0.389*	0.135	0.308	0.104	0.072	0.174	0.217
	(0.349)	(0.419)	(0.092)	(0.056)	(0.176)	(0.183)	(0.141)	(0.201)	(0.052)	(0.043)	(0.058)	(0.064)
QMI	0.488	0.686	0.332	*0.299*	0.338	0.313	0.158	0.419	0.085	0.050	0.203	0.217
	(0.377)	(0.399)	(0.090)	(0.070)	(0.218)	(0.222)	(0.180)	(0.266)	(0.051)	(0.049)	(0.083)	(0.049)
SL	0.102		0.020		–0.002		–0.003		–0.008		0.010	
CL	0.102		0.095		–0.002		–0.006		–0.023		0.091	
AL	0.102		0.020		–0.002		–0.003		–0.027		0.010	
KM	0.580 (0.340)		0.089 (0.091)		0.074 (0.161)		0.064 (0.193)		0.064 (0.074)		0.200 (0.062)	

Notes: The highest two measures for each dataset are highlighted in BOLDFACE, and the corresponding standard deviation values are given in parentheses. Note that the ensemble type of Fixed-k and Random-k are denoted shortly by ‘F’ and ‘R’, respectively.

To further evaluate the quality of identified techniques, the number of times that one method is significantly *better* and *worse* (of 95% confidence level) than the others is assessed across datasets. Figure 8 summarises for each method, both better and worse statistics, based on AR validity index. The results shown in this figure indicate the superior effectiveness of LCE*, as compared to LCE and other cluster ensemble techniques. In addition, HGPA and MULTI-K appear to be the most and the least accurate amongst the compared methods, respectively. Similar trends are observed with the statistics collected from CA validity index. See Figure 9 for details.

Figure 8 The statistics of better and worse performance, summarised across all experiment settings, based on AR validity measures (see online version for colours)

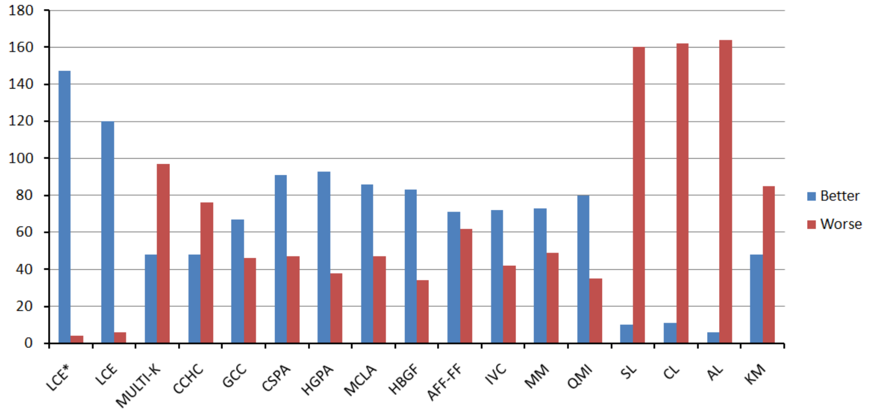
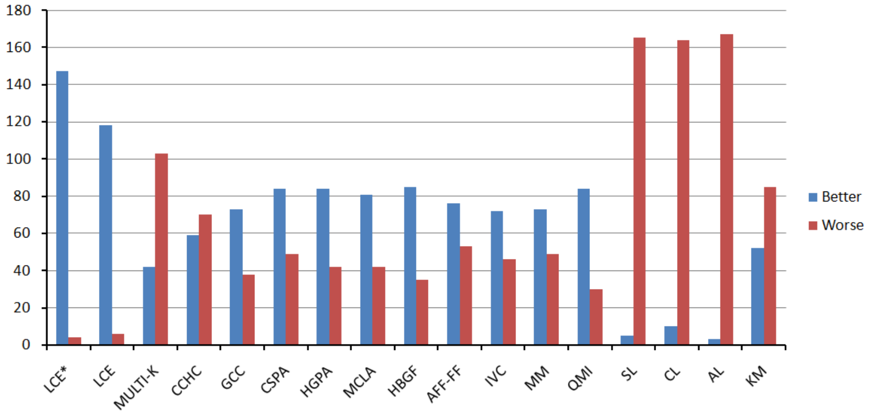


Figure 9 The statistics of better and worse performance, summarised across all experiment settings, based on CA validity measures (see online version for colours)



6 Conclusions

This paper has presented the improved model of link-based cluster ensembles, with a new similarity measure being developed such that additional information within a link network is included. This helps to refine the resulting measures, hence the quality of the information matrix. The new method has proven more effective than the original and several other methods found in the literature. Its performance is also robust to parameter settings, which can be useful for less-experienced users. Despite this encouraging result, there are several subjects needed to be further investigated. First, an automated and data-driven setting of DC can provide a more accurate outcome, as compared to a manual configuration. Following the study of Iam-On et al. (2010), it is interesting observe the performance of this improved framework with the problem of microarray data analysis, with which link-based approach has been successful. Also, new clustering algorithms, e.g., that of Das et al. (2008), can be used to form a more accurate cluster ensemble.

References

Ahmad, A. and Dey, L. (2007) 'A k-mean clustering algorithm for mixed numeric and categorical data', *Data and Knowledge Engineering*, Vol. 63, No. 2, pp.503–527.

Asuncion, A. and Newman, D.J. (2007) *UCI Machine Learning Repository*, University of California, School of Information and Computer Science, Irvine, CA [online] www.ics.uci.edu/~mlearn/MLRepository.html].

Ayad, H. and Kamel, M. (2003) 'Finding natural clusters using multiclusterer combiner based on shared nearest neighbors', in *Proceedings of International Workshop on Multiple Classifier Systems*, pp.166–175.

Bhatia, S.K. and Deogun, J.S. (1998) 'Conceptual clustering in information retrieval', *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 28, pp.427–436.

Boongoen, T., Shang, C., Iam-On, N. and Shen, Q. (2011) 'Extending data reliability measure to a filter approach for soft subspace clustering', *IEEE Transactions on Systems, Man and Cybernetics, Part B*, Vol. 41, No. 6, pp.1705–1714.

Boongoen, T., Shen, Q. and Price, C. (2010) 'Disclosing false identity through hybrid link analysis', *AI and Law*, Vol. 18, No. 1, pp.77–102.

Boulis, C. and Ostendorf, M. (2004) 'Combining multiple clustering systems', in *Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp.63–74.

Costa, J.A.F. and 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.

Das, S., Abraham, A. and Konar, A. (2008) 'Automatic kernel clustering with a multielitist particle swarm optimization algorithm', *Pattern Recognition Letters*, Vol. 29, No. 5, pp.688–699.

de Souto, M., Costa, I., de Araujo, D., Ludermir, T. and Schliep, A. (2008) 'Clustering cancer gene expression data: a comparative study', *BMC Bioinformatics*, Vol. 9, p.497.

Domeniconi, C. and Al-Razgan, M. (2009) 'Weighted cluster ensembles: methods and analysis', *ACM Transactions on Knowledge Discovery from Data*, Vol. 2, No. 4, pp.1–40.

Duda, R.O., Hart, P.E. and Stork, D.G. (2000) *Pattern Classification*, 2nd ed., Wiley-Interscience, New York, USA.

Dudoit, S. and Fridyand, J. (2003) 'Bagging to improve the accuracy of a clustering procedure', *Bioinformatics*, Vol. 19, No. 9, pp.1090–1099.

Fern, X.Z. and 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. and Brodley, C.E. (2004) 'Solving cluster ensemble problems by bipartite graph partitioning', in *Proceedings of International Conference on Machine Learning*, pp.36–43.

Fischer, B. and Buhmann, J.M. (2003) 'Bagging for path-based clustering', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 11, pp.1411–1415.

Fred, A.L.N. and Jain, A.K. (2005) 'Combining multiple clusterings using evidence accumulation', *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 6, pp.835–850.

Gionis, A., Mannila, H. and Tsaparas, P. (2007) 'Clustering aggregation', *ACM Transactions on Knowledge Discovery from Data*, Vol. 1, No. 1, p.4-ex.

Gosling, J. (1995) *Introductory Statistics*, Pascal Press, New South Wales, Australia.

Hadjitodorov, S.T., Kuncheva, L.I. and Todorova, L.P. (2006) 'Moderate diversity for better cluster ensembles', *Information Fusion*, Vol. 7, No. 3, pp.264–275.

Hu, X. and 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. and 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. and Garrett, S. (2010) 'LCE: a link-based cluster ensemble method for improved gene expression data analysis', *Bioinformatics*, Vol. 26, No. 12, pp.1513–1519.

Iam-On, N., Boongoen, T., Garrett, S. and Price, C. (2011) 'A link-based approach to the cluster ensemble problem', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 12, pp.2396–2409.

Jiang, D., Tang, C. and Zhang, A. (2004) 'Cluster analysis for gene expression data: a survey', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 11, pp.1370–1386.

Karypis, G. and Kumar, V. (1998) 'Multilevel k-way partitioning scheme for irregular graphs', *Journal of Parallel Distributed Computing*, Vol. 48, No. 1, pp.96–129.

Kim, E., Kim, S., Ashlock, D. and Nam, D. (2009) 'MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering', *BMC Bioinformatics*, Vol. 10, p.260.

Kim, K. and Ahn, H. (2008) 'A recommender system using GA K-means clustering in an online shopping market', *Expert Systems with Applications*, Vol. 34, pp.1200–1209.

Kittler, J., Hatef, M., Duin, R. and Matas, J. (1998) 'On combining classifiers', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, pp.226–239.

Kuncheva, L.I. and Vetrov, D. (2006) 'Evaluation of stability of k-means cluster ensembles with respect to random initialization', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 11, pp.1798–1808.

McLachlan, G., Bean, R. and Peel, D. (2002) 'A mixture model-based approach to the clustering of microarray expression data', *Bioinformatics*, Vol. 18, No. 3, pp.413–422.

Monti, S., Tamayo, P., Mesirov, J.P. and Golub, T.R. (2003) '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.

Ng, A., Jordan, M. and Weiss, Y. (2001) 'On spectral clustering: analysis and an algorithm', *Advances in Neural Information Processing Systems*, Vol. 14, pp.849–856.

Nguyen, N. and Caruana, R. (2007) 'Consensus clusterings', in *Proceedings of IEEE International Conference on Data Mining*, pp.607–612.

Rand, W.M. (1971) 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical Association*, Vol. 66, pp.846–850.

Spang, R. (2003) 'Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine', *BIOSILICO*, Vol. 1, pp.264–268.

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

Topchy, A.P., Jain, A.K. and Punch, W.F. (2004) 'A mixture model for clustering ensembles', in *Proceedings of SIAM International Conference on Data Mining*, pp.379–390.

Topchy, A.P., Jain, A.K. and Punch, W.F. (2005) 'Clustering ensembles: models of consensus and weak partitions', *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 12, pp.1866–1881.

Wu, R.C., Chen, R.S., Chang, C.C. and 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. and Yang, Q. (2009) 'Discriminatively regularized least-squares classification', *Pattern Recognition*, Vol. 42, No. 1, pp.93–104.

Yu, Z., Wong, H-S. and Wang, H. (2007) 'Graph-based consensus clustering for class discovery from gene expression data', *Bioinformatics*, Vol. 23, No. 21, pp.2888–2896.

Zhang, J., Mostafa, J. and Tripathy, H. (2002) 'Information retrieval by semantic analysis and visualisation of the concept space of D-Lib magazine', *D-Lib Magazine*, Vol. 8, No. 10.