BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 12 2010, pages 1513–1519
doi:10.1093/bioinformatics/btq226

Gene expression Advance Access publication May 5, 2010

LCE: a link-based cluster ensemble method for improved gene expression data analysis

Natthakan Iam-on[1,*], Tossapon Boongoen[1,2] and Simon Garrett[1]

[1]Department of Computer Science, Aberystwyth University, Aberystwyth, Ceredigion, UK and [2]Department of Mathematics and Computer Science, Royal Thai Air Force Academy, Thailand

Associate Editor: Olga Troyanskaya

ABSTRACT

Motivation: It is far from trivial to select the most effective clustering method and its parameterization, for a particular set of gene expression data, because there are a very large number of possibilities. Although many researchers still prefer to use hierarchical clustering in one form or another, this is often sub-optimal. Cluster ensemble research solves this problem by automatically combining multiple data partitions from different clusterings to improve both the robustness and quality of the clustering result. However, many existing ensemble techniques use an association matrix to summarize sample-cluster co-occurrence statistics, and relations within an ensemble are encapsulated only at coarse level, while those existing among clusters are completely neglected. Discovering these missing associations may greatly extend the capability of the ensemble methodology for microarray data clustering.

Results: The link-based cluster ensemble (LCE) method, presented here, implements these ideas and demonstrates outstanding performance. Experiment results on real gene expression and synthetic datasets indicate that LCE: (i) usually outperforms the existing cluster ensemble algorithms in individual tests and, overall, is clearly class-leading; (ii) generates excellent, robust performance across different types of data, especially with the presence of noise and imbalanced data clusters; (iii) provides a high-level data matrix that is applicable to many numerical clustering techniques; and (iv) is computationally efficient for large datasets and gene clustering.

Availability: Online supplementary and implementation are available at: http://users.aber.ac.uk/nii07/bioinformatics2010

Contact: nii07@aber.ac.uk; natthakan@mfu.ac.th

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 14, 2009; revised on March 18, 2010; accepted on April 20, 2010

1 INTRODUCTION

The use of clustering is vital both for visualizing and extracting useful information from microarray data. However, different algorithms (or even the same algorithm with different parameters) often provide distinct clusterings. As a result, it is extremely difficult for users to decide which algorithm and parameters will be *optimal* for a given set of data—this is because no single-pass/simple clustering algorithm can perform the best for all datasets (Kuncheva and Hadjitodorov, 2004), and discovering all types of cluster shapes and structures presented in data is impossible for any known clustering algorithm (Duda *et al.*, 2000; Handl *et al.*, 2005).

Clinical researchers commonly use simple clustering methods, such as agglomerative hierarchical and *k*-means (Bredel *et al.*, 2005; Sorlie *et al.*, 2003) to cluster cancer microarray samples, despite the advent of several new techniques that capitalize on the inherent characteristics of gene expression data (noise and high dimensionality) to improve clustering quality (e.g. Brunet *et al.*, 2004; Liu *et al.*, 2003; McLachlan *et al.*, 2002). de Souto *et al.* (2008) says, this is because the use of such methods is difficult for non-expert users.

Recently, *cluster ensembles* or *consensus clusterings* have emerged as simple, effective, one-stop methods for improving the robustness and quality of clustering results. Cluster ensembles combine multiple clustering decisions (referred to as 'base clusterings' or 'ensemble members') where the base clusterings contain diversity in their choice of clusters by: (i) using a single clustering algorithm with random parameter initializations (Kim *et al.*, 2009; Monti *et al.*, 2003; Yu *et al.*, 2007); (ii) employing multiple clustering algorithms (Swift *et al.*, 2004); (iii) selecting a random number of clusters (Fred and Jain, 2005; Kuncheva and Vetrov, 2006); (iv) using different subsets of gene (Avogadri and Valentini, 2009; Yu *et al.*, 2007); or (v) using data sampling techniques (Dudoit and Fridyand, 2003; Monti *et al.*, 2003). Most existing methods compare cluster associations between each of the N samples in the dataset to produce an $N \times N$ pairwise similarity matrix [i.e. *consensus* (Monti *et al.*, 2003), *agreement* (Swift *et al.*, 2004) and *co-association* (Fred and Jain, 2005) matrices], to which a consensus function (e.g. agglomerative hierarchical clustering) is applied to acquire the final data partition. With the ensemble of two base clusterings $\Pi = \{\pi_1, \pi_2\}$ and five samples $(x_1, \ldots x_5)$ that is given in Figure 1a, the corresponding similarity matrix is shown in Figure 1b.

An alternative approach (Fern and Brodley, 2004; Strehl and Ghosh, 2002) to pairwise similarity methods makes use of an $N \times P$ *binary cluster-association matrix* (*BM*) (where P denotes the number of clusters in an ensemble). Figure 1c shows the example of such matrix that is generated from the ensemble of Figure 1a. Despite reported success and efficiency, these methods generate the ultimate clustering result based on incomplete information of a cluster ensemble. The underlying association matrix presents sample–cluster relations at a coarse level and completely ignores the relations among clusters (Iam-on *et al.*, 2008). As a result, the

*To whom correspondence should be addressed.

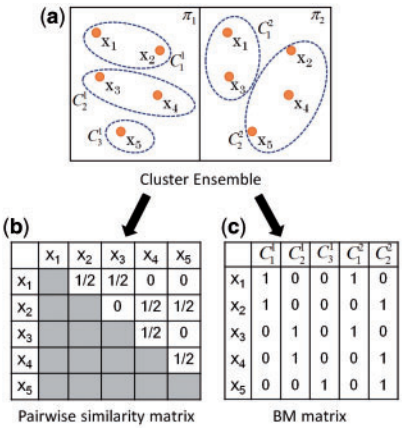


Fig. 1. An example of (**a**) cluster ensemble of samples $\{x_1 \ldots x_5\}$ that consists of two base clusterings ($\pi_1 = \{C_1^1, C_2^1, C_3^1\}$ and $\pi_2 = \{C_1^2, C_2^2\}$), (**b**) the corresponding pairwise similarity matrix and (**c**) BM, respectively.

performance of such techniques may consequently be degraded as many matrix entries are left *unknown*, each presented with zero.

In response, we present a new method—the LCE—for clustering data. It significantly extends the hybrid bipartite graph formulation (HBGF) technique (Fern and Brodley, 2004), by applying a graph-based consensus function to an improved cluster association matrix, instead of the conventional BM. This article explores its application to the problem of clustering cancer microarray samples, and is shown to refine the cluster-association matrix, as well as reducing the number of unknown entries and, therefore, increasing accuracy; moreover, it can easily replace or augment a researcher's existing clustering tools.

2 METHODS

The proposed LCE methodology is illustrated in Figure 2. It includes three major steps: (i) creating M base clusterings to form a cluster ensemble; (ii) creating a refined cluster-association matrix (RM) using a link-based similarity algorithm (Weighted Connected-Triples, WCT); and (iii) generating the final data partition by exploiting the spectral graph partitioning (SPEC) technique as a consensus function. This framework is similar to that of HBGF (Fern and Brodley, 2004), except the second step that is introduced for constructing a refined information matrix. As compared to HBGF that is based on the BM, LCE may enhance effectiveness of the former using a more informative RM.

2.1 Creating a cluster ensemble

Let $X = \{x_1, \ldots, x_N\}$ be a set of N samples and let $\Pi = \{\pi_1, \ldots, \pi_M\}$ be a cluster ensemble with M base clustering results. Each base clustering returns a set of clusters $\pi_i = \{C_1^i, C_2^i, \ldots, C_{k_i}^i\}$, such that $\bigcup_{j=1}^{k_i} C_j^i = X$, where k_i is the number of clusters in the i-th clustering.

As in many previous studies (Fred and Jain, 2005; Kim *et al.*, 2009), the k-means clustering algorithm is used to generate base clusterings, each with random initialization of cluster centers. Euclidean distance is used to measure the dissimilarity between two samples unless stated otherwise. For each base clustering, there are two schemes of selecting the number of clusters: Fixed-k ($k = \sqrt{N}$, where N is the number of samples) and Random-k ($k \in \{2, \ldots, \sqrt{N}\}$). To create diversity in an ensemble, 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). Note that the quality of the final clustering is directly subjected to the diversity among base clusterings (Kuncheva and Vetrov, 2006).

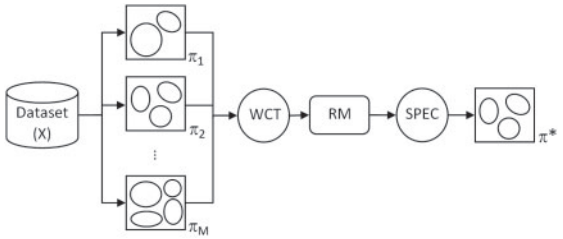


Fig. 2. The LCE framework: (i) a cluster ensemble $\Pi = \{\pi_1, \ldots, \pi_M\}$ is created from M base clusterings; (ii) a refined cluster-association matrix (RM) is then generated from the ensemble using the WCT algorithm; and (iii) a final clustering result (π^*) is produced by a consensus function of the spectral graph partitioning (SPEC).

Another alternative to generate diversity within an ensemble is to exploit a number of different data partitions. To this extent, the cluster ensemble is also established on various data subspaces. Similar to the study of Yu *et al.* (2007), for a given $N \times d$ dataset of N samples and d genes, an $N \times q$ data subspace (where $q < d$) is generated by

$$q = q_{\min} + \lfloor \alpha(q_{\max} - q_{\min}) \rfloor \quad (1)$$

here $\alpha \in [0, 1]$ is a uniform random variable, $q_{\min}$ and $q_{\max}$ are the lower and upper bounds of the generated subspace, respectively. In particular, $q_{\min}$ and $q_{\max}$ are set to $0.75d$ and $0.85d$. A gene is selected one by one from the pool of d genes, until the collection of q is obtained. The index of each selected gene is determined as follows, where h denotes the h-th gene in the pool of d genes and $\beta \in [0, 1)$ is a uniform random variable.

$$h = \lfloor 1 + \beta d \rfloor \quad (2)$$

Generating a refined cluster-association matrix (RM)

In particular to HBGF, BM has been used to summarized information presented in an ensemble Π. Each entry in this matrix $\mathrm{BM}(x_i, C_j) \in \{0, 1\}$ represents a *crisp* association degree between sample $x_i \in X$ and cluster $C_j \in \Pi$. According to Figure 1, which shows an example of cluster ensemble and the corresponding BM, a large number of entries in the BM are *unknown*, each presented with 0. Intuitively, this may limit the quality of a data partition generated by any consensus function. These conditions occur when relations between different clusters of a base clustering are originally assumed to be nil. It is important to note that each sample can associate (to a certain degree within $[0, 1]$) to several clusters of any particular clustering, at the same time. These hidden or unknown associations can be estimated upon the similarity among clusters, discovered from a link network of clusters.

Based on this insight, the refined cluster-association matrix (RM) is put forward as the enhanced variation of the original BM. Its aim is to approximate value of unknown associations ('0') from known ones ('1'), whose association degrees are preserved within the RM (i.e. $\mathrm{BM}(x_i, cl) = 1 \rightarrow \mathrm{RM}(x_i, cl) = 1$). For each clustering $\pi_t, t = 1 \ldots M$ and their corresponding clusters $C_1^t, \ldots, C_{k_t}^t$ (where k_t is the number of clusters in the clustering π_t), the association degree $\mathrm{RM}(x_i, cl) \in [0, 1]$ that sample $x_i \in X$ has with each cluster $cl \in \{C_1^t, \ldots, C_{k_t}^t\}$ is estimated as follows:

$$\mathrm{RM}(x_i, cl) = \begin{cases} 1 & \text{if } cl = C_*^t(x_i) \\ \mathrm{sim}(cl, C_*^t(x_i)) & \text{otherwise} \end{cases} \quad (3)$$

where $C_*^t(x_i)$ is a cluster label (corresponding to a particular cluster of the clustering π_t) to which the sample x_i belongs. In addition, $\mathrm{sim}(C_x, C_y) \in [0, 1]$ denotes the similarity between any two clusters C_x, C_y, which can be discovered using the following link-based algorithm. Note that, for any clustering $\pi_t \in \Pi$, $1 \leq \sum_{\forall C \in \pi_t} \mathrm{RM}(x_i, C) \leq k_t$. Unlike the measure of fuzzy membership, the typical constraint of $\sum_{\forall C \in \pi_t} \mathrm{RM}(x_i, C) = 1$ is

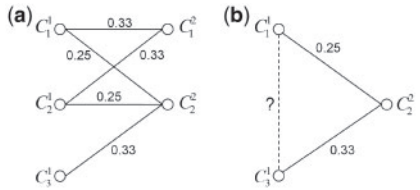


Fig. 3. Examples of (**a**) cluster network and (**b**) connected-triple between vertices C_1^1 and C_3^1, where $w_{C_1^1 C_3^1}=0$.

not appropriate for re-scaling associations within the RM. In fact, such local normalization will significantly distort the true semantics of known associations ('1'), such that their magnitudes become dissimilar, different from one clustering to another. According to our empirical investigation, the quality of RM is usually higher than other soft, fuzzy-like variations of the BM, which can be obtained from sample-to-cluster distances or a fuzzy cluster ensemble. See Supplementary Section 8.1 for details of such methods and associated experimental results.

2.1.1 WCT: a link-based similarity algorithm Given a cluster ensemble Π of data samples X, a weighted graph $G=(V,W)$ can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters. Formally, the weight assigned to the edge $w_{xy}\in W$, that connects clusters $C_x, C_y \in V$, is estimated by

$$w_{xy}=\frac{|L_x \cap L_y|}{|L_x \cup L_y|} \tag{4}$$

where $L_z \subset X$ denotes the set of samples belonging to cluster $C_z \in V$. Figure 3a shows the network of clusters that is generated from the example given in Figure 1. Note that circle nodes represent clusters and edges exist only when the corresponding weights are non-zero.

Given this network formalism, the new WCT algorithm is introduced to disclose the similarity between any pair of clusters. It extends the Connected-Triple method (Reuther and Walter, 2006) that has been originally developed to identify ambiguous author names within publication databases. In particular, the similarity of any $C_x, C_y \in V$ can be estimated by counting the number of Connected-Triples (i.e. triples) they are part of. Formally, a triple, Triple = $(V_{\text{Triple}}, W_{\text{Triple}})$, is a subgraph of G containing three vertices $V_{\text{Triple}}=\{C_x, C_y, C_k\} \subset V$ and two non-zero edges $W_{\text{Triple}}=\{w_{xk}, w_{yk}\} \subset W$, with $w_{xy}=0$. An example of triple within the network of Figure 3a is shown in Figure 3b.

This simple counting might be sufficient for any indivisible object, e.g. name or sample. 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 members). Inspired by this idea, the WCT measure of clusters $C_x, C_y \in V$ with respect to each triple $C_k \in V$, is estimated as

$$\text{WCT}_{xy}^k=\min(w_{xk}, w_{yk}) \tag{5}$$

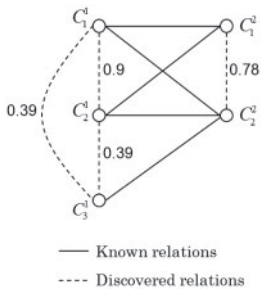
where $w_{xk}, w_{yk} \in W$ are weights of the edges connecting clusters C_x and C_k, and clusters C_y and C_k, respectively. The count of all triples $(1 \ldots q)$ between clusters C_x and C_y can be calculated as follows:

$$\text{WCT}_{xy}=\sum_{k=1}^{q}\text{WCT}_{xy}^k \tag{6}$$

Then, the similarity between clusters C_x and C_y can be estimated by

$$\text{sim}(C_x, C_y)=\frac{\text{WCT}_{xy}}{\text{WCT}_{\max}} \times \text{DC} \tag{7}$$

where $\text{WCT}_{\max}$ is the maximum WCT_{pq} value of any two clusters $C_p, C_q \in V$ and $\text{DC} \in (0,1)$ is a constant decay factor (i.e. confidence level of accepting two non-identical clusters as being similar). Following the example shown in Figs 1 and 3, the discovered link-based similarities/relations and the resulting RM are presented in Figure 4.



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

RM matrix

Fig. 4. Details of disclosed WCT similarities/relations with DC being 0.9, and the resulting RM.

2.2 Applying a consensus function to RM

Having obtained a refined cluster-association matrix (RM) with the aforementioned link-based similarity algorithm, a graph-based partitioning method is exploited to obtain the final clustering. This consensus function requires the underlying matrix to be initially transformed into a weighted bipartite graph. Formally, given an RM representing associations between N samples and P clusters in an ensemble Π, a weighted bipartite graph $G=(V,W)$ can be constructed, 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 that can be defined as follows:

- $w_{ij}=0$ when vertices $v_i, v_j \in V^X$, i.e. correspond to samples.
- $w_{ij}=0$ when vertices $v_i, v_j \in V^C$, i.e. correspond to clusters.
- $w_{ij}=RM(v_i, v_j)$ when vertices $v_i \in V^X$ and $v_j \in V^C$. The bipartite graph G is bi-directional such that w_{ij} is equivalent to w_{ji}.

Given such graph, the spectral graph partitioning (SPEC) method similar to that of Ng *et al.* (2002) is applied to generate a final data partition. This is a powerful method for decomposing an undirected graph, with good performance being exhibited in many application areas, including protein modelling, information retrieval and identification of densely connected on-line hypertextual regions (Luxburg, 2007). Principally, given a graph $G=(V,W)$, SPEC first finds the K largest eigenvectors $u_1, \ldots, u_K$ of W, which are used to formed another matrix U (i.e. $U=[u_1, \ldots, u_K]$), whose rows are then normalized to have unit length. By considering the row of U as K-dimensional embedding of the graph vertices, SPEC applies k-means to these embedded points in order to acquire the final clustering result. Further details of SPEC can be found in Supplementary Section 1.

2.3 Experiment design

The experiments set out to investigate the performance of LCE compared to a number of different simple/standard clustering algorithms and state-of-the-art cluster ensemble methods, over real gene expression and synthetic datasets. The compared techniques include: (i) HBGF that is the baseline model of LCE; (ii) four simple clustering techniques that are usually used by clinical researchers to analyse microarray data [k-means (KM), single-linkage (SL), complete-linkage (CL) and average-linkage (AL)]; (iii) three pairwise similarity-based cluster ensemble algorithms that have been developed so far for gene expression data analysis [MULTI-K, consensus clustering with hierarchical clustering (CC_{HC}) and graph-based consensus clustering (GCC)], and three graph-based cluster ensemble techniques that have been considered as benchmarks in the literature [Cluster-based Similarity Partitioning Algorithm (CSPA), Hyper-Graph Partitioning Algorithm (HGPA) and Meta-Clustering Algorithm (MCLA)]. Details of examined cluster ensemble techniques are given below.

- *Pairwise similarity-based cluster ensemble methods* are based principally on the pairwise similarity among samples. Given a cluster ensemble $\Pi=\{\pi_1, \ldots, \pi_M\}$ of a dataset $X=\{x_1, \ldots, x_N\}$, an $N \times N$

Table 1. Description of real gene expression datasets: tissue type, microarray chip type, number of samples (N), number of original genes (d^*), number of selected genes (d), number of classes (K) and class distribution

Dataset	Tissue	Chip	Samples (N)	Original genes (d^*)	Selected genes (d)	Classes (K)	Class distribution
Leukemia1 (Golub *et al.*, 1999)	Bone marrow	Affy	72	7129	1877	2	47, 25
Leukemia2 (Golub *et al.*, 1999)	Bone marrow	Affy	72	7129	1877	3	38, 9, 25
Leukemia3 (Armstrong *et al.*, 2002)	Blood	Affy	72	12582	2194	3	20, 24, 28
Breast-Colon tumors (Chowdary *et al.*, 2006)	Breast and colon	Affy	104	22283	182	2	62, 42
Brain Tumor (Nutt *et al.*, 2003)	Brain	Affy	50	12625	1377	4	14, 14, 7, 15
Central nervous system (Pomeroy *et al.*, 2002)	Brain	Affy	42	7129	1379	5	10, 8, 10, 10, 4
Multi-tissue1 (Ramaswamy *et al.*, 2001)	Multi-tissue	Affy	190	16063	1363	14	11, 10, 11, 11, 22, 11, 10, 10, 30, 11, 11, 11, 11, 20
Multi-tissue2 (Su *et al.*, 2001)	Multi-tissue	Affy	174	12533	1571	10	26, 8, 26, 23, 12, 11, 7, 27, 6, 28
Hepatocellular carcinoma (Chen *et al.*, 2002)	Liver	cDNA	180	22699	85	2	104, 76
Small, round blue-cell tumors (Khan *et al.*, 2001)	Multi-tissue	cDNA	83	6567	1069	4	29, 11, 18, 25

See Supplementary Material for further details.

similarity matrix (CO) is constructed as $CO(x_i,x_j)=\frac{1}{M}\sum_{m=1}^{M}S_m(x_i,x_j)$, where $\mathrm{CO}(x_i,x_j)\in[0,1]$ represents the similarity measure between samples $x_i,x_j\in X$. In addition, $S_m(x_i,x_j)=1$ if $C^m(x_i)=C^m(x_j)$, and $S_m(x_i,x_j)=0$ otherwise. Note that $C^m(x_i)$ denotes the cluster label of the m-th clustering to which a sample $x_i\in X$ belongs.

Since co-association matrix (CO) is a similarity matrix, any similarity-based clustering algorithm (referred to as 'consensus function') can be applied to this matrix to yield the final partition π^* (Fred and Jain, 2005). Among several existing similarity-based techniques, the most well-known is agglomerative hierarchical clustering algorithm. Specifically to the problem of clustering cancer samples, *MULTI-K* (Kim *et al.*, 2009) and CC_{HC} (Monti *et al.*, 2003) methods make use the SL and AL agglomerative hierarchical clusterings as consensus functions, respectively. In addition, to obtain π^*, the *GCC* approach (Yu *et al.*, 2007) transforms the CO matrix into a graph of samples to which the normalized cut algorithm (Shi and Malik, 2000) is applied.

- *Graph-based cluster ensemble algorithms* that are investigated herein include the methods of Fern and Brodley (2004) (HBGF) and Strehl and Ghosh (2002) (CSPA, HGPA and MCLA). Note that HBGF is included as the baseline model of LCE. It makes use of the bipartite graph that is generated from the BM. There is no edge connecting vertices of the same object type, and the weight of an edge between any data point and cluster is either 1 (when the sample belongs to the cluster) or 0 (otherwise). SPEC (Ng *et al.*, 2002) is exploited to obtain the final clustering result from this graph. This effectively allows the quality of the two cluster-association matrices (i.e. BM and RM) to be compared. CSPA creates a similarity graph, where vertices represent samples and edges' weight represent similarity scores obtained from the CO matrix. Afterwards, a graph partitioning algorithm called METIS (Karypis and Kumar, 1998) is used to partition the similarity graph into K clusters. HGPA constructs a hyper-graph, where vertices represent samples and the same-weighted hyper-edges represent clusters in the ensemble. Then, HMETIS (Karypis *et al.*, 1999) is applied to partition the underlying hyper-graph into K parts. MCLA creates a graph where each vertex corresponds to each cluster in the ensemble and each edge's weight between any two cluster vertices is computed using the binary Jaccard measure. METIS is also employed to partition the meta-level graph into K meta-clusters. The final clustering is produced by assigning each sample to the meta-cluster with which it is most frequently associated.

Note that the performance of SL, CL, AL and KM are always assessed over the original data, without using any information of cluster ensemble. The effectiveness of LCE and other cluster ensemble methods with different ensemble sizes and types are also empirically examined. Details of gene expression datasets and experiment setting are presented below.

2.3.1 Real gene expression datasets This evaluation is based on real gene expression data, obtained from nine published microarray studies, and summarized in Table 1. The experiments were conducted over filtered datasets as given in the empirical study of de Souto *et al.* (2008), where uninformative genes are removed for a better quality of clustering result. Details of the types of datasets, data preprocessing and the gene selection method are given in Supplementary Sections 2.1–2.2. To rigourously evaluate the robustness of LCE and its compared techniques, they are also assessed on both simulated gene expression data (with noise and imbalanced clusters) and geometrically complicated datasets (see Supplementary Sections 3–4 for data descriptions).

2.3.2 Experiment setting The proposed LCE method and its competitors are evaluated, using the experiment setting illustrated below.

- Each cluster ensemble method is evaluated over four different types of ensemble: (i) Fixed-k with full-space data (with d genes), (ii) Fixed-k with subspace data (with q genes), (iii) Random-k with full-space data and (iv) Random-k with subspace data, respectively.
- An ensemble size (M) of only 10 base clusterings was used.
- To generate a refined cluster-association matrix (RM), the constant decay factor (DC) of 0.9 is exploited with the underlying link-based similarity algorithm (i.e. WCT).
- For a comparison purpose, as in Fern and Brodley (2004) and Kim *et al.* (2009), each clustering method divides data points into a partition of K (the number of *true classes* for each dataset, known as 'gold standard') clusters, which is then evaluated against the corresponding true partition using a set of well-known evaluation indices. Note that, true classes are known for all datasets but are absolutely not used in any way by the cluster ensemble process; they are only used to evaluate the quality of the clustering results after clustering is complete. This assessment framework has been successfully adopted in de Souto *et al.* (2008) to compare the performance of different simple clustering algorithms over a large number of gene expression datasets.
- The current research follows several previous studies (Kim *et al.*, 2009; Monti *et al.*, 2003; Yu *et al.*, 2007) that focus on clustering samples of a given microarray data into known groups, i.e. class prediction. In particular to these methods, the quality of data partition π^* generated by a clustering technique is directly compared against the

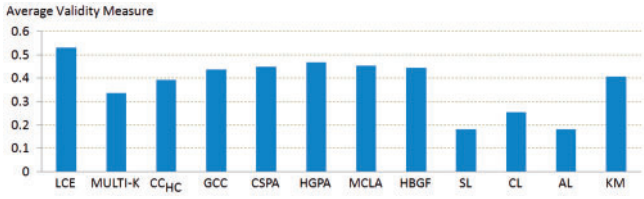


Fig. 5. Average validity measures of different clustering methods, across all validity indices (CA, NMI, AR) and experimental settings.

known partition Π' (i.e. class labels), using external validity indices such as Adjusted Rand (AR; Hubert and Arabie, 1985), Normalized Mutual Information (NMI; Strehl and Ghosh, 2002) and Classification Accuracy (CA; Nguyen and Caruana, 2007). These specific indices are exploited for evaluating the performance of the proposed LCE method, against several other clustering techniques.

The limitation of this evaluation is that the capability of examined methods for 'class discovery' has not been reviewed. Unlike the task of class prediction, the quality of data partition is determined by a structural properties of clusters, e.g. a compactness of samples in a cluster and a distance between clusters. To this extent, an initial study regarding of LCE for the task of class discovery is provided in Supplementary Section 8.2. In addition, the analysis of gene domain is another prominent research, in which LCE may prove to be useful. In particular, a better quality assessment should make use of a validity index that takes into account known gene functions, instead of simple external or internal indices mentioned earlier. Here, the performance of a given clustering algorithm is justified in terms of its ability to produce biologically meaningful clusters using a reference set of functional classes, which can be obtained from prior biological knowledge specific to a microarray study or may be formed using the growing databases of Gene Ontologies.

- The quality of each cluster ensemble method with respect to a specific ensemble setting is generalized as the average of 50 runs.

3 RESULTS

The results[1] with real gene expression data are summarized in Figure 5, where each investigated clustering method is represented with its average validity measure across all validity indices, datasets and ensemble types. It is clear that LCE regularly performs better than any of these clustering methods. It also enhances the performance of KM, which is used as base clusterings. In particular, HBGF is apparently less effective than LCE. This information suggests that the quality of the refined cluster-association matrix (RM) is superior than the original BM counterpart. See the full results with real gene expression datasets in Supplementary Section 2.3.

Following the study of Kuncheva *et al.* (2006), to rigourously evaluate the quality of investigated clustering techniques, the number of times that one method is significantly *better* and *worse* (to 95% confidence level) than the others are assessed across all experiment settings. Let $\overline{X}_C(i,\beta)$ be the average value of validity index $C \in \{\text{CA, NMI, AR}\}$ across n runs ($n=50$ in this evaluation) for a cluster ensemble method $i \in \text{CM}$ (CM is a set of 12 experimented clustering methods), on a specific experiment

[1]This section only contains a summary of our empirical evaluation over real gene expression data. For more detailed results and experiments with other data collections, please see Supplementary Sections 2-4.

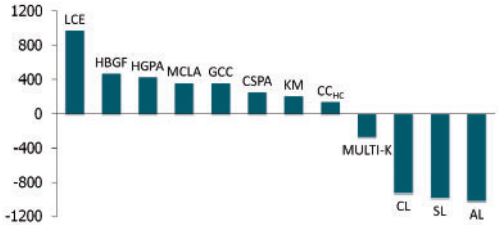


Fig. 6. The statistics of total performance, summarized across all evaluation indices, i.e. $(B-W)_i, \forall i \in \text{CM}$.

setting $\beta \in \text{ST}$ (ST is a set of 40 unique combination of four ensemble types and ten real gene expression datasets). The 95% CI, $[L_{\overline{X}_C(i,\beta)}, U_{\overline{X}_C(i,\beta)}]$, for the mean $\overline{X}_C(i,\beta)$ of each validity criterion C is calculated by $L_{\overline{X}_C(i,\beta)} = \overline{X}_C(i,\beta) - 1.96\frac{S_C(i,\beta)}{\sqrt{n}}$ and $U_{\overline{X}_C(i,\beta)} = \overline{X}_C(i,\beta) + 1.96\frac{S_C(i,\beta)}{\sqrt{n}}$. Note that $S_C(i,\beta)$ denotes the SD of the validity index C across n runs for a clustering method i and an experiment setting β. In addition, multiple runs of any setting $\beta \in \text{ST}$ are different and independent—each with a unique ensemble that is generated by randomly selected parameters, and possibly dissimilar gene subsets.

The number of times that one method $i \in \text{CM}$ is significantly *better* than its competitors, $B_C(i)$ (in accordance with the validity criterion C, across all experiment settings), can be defined as

$$B_C(i) = \sum_{\forall \beta \in \text{ST}} \sum_{\forall i^* \in \text{CM}, i^* \neq i} \text{better}_C^{\beta}(i, i^*) \tag{8}$$

$$\text{better}_C^{\beta}(i, i^*) = \begin{cases} 1 & \text{if } L_{\overline{X}_C(i,\beta)} > U_{\overline{X}_C(i^*,\beta)} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Similarly, the number of times that one method $i \in \text{CM}$ is significantly *worse* than its competitors, $W_C(i)$, in accordance with the validity criterion C, can be computed as

$$W_C(i) = \sum_{\forall \beta \in \text{ST}} \sum_{\forall i^* \in \text{CM}, i^* \neq i} \text{worse}_C^{\beta}(i, i^*) \tag{10}$$

$$\text{worse}_C^{\beta}(i, i^*) = \begin{cases} 1 & \text{if } U_{\overline{X}_C(i,\beta)} < L_{\overline{X}_C(i^*,\beta)} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

Using the aforementioned assessment formalism, Figure 6 illustrates for each method $i \in CM$ the statistics of total performance $(B-W)_i = \sum_{\forall C \in \{\text{CA,NMI,AR}\}} B_C(i) - W_C(i)$. The results shown in this figure indicate that LCE is more effective than other clustering techniques included in this experiment. Since SL and AL do not perform well over the examined datasets, MULTI-K and CC_{HC} that use the former and latter as a consensus function, respectively, are less accurate than other cluster ensemble methods and KM. However, their performance may improve with an ensemble that is much larger than the one investigated herein (i.e. $M >> 10$). In addition to this evaluation scheme, a further performance analysis with a paired t-test is discussed in Supplementary Section 2.4.

Another important investigation is on the subject of relations between performance of experimented cluster ensemble methods and different types of ensemble being explored in the present evaluation. Figure 7 shows the average validity measures of different cluster ensemble methods across all validity indices and real gene expression datasets. For each method, its performance with

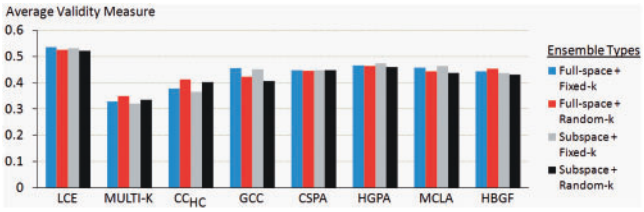


Fig. 7. Average validity measures of different cluster ensemble methods across all validity indices and real gene expression datasets, categorized in accordance with four types of ensemble.

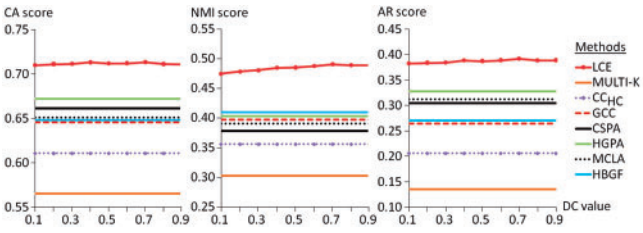


Fig. 8. The relations between $DC \in \{0.1, 0.2, \ldots, 0.9\}$ and the performance of the LCE method (the averages of CA, NMI and AR over 10 real gene expression datasets and 4 ensemble types), whose values are presented in X-axis and Y-axis, respectively. Note that the averaged performance of other cluster ensemble methods are also included for a comparison purpose.

four ensemble types (Full-space + Fixed-k, Full-space + Random-k, Subspace + Fixed-k, and Subspace + Random-k) are compared. It is clear that LCE is more effective than other cluster ensemble techniques over all ensemble types, with its best performance being generated from a 'Full-space + Fixed-k' ensemble. Most methods work better with Full-space ensembles, as compared to Subspace alternatives. Unlike LCE, GCC and other graph-based techniques that usually produce a superior performance with a Fixed-k ensemble type, MULTI-K and CC_{HC} are best when coupled with Random-k ensembles.

Although the results are impressive, on several datasets, it is important to ensure they are obtainable in a wide range of conditions. To this end the LCE algorithm's response was examined to perturbations in its parameters, and by investigating its time and space complexity.

3.1 Parameter analysis

The parameter that has any effect on the results of LCE is DC [see Equation (7)]. With the ensemble size of 10, we varied this value from 0.1 through 0.9, in steps of 0.1, for three validity measures, and obtained the results in Figure 8. This figure clearly shows that the results are robust, and do not depend strongly on any particulary value of DC. This makes it easy for users to obtain high-quality, reliable results when using LCE, particularly since values of DC near 0.7 generally produce the best results. Although there is variation in response across the DC values, the performance of LCE is always better than any of the other cluster ensemble methods included in this assessment.

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.

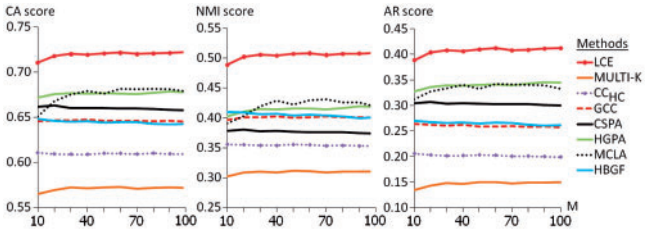


Fig. 9. Performance of different cluster ensemble methods in accordance with ensemble size ($M \in \{10, 20, \ldots, 100\}$), as the averages of validity measures (CA, NMI and AR) across all real gene expression datasets and ensemble types.

According to Figure 9 in which DC = 0.9, this heuristic is applicable to LCE, where its validity measures (averages of CA, NMI and AR across all experimental settings) gradually incline to the increasing value of $M \in \{10, 20, \ldots, 100\}$. Furthermore, LCE performs better than its competitors with all different ensemble sizes. Note that a bigger ensemble leads to an improved accuracy, but with the tradeoff of run time—but, again, even the worst results for LCE are better than the best results of the other methods. These findings regarding the relation between LCE and its parameters have also been observed when both DC and M are simultaneously analysed (see details in Supplementary Section 7).

3.2 Complexity analysis

The space and time complexity of creating a refined cluster-association matrix (RM) are $O(P^2 + \text{NP})$ and $O(P^2 l + \text{NP})$, where N is the number of samples, 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 neighbors of C_x (or C_y) to identify connected triples. Following this, the RM of size $N \times P$ is created using the aforementioned similarity matrix. As a result, LCE is computationally efficient with the time complexity generally converging to $O(N)$. Please consult Supplementary Section 6 for details of the scalability test.

3.3 Additional utilization of RM with simple clusterings

Besides its current utilization through the formation of a weighted bipartite graph, the RM can also be regarded as a 'high-level' data matrix to which any simple clustering algorithm can be directly applied. Promising results have been obtained from the exploitation of six simple clustering techniques with RM: RM + SL, RM + CL, RM + AL, RM + KM, RM + Partitioning Around Medoids and RM + spectral clustering, respectively (see detailed results in Supplementary Section 5).

4 CONCLUSION

A new LCE method has been introduced for clustering gene expression data samples that has greatly improved accuracy and efficiency. The performance of LCE is usually superior than existing graph-based ensemble techniques, and those that are particularly developed for gene data analysis. LCE is highly effective over real gene expression datasets and synthetic data collections (with

the presence of noise and non-equal-size clusters). Unlike existing pairwise similarity based counterparts, LCE is efficient for clustering large-size datasets, including the clustering of genes. Specifically, the refined cluster-association matrix (RM) used by LCE is able to recover and account for unknown entries in the original BM counterpart, and hence, delivers a superior clustering performance. With its consistent performance over settings of parameter, ensemble type and size, LCE also proves to be a user-friendly data analysis tool, especially for non-expert users.

Funding: Scholarship of the Ministry of Science and Technology, Royal Thai Government (to N.I.-O.).

Conflict of Interest: none declared.

REFERENCES

Armstrong,S. *et al*. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet*., **30**, 41–47.

Avogadri,R. and Valentini,G. (2009) Fuzzy ensemble clustering based on random projections for DNA microarray data analysis. *Artif. Intell. Med.*, **45**, 173–183.

Bredel,M. *et al*. (2005) Functional network analysis reveals extended gliomagenesis pathway maps and three novel MYC-interacting genes in human gliomas. *Cancer Res.*, **65**, 8679–8689.

Brunet,J. *et al*. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Chen,X. *et al*. (2002) Gene expression patterns in human liver cancers. *Mol. Biol. Cell.*, **13**, 1929–1939.

Chowdary,D. *et al*. (2006) Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *J. Mol. Diagn.*, **8**, 31–39.

de Souto,M. *et al*. (2008) Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics*, **9**, 497.

Duda,R.O. *et al*. (2000). *Pattern Classification*. 2nd edn. Wiley-Interscience, New York.

Dudoit,S. and Fridyand,J. (2003) Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, **19**, 1090–1099.

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

Fred,A.L.N. and Jain,A.K. (2005) Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 835–850.

Golub,T. *et al*. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hadjitodorov,S.T. *et al*. (2006) Moderate diversity for better cluster ensembles. *Inform. Fusion*, **7**, 264–275.

Handl,J. *et al*. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Iam-on,N. *et al*. (2008) Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. In *Proceedings of Eleventh International Conference on Discovery Science*, Springer, Budapest, Hungary, pp. 222–233.

Karypis,G. and Kumar,V. (1998) Multilevel k-way partitioning scheme for irregular graphs. *J. Parallel Distrib. Comput.*, **48**, 96–129.

Karypis,G. *et al*. (1999) Multilevel hypergraph partitioning: applications in VLSI domain. *IEEE Trans. VLSI Syst.*, **7**, 69–79.

Khan,J. *et al*. (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 673–679.

Kim,E. *et al*. (2009) MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, **10**, 260.

Kuncheva,L.I. and Hadjitodorov,S.T. (2004) Using diversity in cluster ensembles. In *Proceedings of the IEEE International Conference on Systems, Man & Cybernetics*, IEEE, The Hague, The Netherlands, pp. 1214–1219.

Kuncheva,L.I. and Vetrov,D. (2006) Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **28**, 1798–1808.

Kuncheva,L.I. *et al*. (2006) Experimental comparison of cluster ensemble methods. In *Proceedings of International Conference on Fusion*, International Society of Information Fusion, Florence, Italy, pp. 105–115.

Liu,L. *et al*. (2003) Robust singular value decomposition analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 13167–13172.

Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.

McLachlan,G. *et al*. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Monti,S. *et al*. (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Ng,A. *et al*. (2002) On spectral clustering: analysis and an algorithm. *NIPS*, **14**, 849–856.

Nutt,C. *et al*. (2003) Gene expressionbased classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res*., **63**, 1602–1607.

Pomeroy,S. *et al*. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Ramaswamy,S. *et al*. (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Reuther,P. and Walter,B. (2006) Survey on test collections and techniques for personal name matching. *Int. J. Metadata Semantics Ontologies*, **1**, 89–99.

Shi,J. and Malik,J. (2000) Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, **22**, 888–905.

Sorlie,T. *et al*. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA*, **100**, 8418–8423.

Strehl,A. and Ghosh,J. (2002) Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, **3**, 583–617.

Su,A. *et al*. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.*, **61**, 7388–7393.

Swift,S. *et al*. (2004) Consensus clustering and functional interpretation of gene-expression data. *Genome Biol.*, **5**, R94.

Yu,Z. *et al*. (2007) Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, **23**, 2888–2896.