Spectral Ensemble Clustering via Weighted K-Means: Theoretical and Practical Evidence

Hongfu Liu, Junjie Wu, Tongliang Liu, Dacheng Tao, *Fellow, IEEE*, and Yun Fu, *Senior Member, IEEE*

Abstract—As a promising way for heterogeneous data analytics, consensus clustering has attracted increasing attention in recent decades. Among various excellent solutions, the co-association matrix based methods form a landmark, which redefines consensus clustering as a graph partition problem. Nevertheless, the relatively high time and space complexities preclude it from wide real-life applications. We, therefore, propose Spectral Ensemble Clustering (SEC) to leverage the advantages of co-association matrix in information integration but run more efficiently. We disclose the theoretical equivalence between SEC and weighted K-means clustering, which dramatically reduces the algorithmic complexity. We also derive the latent consensus function of SEC, which to our best knowledge is the first to bridge co-association matrix based methods to the methods with explicit global objective functions. Further, we prove in theory that SEC holds the robustness, generalizability, and convergence properties. We finally extend SEC to meet the challenge arising from incomplete basic partitions, based on which a row-segmentation scheme for big data clustering is proposed. Experiments on various real-world data sets in both ensemble and multi-view clustering scenarios demonstrate the superiority of SEC to some state-of-the-art methods. In particular, SEC seems to be a promising candidate for big data clustering.

Index Terms—Consensus clustering, ensemble clustering, spectral clustering, co-association matrix, weighted K-means

1 Introduction

Consensus clustering, also known as *ensemble clustering*, emerges as a promising way for multi-source, heterogeneous data clustering, and recently attracts increasing academic attention. It aims to find a single partition that mostly agrees with multiple existing basic ones [1]. It is of recognized benefits in generating robust partitions, discovering novel structures, handling noisy features, and integrating solutions from multiple sources [2].

Generally speaking, consensus clustering can be roughly divided into two categories, i.e., those with or without explicit global objective functions. Methods that do not set objective functions make use of some heuristics or meta-heuristics to find approximate solutions. Representative methods include co-association matrix-based methods [3], [4], [5], [6], graph-based algorithms [1], [7], relabeling and voting methods [8], [9] and locally adaptive cluster-based methods [10]. Methods with explicit objectives employ global objective functions to measure the similarity between basic partitions and the consensus one. Representative solutions with different objective functions include K-means-like algorithm [11], [12], NMF [13], EM algorithm [14], simulated annealing [15] and combination regularization [16].

Of the above-mentioned methods, the co-association matrix-based methods form a landmark, where a co-association matrix is constructed to summarize basic partitions via measuring how many times a pair of instances occur simultaneously in a same cluster. The main contribution of these methods is the redefinition of the consensus clustering problem as a classical graph partition problem on the co-association matrix, so that agglomerative hierarchical clustering, spectral clustering, or other algorithms can be employed directly to find the consensus partition. It has been well informed that the co-association matrix-based methods can achieve excellent performances [5], [6], but they also suffer from some non-ignorable drawbacks. Particularly, the high time and space complexities prevent it from handling real-life large-scale data, and no explicit global objective function to guide consensus learning might lead to consensus partitions of unstable qualities when facing data sets of different characteristics.

In light of this, we propose Spectral Ensemble Clustering (SEC), which conducts spectral clustering on the co-association matrix to find the consensus partition. Our main contributions are summarized as follows. First, we formally prove that the spectral clustering of a co-association matrix is equivalent to the weighted K-means clustering of a binary matrix, which decreases the time and space complexities of SEC dramatically to roughly linear ones. Second, we derive the intrinsic consensus objective for SEC, which to our best knowledge is the first to give explicit global objective function to a co-association matrix based method, and thus could give clues to its theoretical foundation. Third, we prove theoretically the fine properties of SEC, including its robustness,

- *H. Liu and Y. Fu are with the Department of Electrical & Computer Engineering, Northeastern University, Boston, MA 02115. E-mail: liu.hongf@husky.neu.edu, yunfu@ece.neu.edu.*
- *J. Wu is with the School of Economics and Management, Beihang University, Beijing Shi 100191, China. E-mail: wujj@buaa.edu.cn.*
- *T. Liu is with the Centre for Artificial Intelligence, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia. E-mail: tliang.liu@gmail.com.*
- *D. Tao is with the School of Information Technologies and the Faculty of Engineering and Information Technologies, University of Sydney, J12/318 Cleveland St, Darlington, NSW 2008, Australia. E-mail: dacheng.tao@sydney.edu.au.*

Manuscript received 30 Mar. 2016; revised 19 Dec. 2016; accepted 3 Jan. 2017. Date of publication 9 Jan. 2017; date of current version 30 Mar. 2017. Recommended for acceptance by J. Gama.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2650229

generalizability and convergence, which are further verified empirically by extensive experiments. Fourth, we extend SEC so as to adapt to incomplete basic partitions (IBP), which enables a row-segmentation scheme suitable for big data clustering. Experimental results on various real-world data sets in both ensemble and multi-view clustering scenarios demonstrate that SEC outperforms some state-of-the-art baselines by delivering higher quality consensus partitions in an efficient way. Besides, SEC is very robust to incomplete basic partitions with many missing values. Finally, the promising ability of SEC in big data clustering is validated via a whole day collection of Weibo data.

This paper is an extension of our conference paper [17]. The new contents include (1) two more theorems (i.e., Theorems 5 and 8) on the convergence of SEC with either complete or incomplete basic partitions, (2) newly designed experiments to validate the robustness, generalizability and convergence of SEC and its ability for multi-view clustering, (3) substantially more competitive methods involved to demonstrate the effectiveness and efficiency of SEC, and (4) more comprehensive related work illustrated in terms of both ensemble clustering and multi-view clustering.

2 Related Work

In this section, we briefly introduce existing work most related to our study in this paper.

2.1 Ensemble Clustering

Ensemble clustering aims to fuse various existing basic partitions into a consensus one, which can be divided into two categories: with or without explicit global objective functions. In a global objective function, usually a utility function is employed to measure the similarity between a basic partition and the consensus one at the partition level. Then the consensus partition is achieved by maximizing the summarized utility function. In the inspiring work, Ref. [11] proposed a Quadratic Mutual Information based objective function for consensus clustering, and used K-means clustering to find the solution. Further, they used the expectation-maximization algorithm with a finite mixture of multinomial distributions for consensus clustering [14]. Wu et al. put forward a theoretic framework for K-means-based Consensus Clustering (KCC), and gave the sufficient and necessary condition for KCC utility functions that can be maximized via a K-means-like iterative process [12], [18], [19], [20]. In addition, there are some other interesting objective functions for consensus clustering, such as the ones based on nonnegative matrix factorization [13], kernel-based methods [21], simulated annealing [15], etc.

Another kind of methods do not set explicit global objective functions for consensus clustering. In one pioneer work, Ref. [1] (GCC) developed three graph-based algorithms for consensus clustering. More methods, however, employ co-association matrix to calculate how many times two instances jointly belong to the same cluster. By this means, some traditional graph partitioning methods can be called to find the consensus partition. Ref. [5] (HCC) is the most representative one in the link-based methods, which applied the agglomerative hierarchical clustering on the co-association matrix to find the consensus partition. Huang et al. employed the micro-cluster concept to summarize the basic partitions into a small core co-association matrix, and applied different partitioning methods, such as probability trajectory accumulation (PTA) and probability trajectory based graph partitioning (PTGP) [22], and graph partitioning with multi-granularity link analysis (MGLA) [23], for the final partition. Other methods include Relabeling and Voting [8], Robust Evidence Accumulation with weighting mechanism [24], Locally Adaptive Cluster based methods [10], Robust Spectral Ensemble Clustering [25] and Simultaneous Clustering and Ensemble [26], etc. There are still many other algorithms for ensemble clustering. Readers with interests can refer to some survey papers for more comprehensive understanding [27]. Most of the existing works focus on the process of the clustering on the (modified) co-assocaition matrix, while our method exactly decomposes the co-association matrix into a binary matrix and provides a high efficient solution for consensus clustering.

2.2 Multi-View Clustering

Multi-view clustering is to separate instances into different groups based on multiple representations. Existing multi-view clustering algorithms can be generally divided into four groups. The simplest way is to concatenate multi-view features directly and conduct traditional clustering via optimizing certain loss functions [28], [29], like ConKM employing K-means and ConNMF employing Non-negative matrix factorization. Another group of algorithms aim to find a common low-dimension latent subspace, which shares the most possible consistency between different views. Any clustering algorithms can then be applied to obtain the final partition [30], [31]. For instance, Ref. [32] proposed Collective NMF (ColNMF) to find a shared coefficient matrix with different basis matrices across views, and similarly, Ref. [33] leveraged $l_{2,1}$ norm to obtain a shared indicator matrix.

The third category generates basic partitions from each view individually and then calls consensus clustering to fuse them, which is very similar to consensus clustering mentioned previously. Recently, some methods interactively learn basic partitions and the consensus subspace by pushing basic clustering solution of each view towards the common one. In [34], the authors proposed Co-Regularized Spectral Clustering (CRSC), which enforces the similarity of eigenvectors learnt from different views and integrates it within the spectral clustering framework. Ref. [35] put forward a joint matrix factorization algorithm (MultiNMF), which incorporates both individual matrix factorizations and the inconsistency between each view's coefficient matrix and the consensus one. Recently, Ref. [36] extended ColNMF to handle partial multi-view clustering (PVC), and dealt with incomplete two-view clustering based on NMF. Other methods include local learning [37], pareto optimization [38], spectral embedding [39], etc.

2.3 Weighted K-Means

Weighted K-means is a variant of K-means [40] by assigning predefined weights to data instances. Similar to K-means, a two-phase iterative heuristic is employed for the partition. In the data assignment phase, each data instance is assigned to the nearest centroid. In the centroid update phase, each centroid is updated by the weighted arithmetic of instances. The two phases iteratively optimize the objective function until convergence.

X	π_1	π_2	π_3	π_4
x_1	1	2	1	1
x_2	1	2	1	1
x_3	1	2	2	1
x_4	2	3	2	1
x_5	2	3	2	2
x_6	3	1	3	2
x_7	3	1	3	2

⇨

X	π_1			π_2			π_3			π_4		$w_{b(x)}$
x_1	1	0	0	0	1	0	1	0	0	1	0	12
x_2	1	0	0	0	1	0	1	0	0	1	0	12
x_3	1	0	0	0	1	0	0	1	0	1	0	13
x_4	0	1	0	0	0	1	0	1	0	1	0	11
x_5	0	1	0	0	0	1	0	1	0	0	1	10
x_6	0	0	1	1	0	0	0	0	1	0	1	9
x_7	0	0	1	1	0	0	0	0	1	0	1	9

Fig. 1. Illustration of binary matrix and instance weights.

3 Spectral Ensemble Clustering

Let $X = \{x_1, \ldots, x_n\}^\top \in \mathbb{R}^{n\times d}$ represent the data matrix containing n instances in d dimensions. π_i is a crisp *basic partition* of X with K_i clusters generated by some traditional clustering algorithm, and $\pi_i(x) \in \{1, 2, \ldots, K_i\}$ represents the cluster label of instance x. Given r basic partitions of X in $\Pi = \{\pi_1, \pi_2, \ldots, \pi_r\}$, a *co-association matrix* $\mathbf{S}_{n\times n}$ is defined as follows [5]:

$$\mathbf{S}(x,y) = \sum_{i=1}^{r} \delta(\pi_i(x), \pi_i(y)), \quad \delta(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b. \end{cases}$$

In essence, the co-association matrix measures the similarity between each pair of instances, which is the co-occurrence counts of two instances in the same cluster in Π.

Spectral Ensemble Clustering applies spectral clustering on the co-association matrix $\mathbf{S}$ for the final consensus partition π, which is formulated as follows:

Let $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_K]$, a $n \times K$ partition matrix, be the 1-of-K coding of π, where K is the user-specified cluster number. The objective function of normalized-cut spectral clustering of $\mathbf{S}$ is the following trace maximization problem

$$\max_{\mathbf{Z}} \frac{1}{K} \mathrm{tr}(\mathbf{Z}^\top \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2} \mathbf{Z}), \text{ s.t. } \mathbf{Z}^\top \mathbf{Z} = \mathbf{I}, \tag{1}$$

where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{ll} = \sum_q \mathbf{S}_{lq}, 1 \leq l, q \leq n$, and $\mathbf{Z} = \mathbf{D}^{1/2}\mathbf{H}(\mathbf{H}^\top \mathbf{D}\mathbf{H})^{-1/2}$. A well-known solution to Eq. (1) is to run K-means on the top largest K eigenvectors of $\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ for the final consensus partition π [41], which consists of K cluster $C_1, C_2, \ldots, C_K$.

3.1 From SEC to Weighted K-Means

Performing spectral clustering on the co-association matrix, however, suffers from huge time complexity originated from both building the matrix and conducting the clustering. To meet this challenge, one feasible way is to find a more efficient yet equivalent solution for SEC. In what follows, we propose to solve SEC by a weighted K-means clustering on a binary matrix.

Let $\mathbf{B}_{n\times(\sum_{i=1}^{r} K_i)}$ be a binary matrix derived from the set of r basic partitions in Π as follows:

$$\begin{aligned} &\mathbf{B}(x,\cdot) = b(x) = \langle b(x)_1, \ldots, b(x)_r \rangle, \\ &b(x)_i = \langle b(x)_{i1}, \ldots, b(x)_{iK_i} \rangle, \\ &b(x)_{ij} = \begin{cases} 1, & \text{if } \pi_i(x) = j \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where $\langle\cdot\rangle$ indicates a transverse vector. Apparently, $|b(x)_i| = 1, \ \forall\, i$, where $|\cdot|$ is the L_1-norm. Fig. 1 shows an example of the binary matrix and the weight for each data instance. The binary matrix is just to concatenate all the 1-of-K_i codings of basic partitions. Based on $\mathbf{B}$, we provide the theorem to connect SEC and classical weighted K-means clustering, from which the calculation of the weights will be also given. Note that *all proofs of theorems will be given in the appendices for clarity*.

Theorem 1. *Given Π, the spectral clustering of $\mathbf{S}$ is equivalent to the weighted K-means clustering of $\mathbf{B}$; that is,*

$$\max \frac{1}{K} \mathrm{tr}(\mathbf{Z}^\top \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2} \mathbf{Z}) \Leftrightarrow \sum_{x\in X} f_{m_1,\ldots,m_K}(x),$$

where $f_{m_1,\ldots,m_K}(x) = \min_k w_{b(x)} \| \frac{b(x)}{w_{b(x)}} - m_k \|^2$, $m_k = \frac{\sum_{x\in C_k} b(x)}{\sum_{x\in C_k} w_{b(x)}}$, *and* $w_{b(x)} = \mathbf{D}(x,x) = \sum_{i=1}^{r} \sum_{l=1}^{n} \delta(\pi_i(x), \pi_i(x_l))$.

Remark 1. By Theorem 1, we explicitly transform SEC into a weighted K-means clustering in a theoretically equivalent way. Without considering the dimensionality, the time complexity of weighted K-means is roughly $\mathcal{O}(InrK)$, where I is the number of iterations. Thus, the transformation dramatically reduces the time and space complexities from $\mathcal{O}(n^3)$ and $\mathcal{O}(n^2)$, respectively, to roughly $\mathcal{O}(n)$. Note that there is only one non-zero element in $b(x)_i$. Accordingly, while the weighted K-means is conducted on $\mathbf{B}$, a $n \times \sum_{i=1}^{r} K_i$ binary matrix, the real dimensionality in computation is merely r.

Remark 2. In [42], the authors uncovered the connection between spectral clustering and weighted kernel K-means. Differently, for SEC we actually figure out the mapping function of the kernel, which turns out to be the binary data dividing its corresponding weight. By this means, we transform SEC into weighted K-means rather than weighted kernel K-means, which is crucial for gaining high efficiency for SEC and making it practically feasible.

Algorithm 1 gives the pseudocodes of SEC. It is worthy to note that in Line 2, $\sum_{l=1}^{n} \delta(\pi_i(x), \pi_i(x_l))$ calculates the size of the cluster where x belongs to in the ith basic partition. Moreover, the binary matrix $\mathbf{B}$ is highly sparse, with only r non-zero elements existing in each row. The weighted K-means is finally called for the solution.

Algorithm 1. Spectral Ensemble Clustering (SEC)

Input: $\Pi = \{\pi_1, \pi_2, \ldots, \pi_r\}$: r basic partitions.
K: the number of clusters.
Output: π: the consensus partition.
1: Build the binary matrix $\mathbf{B} = [b(x)]$ by Eq. (2);
2: Calculate the weight for each instance x by $w_{b(x)} = \sum_{i=1}^{r} \sum_{l=1}^{n} \delta(\pi_i(x), \pi_i(x_l))$;
3: Call weighted K-means on $\mathbf{B}' = [b(x)/w_{b(x)}]$ with the weight $w_{b(x)}$ and return the partition π;

3.2 Intrinsic Consensus Objective Function

By the transformation in Theorem 1, we give a new insight of the objective function of SEC. Here we derive the *intrinsic* consensus objective function of SEC to measure the similarity in the partition level. For two partitions π and π_i containing K and K_i clusters, respectively, let $n_{kj}^{(i)}$ denote the number of data instances belonging to both cluster $C_j^{(i)}$ in π_i

TABLE 1
Contingency Matrix

		π_i				
		$C_1^{(i)}$	$C_2^{(i)}$	$\cdots$	$C_{K_i}^{(i)}$	$\sum$
π	C_1	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$\cdots$	$n_{1K_i}^{(i)}$	n_{1+}
	C_2	$n_{21}^{(i)}$	$n_{22}^{(i)}$	$\cdots$	$n_{2K_i}^{(i)}$	n_{2+}
	$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$
	C_K	$n_{K1}^{(i)}$	$n_{K2}^{(i)}$	$\cdots$	$n_{KK_i}^{(i)}$	n_{K+}
	$\sum$	$n_{+1}^{(i)}$	$n_{+2}^{(i)}$	$\cdots$	$n_{+K_i}^{(i)}$	n

and cluster C_k in π, $n_{k+} = \sum_{j=1}^{K_i} n_{kj}^{(i)}$, and $n_{+j}^{(i)} = \sum_{k=1}^{K} n_{kj}^{(i)}$, $1 \le j \le K_i$, $1 \le k \le K$ (See Table 1). Let $p_{kj}^{(i)} = n_{kj}^{(i)}/n$, $p_{k+} = n_{k+}/n$, and $p_{+j}^{(i)} = n_{+j}^{(i)}/n$, then we can define a wide range of utility functions to measure the similarity between two partitions by Table 1. For example, the widely-used category utility function [43] is

$$U_c(\pi, \pi_i) = \sum_{k=1}^{K} p_{k+} \sum_{j=1}^{K_i} \left(\frac{p_{kj}^{(i)}}{p_{k+}}\right)^2 - \sum_{j=1}^{K_i} \left(p_{+j}^{(i)}\right)^2. \quad (3)$$

Theorem 2. *If a utility function takes the form as*

$$U(\pi, \pi_i) = \sum_{k=1}^{K} \frac{n_{k+}}{w_{C_k}} p_{k+} \sum_{j=1}^{K_i} \left(\frac{p_{kj}^{(i)}}{p_{k+}}\right)^2, \quad (4)$$

where $w_{C_k} = \sum_{x \in C_k} w_{b(x)}$, then it satisfies

$$\max_{\mathbf{Z}} \frac{1}{K} \mathrm{tr}(\mathbf{Z}^\top \mathbf{D}^{-1/2} \mathbf{S} \mathbf{D}^{-1/2} \mathbf{Z}) \Leftrightarrow \max_{\pi} \sum_{i=1}^{r} U(\pi, \pi_i). \quad (5)$$

Remark 3. The utility function U of SEC in Eq. (4) actually defines a family of utility functions to supervise the consensus learning process. Obviously, $f(U)$ also satisfies Eq. (5) if f is a strictly increasing function. Compared with the categorical utility function, the utility function U of SEC enforces the weights of the instances in large clusters in a quite natural way. Recall that the co-association matrix measures the similarity at the *instance* level; by Theorem 2, we derive the utility function to measure the similarity at the *partition* level. This indicates that two kinds of similarities at different levels are essentially inter-convertible, which to the best of our knowledge is the first claim in consensus clustering.

Remark 4. Theorem 2 gives a way of incorporating the weights of basic partitions into the ensemble learning process as follows:

$$\max_{\pi} \sum_{i=1}^{r} \mu_i U(\pi, \pi_i) \Leftrightarrow \sum_{x \in X} f_{m_1, \ldots, m_K}(x),$$

where $\boldsymbol{\mu}$ is the weight vector of basic partitions, $f_{m_1,\ldots,m_K}(x) = \min_k w_{b(x)} \sum_{i=1}^{r} \mu_i \| \frac{b(x)_i}{w_{b(x)}} - m_{k,i} \|^2$, and $m_{k,i} = \sum_{x \in C_k} b(x)_i / \sum_{x \in C_k} w_{b(x)}$. By this means, we can extend SEC to incorporate the weights of both instances and basic partitions in the ensemble learning process. In what follows, without loss of generality, we set $\mu_i = 1, \forall\, i$.

4 Theoretical Properties

Here, we analyze the learning ability of SEC by exploiting its robustness, generalizability and convergence in theory.

4.1 Robustness

Robustness that measures the tolerance of learning algorithms to perturbations (noise) is a fundamental property for learning algorithms. If a new instance is close to a training instance, a good learning algorithm should make their errors similar. This property of algorithms is formulated as robustness by the following definition [44].

Definition 1 (Robustness). *Let $\mathcal{X}$ be the training example space. An algorithm is $(K, \epsilon(\cdot))$ robust, for $K \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{X}^n \mapsto \mathbb{R}$, if $\mathcal{X}$ can be partitioned into K disjoint sets, denoted by $\{C_i\}_{i=1}^{K}$, such that the following holds for all $X \in \mathcal{X}^n, \forall x \in X, \forall x' \in \mathcal{X}, \forall i = 1, \ldots, K$: if $x, x' \in C_i$, then $|f_{m_1,\ldots,m_K}(x) - f_{m_1,\ldots,m_K}(x')| \le \epsilon(X)$.*

We then have Theorem 3 to measure the robustness of SEC as follows:

Theorem 3. *Let $\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_2)$ be a covering number of $\mathcal{X}$, which is defined to be the minimal integer $m \in \mathbb{N}$ such that there exist m disks with radius γ (measured by the metric $\|\cdot\|_2$) covering $\mathcal{X}$. For any $x, x' \in \mathcal{X}, \|x - x'\|_2 \le \gamma$, we define $\|b(x)_i - b(x')_i\|_2 \le \gamma_i$ and $|w_{b(x)_i} - w_{b(x')_i}| \le \gamma_{w,i}, i = 1, \ldots, r$, where $w_{b(x)_i} = \sum_{l=1}^{n} \delta(\pi_i(x), \pi_i(x_l))$. Then, for any centroids $m_1, \ldots, m_K$ learned by SEC, we obtain SEC is $(\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_2), \frac{2\sum_{i=1}^{r} \gamma_{w,i}}{r} + \sqrt{\frac{\sum_{i=1}^{r} \gamma_i^2}{r}})$-robust.*

Remark 5. From Theorem 3, we can see that even if $\{\gamma_i\}$ and $\{\gamma_{w,i}\}$ might be large due to some instances "poorly" clustered by some basic partitions, the high-quality performance of SEC will be preserved, provided that these instances are "well" clustered by other majorities. This means that SEC could benefit from the ensemble of basic partitions.

4.2 Generalizability

A small generalization error leads to a small gap between the expected reconstruction error of the learned partition and that of the target one [45]. The generalizability of SEC is highly dependent on the basic partitions. In what follows, we prove that the generalization bound of SEC can converge quickly and SEC can therefore achieve high-quality clustering with a relatively small number of instances.

Theorem 4. *Let π be the partition learned by SEC. For any independently distributed instances $x_1, \ldots, x_n$ and $\delta > 0$, with probability at least $1 - \delta$, the following holds:*

$$\begin{aligned} & E_x f_{m_1,\ldots,m_K}(x) - \frac{1}{n} \sum_{l=1}^{n} f_{m_1,\ldots,m_K}(x_l) \\ & \le \frac{\sqrt{2\pi} r K}{n} \left(\sum_{l=1}^{n} \left(w_{b(x_l)}\right)^{-2} \right)^{\frac{1}{2}} + \frac{\sqrt{8\pi} r K}{\sqrt{n} \min_{x \in X} w_{b(x)}} \\ & \quad + \frac{\sqrt{2\pi} r K}{n \min_{x \in X} \left(w_{b(x)}\right)^2} \left(\sum_{l=1}^{n} \left(w_{b(x_l)}\right)^2 \right)^{\frac{1}{2}} + \left(\frac{\ln(1/\delta)}{2n} \right)^{\frac{1}{2}}. \end{aligned} \quad (6)$$

Remark 6. Theorem 4 shows that if the third term of the upper bound goes to zero when n goes to infinity, the empirical reconstruction error of SEC will reach its expected reconstruction error. So, the convergence of

$$\frac{\sqrt{2\pi}rK}{n}\left(\sum_{l=1}^{n}\left(w_{b(x_l)}\right)^2\right)^{\frac{1}{2}}\frac{1}{\min_{x\in X}\left(w_{b(x)}\right)^2},$$

is a sufficient condition for the convergence of SEC. This sufficient condition is easily achieved by the consistency property of the basic partitions.

Remark 7. The consistency of crisp basic partitions will make $w_{b(x_l)}/|C_k|$ diverge little, where $|C_k|$ denotes the cardinality of the cluster containing x_l. If we further assume that $|C_k| = a_k n$, where $a_k \in (0,1)$, the convergence of SEC can be as fast as $\mathcal{O}(1/\sqrt{n^3})$. The fast convergence rate will result in the expected risk of the learned partition decreasing quickly to the expected risk of the target partition [46]. This verifies the efficiency of SEC. Compared with classical K-means clustering, the fastest known convergence rate is $\mathcal{O}(1/\sqrt{n})$ [46], [47].

4.3 Convergence

Due to the good convergence of weighted K-means, SEC will converge **w.r.t**. n. Here, we show that it will also converge **w.r.t**. r, the number of basic partitions, which means that the final clustering π will become more robust and stable as we keep increasing the number of basic partitions.

Theorem 5. *$\forall\ \lambda > 0$, there exists a clustering π_0 such that*

$$\lim_{r\to\infty}\Pr\{|\pi - \pi_0| \geq \lambda\} \to 0,$$

where π is the final consensus clustering output by SEC and $\Pr\{A\}$ denotes the probability of event A.

Remark 8. Theorem 5 implies that the centroids $m_1,\ldots,m_K$ will converge to $m_1^0,\ldots,m_K^0$ as r goes to infinity. Thus, the output of SEC will converge to the true clustering as we increase the number of basic partitions sufficiently.

5 Incomplete Evidence

In practice, *incomplete basic partitions* are easily met for data collecting device failures or transmission loss. By clustering a data subset $X_i \subseteq X$, $1 \leq i \leq r$, we can obtain an incomplete basic partition π_i of X. Assume r data subsets can cover the whole data set, i.e., $\bigcup_{i=1}^{r} X_i = X$ with $|X_i| = n^{(i)}$. The problem is how to cluster X into K crisp clusters using SEC given r IBPs in $\Pi = \{\pi_1,\ldots,\pi_r\}$.

Due to the missing values in Π, the co-association matrix cannot reflect the similarity of instance pairs any longer. To address this challenge, we start from the objective function of weighted K-means and extend it to handling incomplete basic partitions. It is obvious that missing elements in basic partitions provide no utility in the ensemble process. Consequently, they should not be involved in the weighted K-means for the centroid computation. We therefore have:

Theorem 6. *Given r incomplete basic partitions, we have*

$$\sum_{x\in X} f_{m_1,\ldots,m_K}(x) \Leftrightarrow \max \sum_{i=1}^{r} p^{(i)} \sum_{k=1}^{K} \frac{n_{k+}^{(i)}}{w_{C_k}^{(i)}} p_{k+} \sum_{j=1}^{K_i}\left(\frac{p_{kj}^{(i)}}{p_{k+}}\right)^2, \tag{7}$$

where $f_{m_1,\ldots,m_K}(x) = \min_k \sum_{i,x\in X_i} w_{b(x)} \| \frac{b(x)_i}{w_{b(x)}} - m_{k,i}\|^2$, *with* $m_{k,i} = \sum_{x\in C_k\cap X_i} b(x)_i / \sum_{x\in C_k\cap X_i} w_{b(x)}$, $p^{(i)} = n^{(i)}/n$, $n_{k+}^{(i)} = |C_k \cap X_i|$, $w_{C_k}^{(i)} = \sum_{x\in C_k\cap X_i} w_{b(x)_i}$.

Remark 9. Compared with Theorem 2, the utility function of SEC with IBPs has one more parameter $p^{(i)}$. This indicates that basic partitions with more elements are naturally assigned with higher importance for the ensemble process, which agrees with our intuition. This theorem also demonstrates the advantages of the transformation from co-association matrix to binary matrix; that is, the former cannot reflect the incompleteness of basic partitions while the latter can.

For the convergence of the SEC with IBPs, we have:

Theorem 7. *For the objective function in Eq. (7), SEC with IBPs is guaranteed to converge in finite two-phase iterations of weighted K-means clustering.*

Theorem 8. *SEC with IBPs holds the convergence property as the number of IBPs (r) increases.*

6 Towards Big Data Clustering

When it comes to big data, it is often difficult to conduct traditional cluster analysis due to the huge data volume and/or high data dimensionality. Ensemble clustering like SEC with the ability in handling incomplete basic partitions becomes a good candidate towards big data clustering.

In order to conduct large-scale data clustering, we propose the so-called *row-segmentation strategy*. Specifically, to generate each basic partition, we randomly select a data subset with a certain sampling ratio from the whole data, and run K-means on it to obtain an incomplete basic partition; this process repeats r times, prior to running SEC to obtain the final consensus partition.

The benefit of employing the row-segmentation strategy is two-fold. On one hand, a big data set can be decomposed into several smaller ones, which can be handled independently and separately to obtain IBPs. On the other hand, in the final consensus clustering, no matter how large the dimensionality of the original data is, we only need to conduct weighted K-means on the binary matrix **B** with only r non-zero elements in each row during the ensemble learning process. Note that Ref. [6] made the co-association matrix sparse for a fast decomposition, but we here transform the co-association matrix into the binary matrix directly so that we even do not need to build the co-association matrix. The experimental results in the next section demonstrate that the row-segmentation strategy does work well and even outperforms the basic clustering on the whole data.

7 Experimental Results

In this section, we evaluate SEC on abundant real-world data sets of different domains, and compare it with several state-

TABLE 2
Experimental Data Sets for Scenario I

Data set	Source	#Instances	#Features	#Classes
breast_w	UCI	699	9	2
iris	UCI	150	4	3
wine	UCI	178	13	3
cacmcisi	CLUTO	4,663	14,409	2
classic	CLUTO	7,094	41,681	4
cranmed	CLUTO	2,431	41,681	2
hitech	CLUTO	2,301	126,321	6
k1b	CLUTO	2,340	21,839	6
la12	CLUTO	6,279	31,472	6
mm	CLUTO	2,521	126,373	2
re1	CLUTO	1,657	3,758	25
reviews	CLUTO	4,069	126,373	5
sports	CLUTO	8,580	126,373	7
tr11	CLUTO	414	6,429	9
tr12	CLUTO	313	5,804	8
tr41	CLUTO	878	7,454	10
tr45	CLUTO	690	8,261	10
letter	LIBSVM	20,000	16	26
mnist	LIBSVM	70,000	784	10

of-the-art algorithms across both ensemble clustering and multi-view clustering areas. In the first scenario, each data set is provided with a single view and basic partitions are produced by some random sampling schemes. In the second scenario, however, each data set is provided with multiple views and each view generates either one or multiple basic partitions by random sampling. Finally, a case study on large-scale Weibo data shows the ability of SEC for big data clustering.

7.1 Scenario I: Ensemble Clustering

7.1.1 Experimental Setup

Data. Various real-world data sets with true cluster labels are used for evaluating the experiments in the scenario of ensemble clustering. Table 2 summarizes some important characteristics of these data sets obtained from UCI,[1] CLUTO,[2] and LIBSVM[3] repositories, respectively.

Tool. SEC is coded in MATLAB. The *kmeans* function in MATLAB with either squared Euclidean distance (for UCI and LIBSVM data sets) or cosine similarity (for CLUTO data sets) is run 100 times to obtain basic partitions by varying the cluster number in $[K, \sqrt{n}]$, where K is the true cluster number and n is the data size. For two relatively large data sets *letter* and *mnist*, the cluster numbers of basic partitions vary in $[2, 2K]$ for meaningful partitions. The baseline methods include consensus clustering with category utility function (CCC, a special case of KCC [12]), graph-based consensus clustering methods (GCC, including CSPA, HGPA and MCLA) [1], co-association matrix with agglomerative hierarchical clustering (HCC with group-average, single-linkage and complete-linkage) [5], and probability trajectory based graph partitioning (PTGP) [22]. These baselines are selected for the following reasons: GCC has great impacts in the area of consensus clustering; CCC shares common grounds with SEC by employing a K-means-like algorithm; both HCC and PTGP are co-association matrix based methods, and the former is a very famous one and the latter is newly proposed. All the methods are coded in MATLAB and set with default settings. The cluster number for SEC and all baselines is set to the true one for fair comparison. All basic partitions are equally weighted (i.e., $\boldsymbol{\mu} = \mathbf{1}$). Each algorithm runs 50 times for average results and deviations.

Validation. We employ external measures to assess cluster validity. It is reported that the normalized Rand index (R_n for short) is theoretically sound and shows excellent properties in practice [48], which therefore is adopted in our study. R_n is defined as follows:

$$R_n(\pi, \pi_i) = \frac{m - m_1 m_2 / M}{m_1/2 + m_2/2 - m_1 m_2 / M}, \tag{8}$$

where $m = \sum_{i,j} \binom{n_{ij}}{2}$, $m_1 = \sum_i \binom{n_{i+}}{2}$, $m_2 = \sum_j \binom{n_{+j}}{2}$, $M = \binom{n}{2}$. The definition of these count variables can be found in Table 1. A large R_n indicates a better clustering performance, whereas a negative R_n indicates a result even poorer than random labeling.

Environment. All experiments in Scenarios I&II were run on a PC with an Intel Core i7-3770 3.4 GHz*2 CPU and a 32 GB DDR3 RAM.

7.1.2 Validation of Effectiveness

Here, we compare the performance of SEC with that of baseline methods in consensus clustering. Table 3 (Left side) shows the clustering results, with the best results highlighted in *bold red* and the second best in *italic blue*.

First, it is obvious that SEC shows clear advantages over other consensus clustering baselines, with 10 best and 9 second best results out of the total 19 data sets; in particular, the margins for the three data sets: *wine*, *la12* and *mm* are very impressive. To fully compare the performance of different algorithms, we propose a measurement score as follows: $score(A_i) = \sum_j \frac{R_n(A_i, D_j)}{\max_i R_n(A_i, D_j)}$, where $R_n(A_i, D_j)$ denotes the R_n value of the A_i algorithm on the D_j data set. This score evaluates certain algorithm by the best performance achieved by the state-of-the-art methods. From this score, we can see that SEC exceeds other consensus clustering methods by a large margin.

Let us take a close look at HCC, which as SEC also leverages co-association matrix for consensus clustering. It is obvious that SEC outperforms HCC with group-average (HCC_GA) completely, in 13 out of 19 data sets, although HCC_GA is already the second best among the baselines. The implication is two-fold: First, the superior performances of SEC and HCC_GA indicate that the co-association matrix indeed does well in integrating information for consensus clustering; Second, a spectral clustering is much better than a hierarchical clustering in making the most of a co-association matrix. The reason for the second point is complicated, but the lack of explicit global objective function in HCC variants might be one of them; that is, unlike CCC or SEC, HCC variants have no utility function to supervise the process of consensus learning, and therefore could perform much less stably than SEC. This is supported by the extremely poor performances of HCC_GA on *cacmcisi* and *mm* in Table 3, with negative R_n values even poorer than that of random labeling. Similar observations can be found for the newly proposed algorithm PTGP on *mm*, which

1. https://archive.ics.uci.edu/ml/datasets.html
2. http://glaros.dtc.umn.edu/gkhome/cluto/cluto/download.
3. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

TABLE 3
Clustering Results (by R_n) and Running Time (by sec.) in Scenario I

Data set	Clustering Result									Execution Time						
	SEC	CCC	GCC			HCC			PTGP	SEC	CCC	GCC			HCC	PTGP
			CSPA	HGPA	MCLA	GA	SL	CL				CSPA	HGPA	MCLA	GA	
breast_w	*0.82 ± 0.00*	0.07 ± 0.05	0.47	0.61 ± 0.00	0.57	0.78	−0.01	0.15	**0.88**	*0.05*	**0.03**	1.34	0.51	2.15	4.57	2.85
iris	*0.92 ± 0.02*	0.74 ± 0.02	**0.94**	*0.92 ± 0.00*	*0.92*	0.73	0.57	0.64	0.75	*0.03*	**0.02**	0.75	0.17	1.03	0.34	0.75
wine	**0.33 ± 0.00**	0.14 ± 0.00	0.15	0.16 ± 0.00	0.14	0.15	−0.01	0.07	*0.19*	**0.02**	**0.02**	0.83	0.20	1.44	*0.11*	0.76
cacmcisi	**0.64 ± 0.02**	−0.04 ± 0.00	0.34	0.09 ± 0.02	0.32	−0.03	−0.01	−0.04	*0.57*	**0.25**	**0.25**	18.55	*3.87*	13.15	543.20	117.69
classic	**0.68 ± 0.02**	0.37 ± 0.07	0.45	0.17 ± 0.05	0.38	0.38	−0.01	0.04	*0.65*	**0.62**	*1.15*	32.14	7.33	27.44	1,640.71	524.76
cranmed	*0.95 ± 0.00*	**0.96 ± 0.00**	0.67	0.59 ± 0.02	0.75	0.94	0.01	0.14	0.94	**0.12**	*0.14*	5.27	1.62	5.55	105.41	35.76
hitech	**0.29 ± 0.01**	0.21 ± 0.02	0.25	0.12 ± 0.02	0.16	*0.27*	0.00	0.02	0.22	**0.19**	*0.23*	4.77	1.72	6.47	102.93	36.93
k1b	*0.57 ± 0.08*	0.32 ± 0.08	0.24	0.18 ± 0.03	0.27	**0.64**	−0.05	0.26	0.42	**0.17**	*0.23*	5.66	1.92	6.69	119.36	35.27
la12	**0.51 ± 0.07**	0.32 ± 0.10	0.35	0.09 ± 0.03	0.36	0.36	0.36	0.36	*0.40*	**0.17**	**0.17**	21.48	*5.90*	18.84	1,148.17	44.27
mm	**0.62 ± 0.05**	0.43 ± 0.07	*0.44*	0.00 ± 0.01	0.38	−0.01	−0.01	−0.01	−'0.01	**0.05**	*0.06*	6.57	1.70	5.27	112.34	10.61
re1	**0.28 ± 0.02**	*0.23 ± 0.02*	0.19	0.17 ± 0.01	*0.23*	**0.28**	0.06	0.14	*0.23*	**0.30**	*0.44*	3.32	1.61	6.46	66.16	32.20
reviews	**0.53 ± 0.05**	0.43 ± 0.08	0.33	0.05 ± 0.03	0.39	*0.46*	*0.46*	*0.46*	*0.46*	*0.12*	**0.11**	10.97	3.39	10.90	397.16	26.89
sports	*0.47 ± 0.03*	0.29 ± 0.08	0.26	0.10 ± 0.03	0.29	**0.48**	**0.48**	**0.48**	**0.48**	*0.30*	**0.25**	39.37	8.46	28.44	2,319.06	56.02
tr11	**0.59 ± 0.06**	0.46 ± 0.07	0.37	0.38 ± 0.00	0.38	**0.59**	0.28	0.41	*0.49*	**0.05**	**0.05**	1.68	*0.36*	1.89	1.59	2.45
tr12	*0.46 ± 0.03*	0.43 ± 0.04	0.30	0.42 ± 0.03	**0.47**	0.45	0.35	0.29	0.43	**0.05**	**0.05**	1.03	*0.36*	1.58	0.67	3.31
tr41	**0.45 ± 0.05**	0.38 ± 0.05	0.30	0.36 ± 0.03	0.36	*0.43*	0.15	0.25	*0.43*	**0.08**	**0.08**	1.83	*0.70*	2.53	10.36	7.49
tr45	*0.45 ± 0.05*	0.33 ± 0.04	0.36	0.40 ± 0.03	0.38	**0.46**	0.29	0.23	0.33	**0.06**	*0.08*	1.84	0.51	2.34	5.27	5.30
letter	*0.12 ± 0.01*	*0.12 ± 0.00*	0.10	0.08 ± 0.01	**0.13**	0.11	0.00	0.05	N/A	**4.46**	*10.05*	130.48	14.02	27.61	2,778.01	N/A
mnist	*0.42 ± 0.02*	0.40 ± 0.02	N/A	0.18 ± 0.01	0.37	**0.45**	0.00	0.05	N/A	**6.38**	*8.31*	N/A	21.81	29.50	38,686.17	N/A
score/avg.	**18.60**	12.65	11.88	9.20	13.45	*14.85*	5.49	7.63	13.40	**0.71**	*1.14*	15.96	4.01	10.49	2808.93	55.66

Note: (1) N/A means the out-of-memory failures. (2) We omit the zero standard deviations of CSPA, MCLA, HCC, and PTGP for space concern. (3) In runtime comparison, we omit two variants of HCC with similar performances due to space concern. (4) The best is highlighted in bold, and the second best in italic.

employs the mini-cluster based *core co-association matrix* but also lacks of utility functions for consensus learning.

We finally turn to CCC, which shares with SEC the K-means clustering in consensus clustering but assigns equal weights to instances. From Table 3, the performance of CCC seems much poorer than that of SEC, especially on *breast_w* and *cacmcisi*. This indicates that equally weighting of data instances might not be appropriate for consensus learning. In contrast, starting from the spectral clustering view of a co-association matrix, SEC enforces the weights of the instances in large clusters in a quite natural way, and finally leads to superior performances.

7.1.3 Validation of Efficiency

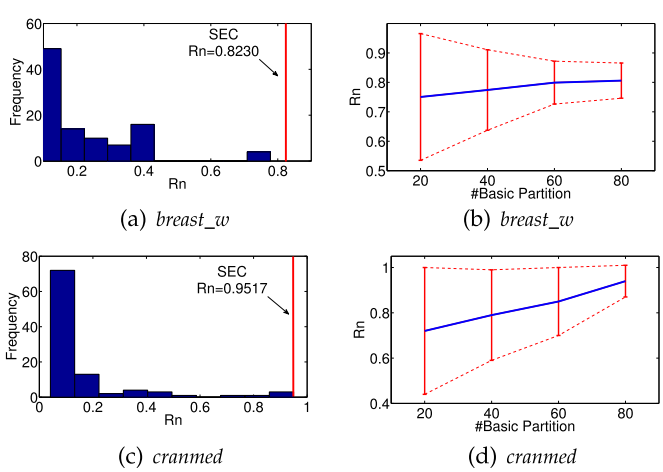
Table 3 (Right side) shows the average execution time of various consensus clustering methods with 50 repetitions. Since HCC variants have similar execution time, we here only report the results of HCC_GA due to limited space. It is obvious that the K-means-like methods, such as SEC and CCC, get clear edges to competitors, and HCC runs the slowest for adopting hierarchical clustering. This indeed demonstrates the value of SEC in transforming spectral clustering of co-association matrix into weighted K-means clustering. On one hand, we make use of co-association matrix to integrate the information of basic partitions nicely. On the other hand, we avoid generating and handling co-association matrix directly but make use of weighted K-means clustering on the binary matrix to gain high efficiency. Although PTGP runs faster than HCC, it needs much more memory and fails to deliver results for two large data sets *letter* and *mnist*.

7.1.4 Validation of Robustness

Figs. 2a and 2c demonstrate the robustness of SEC by taking *breast_w* and *cranmed* as example. We choose these two data sets due to their relatively well-structured clusters—it is often difficult to observe the theoretical properties of an algorithm given very poor performances. We can see that for each data set, the majority of basic partitions are of very low quality. For example, the quality of over 60 basic partitions on *cranmed* is below 0.1 in terms of R_n. Nevertheless, SEC performs excellently (with $R_n > 0.95$) by leveraging the diversity among poor basic partitions. Similar phenomena also occur on some other data sets like *breast_w*, which indicates the power of SEC in fusing diverse information from even poor basic partitions.

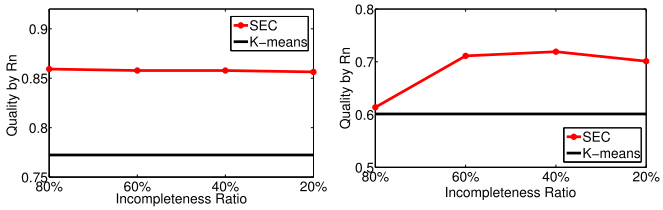
7.1.5 Validation of Generalizability and Convergence

Next, we check the generalizability and convergence of SEC. Figs. 2b and 2d show the results by varying the number of basic partitions from 20 to 80 for *breast_w* and *cranmed*, respectively. Note that the above process is repeated 20



(a) *breast_w* (b) *breast_w* (c) *cranmed* (d) *cranmed*

Fig. 2. Impact of quality and quantity of basic partitions.



(a) *mm* (b) *reviews*

Fig. 3. Performance of SEC with different incompleteness ratios.

times for average results. Generally speaking, it is clear that with the increasing number of basic partitions (i.e., r), the performance of SEC goes up and becomes stable gradually. For instance, SEC achieves satisfactory result from *breast_w* with only 20 basic partitions, but it also suffers from high volatility given such a small r; when r goes up, the variance becomes narrow and stabilizes in a small region.

7.1.6 Effectiveness of Incompleteness Treatment

Here, we demonstrate effectiveness of SEC in handling incomplete basic partitions. The row-segmentation strategy is employed to generate IBPs. In detail, data instances are first randomly sampled with replacement, with the sampling ratio going up from 20 to 80 percent, to form overlapped data subsets and generate IBPs; SEC is then called to ensemble these IBPs and obtain a consensus partition. Note that for each ratio, the above process repeats 100 times to obtain IBPs, and unsampled instances are omitted in the final consensus learning. It is intuitive that a lower sampling ratio leads to smaller overlaps between IBPs and thus worse clustering performances. Fig. 3 shows the sample results on *mm* and *reviews*, where the horizontal line indicates the K-means clustering result on the original data set and serves as the baseline unchanged with the sampling ratio. As can be seen, SEC keeps providing stable and competitive results as the sampling ratio goes down to 20 percent, which demonstrates the effectiveness of incompleteness treatment of SEC.

7.2 Scenario II: Multi-View Clustering

7.2.1 Experimental Setup

Data. Four real-world data sets, i.e., *UCI Handwritten Digit*, *three-Sources*, *Multilingual* and *four-Areas* listed in Table 4, are used in the experiments. *UCI Handwritten Digit*[4] consists of 0-9 handwritten digits obtained from the UCI repository, where each digit has 200 instances with 240 features in pixel view and 76 features in Fourier view. *three-Sources*[5] is collected from three online news sources: BBC, Guardian and Reuter, from February to April in 2009. Of these documents, 169 are reported in all three sources (views). Each document is annotated with one of six categories: business, entertainment, health, politics, sports and technology. *Multilingual*[6] contains the documents written originally in five different languages over 6 categories. We here use the sample suggested by [35], which has 100 documents for each category with three views in English, German and French, respectively. *four-Areas*[7] is derived from 20 conferences in four areas including database, data mining, machine learning and information retrieval. It contains 28,702 authors and 13,214 terms in the abstract. Each author is labeled with one or multiple areas, and the cross-area authors are removed for unambiguous evaluation. The remainder has 4,236 authors in both conference and term views.

Tool. We compare SEC with a number of baseline algorithms including ConKM, ConNMF, ColNMF [32], CRSC [34], MultiNMF [35] and PVC [36]. All the competitors are with default settings whenever possible. Gaussian kernel is used to build the affinity matrix for CRSC. The trade-off parameter λ is set to 0.01 for MultiNMF as suggested in Ref. [35]. For SEC, we employ the *kmeans* function in MATLAB to generate one basic partition for each view, and then call SEC to fuse them with equal weights into a consensus one. Each algorithm is called 50 times for the average results.

Validation. For consistency, we also employ R_n to evaluate cluster validity.

TABLE 4
Experimental Data Sets for Scenario II

View	Digit	3-Sources	Multilingual	4-Areas
1	Pixel (240)	BBC (3,560)	English (9,749)	Conference (20)
2	Fourier (74)	Guardian (3,631)	German (9,109)	Term (13,214)
3	-	Reuters (3,068)	French (7,774)	-
#Instances	2,000	169	600	4,236
#Classes	10	6	6	4

TABLE 5
Clustering Results in Scenario II (by R_n)

Data sets	Digit	3-Sources	Multilingual	4-Areas
ConKM	0.58 ± 0.06	0.16 ± 0.08	0.12 ± 0.04	0.00 ± 0.00
ConNMF	0.49 ± 0.06	0.28 ± 0.09	0.22 ± 0.02	0.03 ± 0.06
ColNMF	0.39 ± 0.03	0.20 ± 0.05	0.22 ± 0.02	*0.11 ± 0.14*
CRSC	*0.64 ± 0.03*	*0.30 ± 0.04*	*0.24 ± 0.01*	0.00 ± 0.00
MultiNMF	**0.65 ± 0.03**	0.22 ± 0.06	0.22 ± 0.02	0.00 ± 0.00
PCV	0.56 ± 0.00	N/A	N/A	0.01 ± 0.00
SEC	0.44 ± 0.05	**0.55 ± 0.09**	**0.25 ± 0.03**	**0.56 ± 0.09**

Note: N/A means no result due to more than two views data sets.

7.2.2 Comparison of Clustering Quality

Table 5 shows the clustering results on four multi-view data sets, with the best results highlighted in bold and the second best in italic. The sign "N/A" indicates PVC cannot handle data with more than two views.

As can be seen from Table 5, SEC generally shows higher clustering performances than the baselines, especially for data sets *three-Sources* and *four-Areas*—actually all baselines seem completely ineffective in inferring the structure of *four-Areas*. This indeed reveals the unique merit of SEC for multi-view clustering; that is, SEC works on new features from basic partitions rather than original features, which might avoid the negative impact of data dimensionality, especially when dealing with data sets such as *four-Areas* that have two views of substantially different dimensionalities.

It is also noteworthy that SEC has poor performance on *Digit*. If we take a close look at the two basic partitions for SEC, we can find the contrastive performances, i.e., $R_n = 0.65$ and 0.32 on "Pixel" and "Fourier", respectively.

4. http://archive.ics.uci.edu/ml/datasets.html
5. http://mlg.ucd.ie/datasets
6. http://www.webis.de/research/corpora
7. http://www.ccs.neu.edu/home/yzsun/data/four_area.zip

TABLE 6
Clustering Results in Scenario II with Pseudo Views (by R_n)

Data sets	Digit	3-Sources	Multilingual	4-Areas
ConKM	0.62 ± 0.09	0.09 ± 0.05	0.15 ± 0.04	0.00 ± 0.00
ConNMF	0.51 ± 0.05	0.25 ± 0.04	0.21 ± 0.00	0.02 ± 0.06
ColNMF	0.43 ± 0.07	0.14 ± 0.09	0.20 ± 0.00	*0.04 ± 0.08*
CRSC	*0.66 ± 0.02*	*0.32 ± 0.02*	*0.25 ± 0.04*	0.00 ± 0.00
MultiNMF	0.65 ± 0.06	0.23 ± 0.08	0.22 ± 0.01	0.00 ± 0.01
PCV	N/A	N/A	N/A	N/A
SEC	$\mathbf{0.69 \pm 0.06}$	$\mathbf{0.62 \pm 0.09}$	$\mathbf{0.29 \pm 0.03}$	$\mathbf{0.67 \pm 0.09}$

Note: N/A means no result due to more than two views data sets.

As a result, given the only two basic partitions, SEC can only find a comprise and thus results in poor performance. One straightforward remedy is to make full use of the robustness of SEC by increasing the number of basic partitions in each view, as suggested by Sections 4.1 and 7.1.4. We give experimental results below.

7.2.3 Robustness Revisited

As mentioned above, sufficient basic partitions could enhance the robustness of SEC via repeating valid local structures. To better understand this, in this experiment, we generate $r = 20$ basic partitions for each view using a random feature selection scheme. That is, for each basic partition, we take all the data instances but sample the features randomly with a ratio r_s so as to form a data subset. We set $r_s = 50\%$ empirically for keeping enough feature information for basic clustering yet without sacrificing the diversity of basic partitions. By this means, SEC gains multiple *pseudo views* of data, which is good to leverage its robustness property.

From Table 6, extra pseudo views indeed boost the performance of multi-view clustering and SEC. Specially, the competitive multi-view clustering methods have slightly improvements on the first three data sets while SEC consistently have significant gains on all four data sets. In particular, SEC with pseudo views performs even better than the baselines on *Digit*. This not only demonstrates the effectiveness of random feature selection for basic partition but also illustrates how to inspire the robustness of SEC in multi-view learning.

7.2.4 Dealing with Partial Multi-View Clustering

In real-world applications, it is common to collect partial multi-view data, i.e., incomplete data in different views, due to device failures or transmission loss [36]. Here we validate the performance of SEC on partial multi-view data, with the well-known PVC designed purposefully for partial multi-view clustering as a baseline.

To simulate the partial multi-view setting, we randomly select a fraction of instances, from 5 to 30 percent with 5 percent as interval from each view. In Fig. 4, the four solid lines in blue represent the performances of SEC on four data sets with varying missing rates, and the two dash lines depict the results of PVC on *Digit* and *four-Areas*. Note that: 1) SEC employs pseudo views with $r = 20$ and $r_s = 50\%$ for each view; 2) PVC only has results on two-view data sets.

From Fig. 4, we can see that the performance of SEC and PVC generally goes down as the missing rate increases.

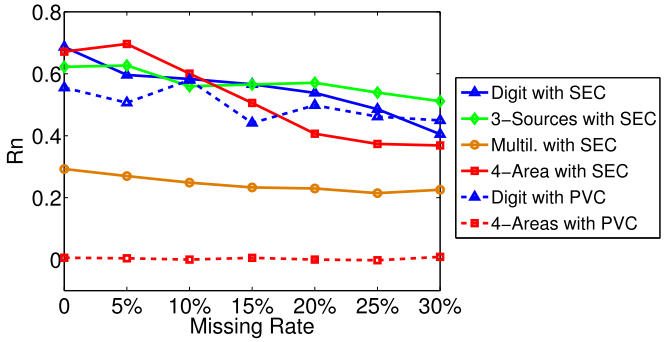


Fig. 4. Clustering results of partial multi-view data.

Nevertheless, SEC behaviors relatively stably on three-view rather than two-view data sets. This is because three-view data sets can provide more information given the same missing rate on each view. More importantly, SEC outperforms PVC by clear margins in nearly all scenarios on *Digit* and *four-Areas*, which again demonstrates the advantage of SEC in handling incomplete basic partitions.

7.3 SEC for Weibo Data Clustering

Sina Weibo,[8] a Twitter-like service launched in 2009, is a popular social media platform in China. It has accumulated more than 500 million users and has around 100 million tweets published everyday, which provides tremendous value for commercial applications and academic research.

Next we illustrate how to employ SEC to cluster the entire Weibo data published on Sept. 1st, 2013, which consist of 97,231,274 Chinese tweets altogether. Python environment is adopted to facilitate text processing. After removing 30 million advertisement related tweets via simple keywords filtering, SCWS[9] is applied to build the vector space model with top 10,000 frequently used terms. By this means, we obtain a text corpus with 61,212,950 instances and 10,000 terms. Next, the row-segmentation strategy proposed in Section 6 is called to acquire 100 data subsets each with 10,000,000 instances, and the famous text clustering tool CLUTO[10] with default settings is then called in parallel to cluster these data subsets into basic partitions.

SEC is finally called to fuse the basic partitions into a consensus one. To achieve this, we build a simple distributed system with 10 servers to accelerate the fusing process. In detail, the binary matrix derived from the 100 IBPs is first horizontally split and distributed to every computational nodes. One server is chosen as the master to broadcast the centroid matrix to all nodes during weighted K-means clustering. Each node then computes the distances between local binary vectors and the centroids, assigns the cluster labels, and summarizes a partial centroid matrix as return to the master server. After receiving all partial centroid matrices in the master node, the centroid matrix is updated and a new iteration begins. Note that the cluster number is set to 100 for both basic and consensus clustering.

The results of some clusters tagged by the representative keywords are shown in Table 7. It can be inferred easily that

8. http://www.weibo.com/
9. http://www.xunsearch.com/scws/
10. http://glaros.dtc.umn.edu/gkhome/views/cluto

TABLE 7
Sample Weibo Clusters Characterized by Keywords

ID	Keywords
Clu.3	*term begins, campus, partner, teacher, school, dormitory*
Clu.21	*Mid-Autumn Festival, September, family, happy, parents*
Clu.40	*China, powerful, history, victory, Japan, shock, harm*
Clu.65	*Meng Ge, mother, apologize, son, harm, regret, anger*
Clu.83	*travel, happy, dream, life, share, picture, plan, haha*

Cluster #3, #21, and #83 represent "the beginning of new semester", "mid-autumn festival", and "travel" events, respectively. In Cluster #40, the tweets reflect the user opinions towards the conflict between China and Japan due to the "September 18th incident"; Cluster #65 reports a hot event that Meng Ge, a famous female singer in China, apologized for her son's crime. In general, although the basic partitions are highly incomplete, some interesting events can still be discovered by using the row-segmentation strategy. SEC appears to be a promising candidate for big data clustering.

8 Conclusion

In this paper, we proposed the Spectral Ensemble Clustering algorithm. By identifying the equivalent relationship between SEC and weighted K-means, we decreased the time and space complexities of SEC dramatically. The intrinsic consensus objective function of SEC was also revealed, which bridges the co-association matrix based methods with the methods with explicit global objective functions. We then investigated the robustness, generalizability and convergence properties of SEC to showcase its superiority in theory, and extended it to handle incomplete basic partitions. Extensive experiments demonstrated that SEC is an effective and efficient algorithm compared with some state-of-the-art methods in both the ensemble and multi-view clustering scenarios. We further proposed a row-segmentation scheme for SEC, and demonstrated its effectiveness via the case of consensus clustering of big Weibo data.

Appendix A
Proof of Theorem 1

Proof. Let $\mathbf{Y} = \{y = b(x)/w_{b(x)}\}$ and $\mathbf{W}_k$ denote the diagonal matrix of the weights in cluster C_k, and $\mathbf{Y}_k$ denote the matrix of binary data associated with cluster C_k. Then the centroid m_k can be rewrote as $m_k = \mathbf{e}^\top \mathbf{W}_k \mathbf{Y}_k / s_k$, where $\mathbf{e}$ is the vector of all ones with appropriate size and $s_k = \mathbf{e}^\top \mathbf{W}_k \mathbf{e}$. According to [42], we have

$$\begin{aligned} SSE_{C_k} &= \sum_{x \in C_k} w_{b(x)} \left\| \frac{b(x)}{w_{b(x)}} - m_k \right\|^2 \\ &= \left\| \left(\mathbf{I} - \frac{\mathbf{W}_k^{1/2} \mathbf{e}\mathbf{e}^\top \mathbf{W}_k^{1/2}}{s_k} \right) \mathbf{W}_k^{1/2} \mathbf{Y}_k \right\|_{\mathrm{F}}^2 \\ &= \mathrm{tr}(\mathbf{Y}_k^\top \mathbf{W}_k^{1/2} \left(\mathbf{I} - \frac{\mathbf{W}_k^{1/2} \mathbf{e}\mathbf{e}^\top \mathbf{W}_k^{1/2}}{s_k} \right)^2 \mathbf{W}_k^{1/2} \mathbf{Y}_k) \\ &= \mathrm{tr}(\mathbf{W}_k^{1/2} \mathbf{Y}_k \mathbf{Y}_k^\top \mathbf{W}_k^{1/2}) - \frac{\mathbf{e}^\top \mathbf{W}_k}{\sqrt{s_k}} \mathbf{Y}_k \mathbf{Y}_k^\top \frac{\mathbf{W}_k \mathbf{e}}{\sqrt{s_k}}. \end{aligned}$$

If we sum up SSE of all the clusters, we have

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{x \in C_k} w_{b(x)} \left\| \frac{b(x)}{w_{b(x)}} - m_k \right\|^2 \\ &= \mathrm{tr}(\mathbf{W}^{\frac{1}{2}} \mathbf{Y}\mathbf{Y}^\top \mathbf{W}^{\frac{1}{2}}) - \mathrm{tr}(\mathbf{G}^\top \mathbf{W}^{\frac{1}{2}} \mathbf{Y}\mathbf{Y}^\top \mathbf{W}^{\frac{1}{2}} \mathbf{G}), \end{aligned}$$

where $\mathbf{G} = diag(\frac{\mathbf{W}_1^{1/2}\mathbf{e}}{\sqrt{s_1}}, \ldots, \frac{\mathbf{W}_k^{1/2}\mathbf{e}}{\sqrt{s_k}})$. Recall that $\mathbf{Y}\mathbf{Y}^\top = \mathbf{W}^{-1}\mathbf{B}\mathbf{B}^\top\mathbf{W}^{-1}$ and $\mathbf{S} = \mathbf{B}\mathbf{B}^\top$, $\mathbf{D} = \mathbf{W}$ and $\mathbf{Z}^\top\mathbf{Z} = \mathbf{G}^\top\mathbf{G} = \mathbf{I}$, so we have

$$\max \mathrm{tr}(\mathbf{Z}^\top \mathbf{D}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}^{-\frac{1}{2}} \mathbf{Z}) \Leftrightarrow \max \mathrm{tr}(\mathbf{G}^\top \mathbf{W}^{-\frac{1}{2}} \mathbf{B}\mathbf{B}^\top \mathbf{W}^{-\frac{1}{2}} \mathbf{G}).$$

The constant $\mathrm{tr}(\mathbf{W}^{-\frac{1}{2}} \mathbf{B}\mathbf{B}^\top \mathbf{W}^{-\frac{1}{2}})$ finishes the proof. □

Appendix B
Proof of Theorem 2

Proof. Given the equivalence of SEC and weighted K-means, we here derive the utility function of SEC. We start from the objective function of weighted K-means as follows:

$$\begin{aligned} &\sum_{k=1}^{K} \sum_{x \in C_k} w_{b(x)} \left\| \frac{b(x)}{w_{b(x)}} - m_k \right\|^2 \\ &= \sum_{k=1}^{K} \left[\sum_{x \in C_k} \frac{||b(x)||^2}{w_{b(x)}} - 2 \sum_{x \in C_k} b(x) m_k^\top + \sum_{x \in C_k} w_{b(x)} ||m_k||^2 \right] \\ &= \sum_{k=1}^{K} \left[\sum_{x \in C_k} \frac{||b(x)||^2}{w_{b(x)}} - 2 \sum_{x \in C_k} w_{b(x)} ||m_k||^2 + \sum_{x \in C_k} w_{b(x)} ||m_k||^2 \right] \\ &= \sum_{k=1}^{K} \sum_{x \in C_k} \frac{||b(x)||^2}{w_{b(x)}} - \sum_{i=1}^{r} \sum_{k=1}^{K} w_{C_k} ||m_{k,i}||^2 \\ &= \underbrace{\sum_{k=1}^{K} \sum_{x \in C_k} \frac{||b(x)||^2}{w_{b(x)}}}_{(\gamma)} - n \sum_{i=1}^{r} \sum_{k=1}^{K} \frac{n_{k+}}{w_{C_k}} p_{k+} \sum_{j=1}^{K_i} \left(\frac{p_{kj}^{(i)}}{p_{k+}} \right)^2. \end{aligned}$$

According to the definition of centroids in K-means, we have $m_{k,ij} = \sum_{x \in C_k} b(x)_{ij} / \sum_{x \in C_k} w_{b(x)} = n_{kj}^{(i)} / w_{C_k} = (n_{kj}^{(i)} / n_{k+})(n_{k+}/w_{C_k}) = (p_{kj}^{(i)} / p_{k+})(n_{k+}/w_{C_k})$. Note that (γ) is a constant, so we get the utility function of SEC. □

Appendix C
Proof of Theorem 3

We first give a lemma as follows.

Lemma 1.

$$f_{m_1,\ldots,m_K}(x) \in [0, 1].$$

Proof. It is easy to show $\|b(x)\|^2 = r$, $w_{b(x)} \in [r, (n-K+1)r]$ and $f_{m_1,\ldots,m_K}(x) \leq \max\{\frac{\|b(x)\|^2}{w_{b(x)}}, w_{b(x)}\|m_k\|^2\}$. We have $\frac{\|b(x)\|^2}{w_{b(x)}} \leq \frac{r}{r} = 1$ and

$$w_{b(x)} \|m_k\|^2 = \frac{w_{b(x)} \| \sum_{b(x) \in C_k} b(x) \|^2}{\left(\sum_{b(x) \in C_k} w_{b(x)} \right)^2} \leq 1. \tag{9}$$

This concludes the proof.

A detailed proof of Eq. (9): If $|C_k| = 1$, the equation holds trivially. When $|C_k| \geq 2$, we have

$$\begin{aligned}
&\frac{w_{b(x)}\|\sum_{b(x)\in C_k} b(x)\|^2}{\left(\sum_{b(x)\in C_k} w_{b(x)}\right)^2} \leq \frac{w_{b(x)}\sum_{b(x)\in C_k}\|b(x)\|^2}{\left(\sum_{b(x)\in C_k} w_{b(x)}\right)^2} \\
&= \frac{w_{b(x)}\sum_{b(x)\in C_k} r}{\left(w_{b(x)} + \sum_{b(x)\in C_k - \{b(x)\}} w_{b(x)}\right)^2} \\
&\leq \frac{w_{b(x)}\sum_{b(x)\in C_k} r}{\left(w_{b(x)} + \sum_{b(x)\in C_k - \{b(x)\}} r\right)^2} = \frac{w_{b(x)}|C_k|r}{\left(w_{b(x)} + (|C_k|-1)r\right)^2} \\
&\leq \frac{w_{b(x)}|C_k|r}{\left(w_{b(x)}\right)^2 + 2w_{b(x)}(|C_k|-1)r} \\
&\leq \frac{|C_k|r}{w_{b(x)} + 2(|C_k|-1)r} \leq \frac{|C_k|r}{|C_k|r + |C_k|r - r} \leq 1.
\end{aligned}$$

The first inequality holds due to the triangle inequality. □

Now we begins the proof of Theorem 3.

Proof. We have

$$\begin{aligned}
&|f_{m_1,\ldots,m_K}(x) - f_{m_1,\ldots,m_K}(x')| \\
&= \left|\min_k w_{b(x)}\left\|\frac{b(x)}{w_{b(x)}} - m_k\right\| - \min_k w_{b(x')}\left\|\frac{b(x')}{w_{b(x')}} - m_k\right\|\right| \\
&\leq \max_k\left|w_{b(x)}\left\|\frac{b(x)}{w_{b(x)}} - m_k\right\| - w_{b(x')}\left\|\frac{b(x')}{w_{b(x')}} - m_k\right\|\right| \\
&= \max_k\Bigg|\frac{r}{w_{b(x)}} - \langle b(x), m_k\rangle + w_{b(x)}\|m_k\|^2 - \frac{r}{w_{b(x')}} \\
&\qquad + \langle b(x'), m_k\rangle - w_{b(x')}\|m_k\|^2\Bigg| \\
&\leq \max_k\Bigg(|\frac{r}{w_{b(x)}} - \frac{r}{w_{b(x')}}| + \|b(x) - b(x')\|\|m_k\| \\
&\qquad + \|m_k\|^2|w_{b(x)} - w_{b(x')}|\Bigg).
\end{aligned}$$

Note that the last inequality holds due to the Cauchy-Schwartz inequality. Recall that we have proved in Lemma 1 that $\|m_k\|^2 \leq \frac{1}{\min_{x\in X} w_{b(x)}}$, we have

$$\begin{aligned}
&|f_{m_1,\ldots,m_K}(x) - f_{m_1,\ldots,m_K}(x')| \\
&\leq \max_k\left(\left|\frac{r}{w_{b(x)}} - \frac{r}{w_{b(x')}}\right| + \|b(x) - b(x')\|\|m_k\|\right. \\
&\qquad \left. + \|m_k\|^2\left|w_{b(x)} - w_{b(x')}\right|\right) \\
&\leq \max_k\left(\frac{r}{\min_{x\in X_n}\left(w_{b(x)}\right)^2} + \|m_k\|^2\right)|w_{b(x)} - w_{b(x')}| \\
&\qquad + \|b(x) - b(x')\|\|m_k\| \\
&\leq \frac{r + \min_{x\in X} w_{b(x)}}{\min_{x\in X}\left(w_{b(x)}\right)^2}\sum_{i=1}^{r}\gamma_{w,i} + \left(\frac{\sum_{i=1}^{r}\gamma_i^2}{\min_{x\in X} w_{b(x)}}\right)^{\frac{1}{2}} \\
&\leq \frac{2\sum_{i=1}^{r}\gamma_{w,i}}{r} + \left(\frac{\sum_{i=1}^{r}\gamma_i^2}{r}\right)^{\frac{1}{2}}.
\end{aligned}$$

This completes the proof. □

APPENDIX D
PROOF OF THEOREM 4

The Glivenko-Cantelli theorem [49] is often used, together with complexity measures, to analyze the non-asymptotic uniform convergence of $E_n f_{m_1,\ldots,m_K}(x)$ to $E_x f_{m_1,\ldots,m_K}(x)$, where $E_n f(x)$ denotes the empirical expectation of $f(x)$. A relatively small complexity of the function class $F_{\Pi_K} = \{f_{m_1,\ldots,m_K} | \pi \in \Pi_K\}$, where Π_K denotes all possible K-means clustering for $\mathcal{X}$ is essential to prove a Glivenko-Cantelli class. Rademacher complexity is one of the most frequently used complexity measures.

Rademacher complexity and *Gaussian complexity* are data-dependent complexity measures. They are often used to derive dimensionality-independent generalization error bounds and defined as follows:

Definition 2. *Let $\sigma_1,\ldots,\sigma_n$ and $\gamma_1,\ldots,\gamma_n$ be independent Rademacher variables and independent standard normal variables, respectively. Let $x_1,\ldots,x_n$ be an independent distributed sample and F a function class. The empirical Rademacher complexity and empirical Gaussian complexity are defined as*

$$\mathfrak{R}_n(F) = E_\sigma \sup_{f\in F}\frac{1}{n}\sum_{l=1}^{n}\sigma_l f(x_l)$$

$$\mathcal{G}_n(F) = E_\gamma \sup_{f\in F}\frac{1}{n}\sum_{l=1}^{n}\gamma_l f(x_l),$$

respectively. The expected Rademacher complexity and Gaussian complexity are defined as

$$\mathfrak{R}(F) = E_x\mathfrak{R}_n(F) \text{ and } \mathcal{G}(F) = E_x\mathcal{G}_n(F).$$

Using the symmetric distribution property of random variables, we have:

Theorem A 1. *Let F be a real-valued function class on $\mathcal{X}$ and $X = (x_1,\ldots,x_n) \in \mathcal{X}^n$. Let*

$$\Phi(X) = \sup_{f\in F}\frac{1}{n}\sum_{l=1}^{n}(E_x f(x) - f(x_l)).$$

Then, $E_x\Phi(X) \leq 2\mathfrak{R}(F)$.

The following theorem [50], proved utilizing Theorem A 1 and McDiarmid's inequality, plays an important role in proving the generalization error bounds:

Theorem A 2. *Let F be an $[a,b]$-valued function class on $\mathcal{X}$, and $X = (x_1,\ldots,x_n) \in \mathcal{X}^n$. For any $f \in F$ and $\delta > 0$, with probability at least $1-\delta$, we have*

$$E_x f(x) - \frac{1}{n}\sum_{l=1}^{n} f(x_l) \leq 2\mathfrak{R}(F) + (b-a)\sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Combining Theorem A 2 and Lemma 1, we have

Theorem A 3. *Let π be any partition learned by SEC. For any independently distributed instances $x_1,\ldots,x_n$ and $\delta > 0$, with probability at least $1-\delta$, the following holds*

$$E_x f_{m_1,\ldots,m_K}(x) - \frac{1}{n}\sum_{l=1}^{n} f_{m_1,\ldots,m_K}(x_l) \leq 2\mathfrak{R}(F_{\Pi_K}) + \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

We use Lemmas 2 and 3 (see proofs in [51]) to upper bound $\mathfrak{R}(F_{\Pi_K})$ by finding a proper Gaussian process which can easily be bounded.

Lemma 2 (Slepian's Lemma). *Let Ω and Ξ be mean zero, separable Gaussian processes indexed by a common set $\mathcal{S}$, such that*

$$E(\Omega_{s_1} - \Omega_{s_2})^2 \leq E(\Xi_{s_1} - \Xi_{s_2})^2, \forall s_1, s_2 \in \mathcal{S}.$$

Then $E\sup_{s\in\mathcal{S}}\Omega_s \leq E\sup_{s\in\mathcal{S}}\Xi_s$.

The Gaussian complexity is related to the Rademacher complexity by the following lemma:

Lemma 3.

$$\mathfrak{R}(F) \leq \sqrt{\pi/2}\mathcal{G}(F).$$

Now, we can upper bound the Rademacher complexity $\mathfrak{R}(F_{\mathcal{W}})$ by finding a proper Gaussian process.

Lemma 4.

$$\mathfrak{R}(F_{\Pi_K}) \leq \frac{\sqrt{\pi/2}rK}{n}\left(\left(\sum_{l=1}^{n}\frac{1}{(w_{b(x_l)})^2}\right)^{\frac{1}{2}} + \frac{2\sqrt{n}}{\min_{x\in X}w_{b(x)}} + \left(\sum_{l=1}^{n}(w_{b(x_l)})^2\right)^{\frac{1}{2}}\frac{1}{\min_{x\in X}(w_{b(x)})^2}\right).$$

Proof. Let $\boldsymbol{M} \in \mathbb{R}^{\sum_{i=1}^{r}K_i\times K}$, whose kth column represents the kth centroid m_k. Define the Gaussian processes indexed by $\boldsymbol{M}$ as

$$\Omega_{\boldsymbol{M}} = \sum_{l=1}^{n}\gamma_l \min_k w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2$$
$$\Xi_{\boldsymbol{M}} = \sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk} w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2,$$

where γ_l and γ_{lk} are independent Gaussian random variables indexed by l and k. And e_k are the natural bases indexed by k.

For any $\boldsymbol{M}$ and $\boldsymbol{M}'$, we have

$$\begin{aligned}
&E(\Omega_{\boldsymbol{M}} - \Omega_{\boldsymbol{M}'})^2 \\
&= \sum_{l=1}^{n}\left(\min_k w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2 - \min_k w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}'e_k\right\|^2\right)^2 \\
&\leq \sum_{l=1}^{n}\max_k\left(w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2 - w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}'e_k\right\|^2\right)^2 \\
&\leq \sum_{l=1}^{n}\sum_{k=1}^{K}\left(w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2 - w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}'e_k\right\|^2\right)^2 \\
&= E(\Xi_{\boldsymbol{M}} - \Xi_{\boldsymbol{M}'})^2.
\end{aligned}$$

Note that the first and last inequalities hold because of the orthogaussian properties.

Using Slepian's Lemma and Lemma 3, we have

$$\begin{aligned}
&\mathfrak{R}(F_{\Pi_K}) \\
&= E_\sigma \sup_{\boldsymbol{M}}\frac{1}{n}\sum_{l=1}^{n}\sigma_l \min_k w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2 \\
&\leq E_\gamma \frac{\sqrt{\pi/2}}{n}\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\gamma_l \min_k w_{b(x_l)}\left\|\frac{b(x_l)}{w_{b(x_l)}} - \boldsymbol{M}e_k\right\|^2 \\
&= \frac{\sqrt{\pi/2}}{n}E_\gamma\left(\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}w_{b(x_l)}\left(\frac{\|b(x_l)\|^2}{(w_{b(x_l)})^2} - 2\left\langle\frac{b(x_l)}{w_{b(x_l)}}, \boldsymbol{M}e_k\right\rangle + \|\boldsymbol{M}e_k\|^2\right)\right) \\
&\leq \frac{\sqrt{\pi/2}}{n}\left(E_\gamma\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}\frac{r}{w_{b(x_l)}} + 2E_\gamma\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}\langle b(x_l), \boldsymbol{M}e_k\rangle + E_\gamma\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}w_{b(x_l)}\|\boldsymbol{M}e_k\|^2\right).
\end{aligned}$$

We give upper bounds to the three terms respectively

$$\begin{aligned}
&E_\gamma\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}\frac{r}{w_{b(x_l)}} \\
&= E_\gamma r\sum_{k=1}^{K}\sum_{l=1}^{n}\frac{\gamma_{lk}}{w_{b(x_l)}} = E_\gamma r\sum_{k=1}^{K}\sqrt{\left(\sum_{l=1}^{n}\frac{\gamma_{lk}}{w_{b(x_l)}}\right)^2} \\
&\leq r\sum_{k=1}^{K}\sqrt{\sum_{l=1}^{n}\frac{1}{(w_{b(x_l)})^2}} = rK\sqrt{\sum_{l=1}^{n}\frac{1}{(w_{b(x_l)})^2}}.
\end{aligned}$$

Note that the last inequality holds for the Jensen's inequality and the orthogaussian property of the Gaussian random variable. We therefore have

$$\begin{aligned}
&2E_\gamma\sup_{\boldsymbol{M}}\sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}\langle b(x_l), \boldsymbol{M}e_k\rangle \\
&\leq 2E_\gamma\sum_{k=1}^{K}\left\|\sum_{l=1}^{n}\gamma_{lk}b(x_l)\right\|\frac{\sqrt{r}}{\min_{x\in X}w_{b(x)}} \\
&\leq 2\sum_{k=1}^{K}\left(\sum_{l=1}^{n}\|b(x_l)\|^2\right)^{\frac{1}{2}}\frac{\sqrt{r}}{\min_{x\in\ X}w_{b(x)}} = \frac{2\sqrt{n}rK}{\min_{x\in X}w_{b(x)}}.
\end{aligned}$$

The second inequality holds because

$$\begin{aligned}
\|\boldsymbol{M}e_k\| &= \left\|\frac{\sum_{b(x)\in C_k}b(x)}{\sum_{b(x)\in C_k}w_{b(x)}}\right\| \leq \frac{\sum_{b(x)\in C_k}\|b(x)\|}{\sum_{b(x)\in C_k}w_{b(x)}} \\
&\leq \frac{\max_x\|b(x)\|}{\min_{x\in X}w_{b(x)}} = \frac{\sqrt{r}}{\min_{x\in X}w_{b(x)}}.
\end{aligned}$$

For the upper bound $E_\gamma \sup_{\boldsymbol{M}} \sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}w_{b(x_l)}\|\boldsymbol{M}e_k\|^2$,

$$\begin{aligned}
&E_\gamma \sup_{\boldsymbol{M}} \sum_{l=1}^{n}\sum_{k=1}^{K}\gamma_{lk}w_{b(x_l)}\|\boldsymbol{M}e_k\|^2 \\
&\le E_\gamma \sum_{k=1}^{K}\left|\sum_{l=1}^{n}\gamma_{lk}w_{b(x_l)}\right|\left(\frac{\sqrt{r}}{\min_{x\in X}(w_{b(x)})^2}\right)^2 \\
&\le \sum_{k=1}^{K}\left(\sum_{l=1}^{n}(w_{b(x_l)})^2\right)^{\frac{1}{2}}\left(\frac{\sqrt{r}}{\min_{x\in X}(w_{b(x)})^2}\right)^2 \\
&= rK\left(\sum_{l=1}^{n}(w_{b(x_l)})^2\right)^{\frac{1}{2}}\frac{1}{\min_{x\in X}(w_{b(x)})^2}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\mathfrak{R}(F_{\Pi_K}) \\
&\le \frac{\sqrt{\pi/2}}{n}\left(rK\left(\sum_{l=1}^{n}\frac{1}{(w_{b(x_l)})^2}\right)^{\frac{1}{2}}+\frac{2\sqrt{n}rK}{\min_{x\in X}w_{b(x)}}\right. \\
&\quad \left.+rK\left(\sum_{l=1}^{n}(w_{b(x_l)})\right)^2\right)^{\frac{1}{2}}\frac{1}{\min_{x\in X}w_{b(x)}^2}\Bigg) \\
&= \frac{\sqrt{\pi/2}rK}{n}\left(\sum_{l=1}^{n}\frac{1}{(w_{b(x_l)})^2}\right)^{\frac{1}{2}}+\frac{2\sqrt{n}}{\min_{x\in X}w_{b(x)}} \\
&\quad +\left(\sum_{l=1}^{n}(w_{b(x_l)})\right)^2\Bigg)^{\frac{1}{2}}\frac{1}{\min_{x\in X}(w_{b(x)})^2}\Bigg).
\end{aligned}$$

This concludes the proof of Lemma 4. □

Theorem 4 in the paper thus follows according to Theorem A 3 and Lemma 4.

APPENDIX E
PROOF OF THEOREMS 5 AND 8

Proof. It has been proven that the co-associate matrix $\mathbf{S}$ will converge w.r.t. r, the number of basic crisp partitions [52]. That is, for any $\lambda_1 > 0$, there exists a matrix $\mathbf{S}_0$, such that

$$\lim_{r\to\infty}\Pr\{|\mathbf{S}-\mathbf{S}_0|\ge\lambda_1\}\to 0.$$

Thus, according to the definition of $b(x)$ and Theorem 1, we can claim that $b(x)$ and the centroids $m_1,\ldots,m_K$ will converge to some $b_0(x)$ and $m_1^0,\ldots,m_K^0$, respectively, as r goes to infinity. Then, for any $\lambda > 0$, there exist a clustering π_0 such that

$$\lim_{r\to\infty}\Pr\{|\pi-\pi_0|\ge\lambda\}\to 0,$$

which concludes the proof of Theorem 5. □

Since the proof of the convergence property of the co-associate matrix $\mathbf{S}$ also holds for the incomplete basic partitions, Theorem 8 can be easily proven by the same proof method of Theorem 5.

APPENDIX F
PROOF OF THEOREM 6

Proof. The proof of Theorem 6 is similar to the proof of Theorem 2, with the only difference being that the missing elements are not taken into account in the objective function of weighted K-means clustering. We therefore have

$$\begin{aligned}
&\sum_{x\in X} f_{m_1,\ldots,m_K}(x) = \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i} w_{b(x)}\left\|\frac{b(x)_i}{w_{b(x)}}-m_{k,i}\right\|^2 \\
&= \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i}\left[\frac{||b(x)_i||^2}{w_{b(x)}}-2b(x)_i m_{k,i}^\top+w_{b(x)}||m_{k,i}||^2\right] \\
&= \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i}\frac{||b(x)_i||^2}{w_{b(x)}}-\sum_{i=1}^{r}\sum_{k=1}^{K}w_{C_k}^{(i)}||m_{k,i}||^2 \\
&= \underbrace{\sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i}\frac{||b(x)_i||^2}{w_{b(x)}}}_{(\gamma)} - n\sum_{i=1}^{r}p^{(i)}\sum_{k=1}^{K}\frac{n_{k+}}{w_{C_k}^{(i)}}p_{k+}\sum_{j=1}^{K_i}\left(\frac{p_{kj}^{(i)}}{p_{k+}}\right)^2.
\end{aligned}$$

According to the definition of centroids in K-means clustering, we have $m_{k,i}=\sum_{x\in C_k\cap X_i}b(x)_i/\sum_{x\in C_k\cap X_i}w_{b(x)}$, $m_k=\langle m_{k,1},\ldots,m_{k,r}\rangle$, $p^{(i)}=|X_i|/|X|=n^{(i)}/n$, $n_{k+}^{(i)}=|C_k\cap X_i|$, $w_{C_k}^{(i)}=\sum_{x\in C_k\cap X_i}w_{b(x)}$. By noting that (γ) is a constant, we get the utility function of SEC with incomplete basic partitions and complete the proof. □

APPENDIX G
PROOF OF THEOREM 7

Proof. The weighted K-means iterates the assigning and updating phase. In the assigning phase, each instance is assigned to the nearest centroid and so the objective function decreases. Thus, we analyze the change of objective function during updating phase under the circumstance of SEC with incomplete basic partitions. For any centroid $g=\langle g_1,\ldots,g_k\rangle$, $g_k=\langle g_{k,i},\ldots,g_{k,r}\rangle$, and $g_k\neq m_k$,

$$\Delta=\sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i}w_{b(x)}[||b(x)_i-g_{k,i}||^2-||b(x)_i-m_{k,i}||^2]. \tag{10}$$

According to the Bergman divergence [53], $f(a,b)=||a-b||^2=\phi(a)-\phi(b)-(a-b)^\top\nabla\phi(b)$, where $\phi(a)=||a||^2$, Eq. (10) can be rewritten as follows:

$$\begin{aligned}
\Delta &= \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i}w_{b(x)}[\phi(b(x)_i)-\phi(g_{k,i}) \\
&\quad +(b(x)_i-g_{k,i})^\top\nabla(g_{k,i})-\phi(b(x)_i)+\phi(m_{k,i}) \\
&\quad -(b(x)_i-m_{k,i})^\top\nabla(m_{k,i})] \\
&= \sum_{i=1}^{r}\sum_{k=1}^{K}\sum_{x\in C_k\cap X_i}w_{b(x)}[\phi(m_{k,i})-\phi(g_{k,i}) \\
&\quad +(b(x)_i-g_{k,i})^\top\nabla(g_{k,i})] \\
&= \sum_{i=1}^{r}\sum_{k=1}^{K}w_{C_k}^{(i)}||m_{k,i}-g_{k,i}||^2 > 0.
\end{aligned}$$

Hence, the objective value will decrease during the update phase as well. Given the finite solution space, the iteration will converge within finite steps. We complete the proof. □

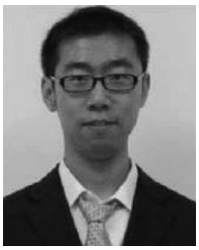
ACKNOWLEDGMENTS

This work was partially supported by NSFC (71531001, 71322104, 71471009, U1636210, 71490723, and 71171007), the National High Technology Research and Development Program of China (SS2014AA012303), the National Center for International Joint Research on E-Business Information Processing (2013B01035), and the Fundamental Research Funds for the Central Universities. Dr. Dacheng Tao's work was supported by Australian Research Council Projects (FT-130101457, DP-140102164, LP-150100671). Dr. Yun Fu's work was supported by US National Science Foundation CNS award (1314484) and US National Science Foundation IIS award (1651902). Junjie Wu is the corresponding author.

REFERENCES

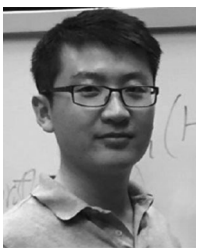
[1] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining partitions," *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2003.
[2] N. Nguyen and R. Caruana, "Consensus clusterings," in *Proc. 7th IEEE Int. Conf. Data Mining*, 2007, pp. 607–612.
[3] X. Wang, C. Yang, and J. Zhou, "Clustering aggregation by probability accumulation," *Pattern Recognit.*, vol. 42, pp. 668–675, 2009.
[4] F. Wang, X. Wang, and T. Li, "Generalized cluster aggregation," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2009, pp. 1279–1284.
[5] A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.
[6] A. Lourenco, S. Bulò, N. Rebagliati, A. Fred, M. Figueiredo, and M. Pelillo, "Probabilistic consensus clustering using evidence accumulation," *Mach. Learn.*, vol. 98, pp. 331–357, 2015.
[7] X. Fern and C. Brodley, "Solving cluster ensemble problems by bipartite graph partitioning," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004, Art. no. 36.
[8] H. Ayad and M. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 160–173, Jan. 2008.
[9] R. Fischer and J. Buhmann, "Path-based clustering for grouping of smooth curves and texture segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 4, pp. 513–518, Apr. 2003.
[10] C. Domeniconi and M. Al-Razgan, "Weighted cluster ensembles: Methods and analysis," *ACM Trans. Knowl. Discovery Data*, vol. 2, 2009, Art. no. 17.
[11] A. Topchy, A. Jain, and W. Punch, "Combining multiple weak clusterings," in *Proc. 3rd IEEE Int. Conf. Data Mining*, 2003, Art. no. 331.
[12] J. Wu, H. Liu, H. Xiong, and J. Cao, "A theoretic framework of K-means-based consensus clustering," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1799–1805.
[13] T. Li, D. Chris, and I. Michael, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in *Proc. 7th IEEE Int. Conf. Data Mining*, 2007, pp. 577–582.
[14] A. Topchy, A. Jain, and W. Punch, "A mixture model for clustering ensembles," in *Proc. 4th SIAM Conf. Data Mining*, 2004, pp. 379–390.
[15] Z. Lu, Y. Peng, and J. Xiao, "From comparing clusterings to combining clusterings," in *Proc. 23rd Nat. Conf. Artif. Intell.*, 2008, pp. 665–670.
[16] S. Xie, J. Gao, W. Fan, D. Turaga, and P. Yu, "Class-distribution regularized consensus maximization for alleviating overfitting in model combination," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 303–312.
[17] H. Liu, T. Liu, J. Wu, D. Tao, and Y. Fu, "Spectral ensemble clustering," in *Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 715–724.
[18] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 155–169, Jan. 2015.
[19] H. Liu, J. Wu, D. Tao, Y. Zhang, and Y. Fu, "DIAS: A disassemble-assemble framework for highly sparse text clustering," in *Proc. SIAM Int. Conf. Data Mining*, 2015, pp. 766–774.
[20] H. Liu, M. Shao, S. Li, and Y. Fu, "Infinite ensemble for image clustering," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1745–1754.
[21] S. Vega-Pons, J. Correa-Morris, and J. Ruiz-Shulcloper, "Weighted partition consensus via kernels," *Pattern Recognit.*, vol. 43, pp. 2712–2724, 2010.
[22] D. Huang, J. Lai, and C. Wang, "Robust ensemble clustering using probability trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1312–1326, May 2016.
[23] D. Huang, J. Lai, and C. Wang, "Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis," *Neurocomputing*, vol. 170, pp. 240–250, 2015.
[24] A. Loureno, S. Bul, A. Fred, and M. Pelillo, "Consensus clustering with robust evidence accumulation," in *Proc. 9th Int. Conf. Energy Minimization Methods Comput. Vis. Pattern Recognit.*, 2013, pp. 307–320.
[25] Z. Tao, H. Liu, S. Li, and Y. Fu, "Robust spectral ensemble clustering," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 367–376.
[26] Z. Tao, H. Liu, and Y. Fu, "Simultaneous clustering and ensemble," in *Proc. Nat. Conf. Artif. Intell.*, 2017.
[27] S. Vega-Pons and J. Ruiz-Shulcloper, "A survey of clustering ensemble algorithms," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 25, 2011, Art. no. 337.
[28] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. 4th IEEE Int. Conf. Data Mining*, 2004, pp. 19–26.
[29] A. Kumar and H. Daume, "A co-training approach for multi-view spectral clustering," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 393–400.
[30] M. Blaschko and C. Lampert, "Correlational spectral clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[31] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan, "Multi-view clustering via canonical correlation analysis," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 129–136.
[32] A. Singh and G. Gordon, "Relational learning via collective matrix factorization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, 650–658.
[33] X. Cai, F. Nie, and H. Huang, "Multi-view K-means clustering on big data," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2598–2604.
[34] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1413–1421.
[35] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 252–260.
[36] S. Li, Y. Jiang, and Z. Zhou, "Partial multi-view clustering," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 1968–1974.
[37] D. Zhang, F. Wang, C. Zhang, and T. Li, "Multi-view local learning," in *Proc. 23rd Nat. Conf. Artif. Intell.*, 2008, pp. 752–757.
[38] X. Wang, B. Qian, J. Ye, and I. Davidson, "Multi-objective multi-view spectral clustering via Pareto optimization," in *Proc. SIAM Int. Conf. Data Mining*, 2013, pp. 234–242.
[39] T. Xia, D. Tao, T. Mei, and Y. Zhang, "Multiview spectral embedding," *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, vol. 40, no. 6, pp. 1438–1446, Dec. 2010.
[40] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probab.*, 1967, pp. 281–297.
[41] S. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, Art. no. 313.
[42] I. Dhillon, Y. Guan, and B. Kulis, "Kernel K-means: Spectral clustering and normalized cuts," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 551–556.
[43] B. Mirkin, "Reinterpreting the category utility function," *Mach. Learn.*, vol. 45, pp. 219–228, 2001.
[44] H. Xu and S. Mannor, "Robustness and generalization," *Mach. Learn.*, vol. 86, pp. 391–423, 2012.
[45] T. Liu, D. Tao, and D. Xu, "Dimensionality-dependent generalization bounds for k-dimensional coding schemes," *Neural Comput.*, vol. 28, no. 10, pp. 2213–2249, 2016.

[46] G. Biau, L. Devroye, and G. Lugosi, "On the performance of clustering in Hilbert spaces," *IEEE Trans. Inf. Theory*, vol. 54, no. 2, pp. 781–790, Feb. 2008.
[47] P. Bartlett, T. L. T, and G. Lugosi, "The minimax distortion redundancy in empirical quantizer design," *IEEE Trans. Inf. Theory*, vol. 44, no. 5, pp. 1802–1813, Sep. 1998.
[48] J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for K-means clustering," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 877–886.
[49] S. Mendelson, "A few notes on statistical learning theory," in *Advanced Lectures on Machine Learning*. Berlin, Germany: Springer, 2003.
[50] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.
[51] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. Berlin, Germany: Springer, 2013.
[52] D. Luo, C. Ding, H. Huang, and F. Nie, "Consensus spectral clustering in near-linear time," in *Proc. IEEE 27th Int. Conf. Data Eng.*, 2011, pp. 1079–1090.
[53] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.



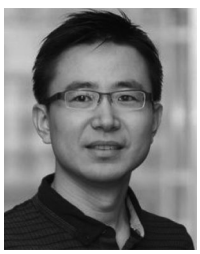
Hongfu Liu received the bachelor's and master's degrees in management information systems from the School of Economics and Management, Beihang University, in 2011 and 2014, respectively. He is currently working toward the PhD degree at Northeastern University, Boston. His research interests generally focus on data mining and machine learning, with special interests in ensemble learning.

Junjie Wu received the BE degree in civil engineering and the PhD degree in management science and engineering from Tsinghua University. He is currently a full professor in the Information Systems Department, School of Economics and Management, Beihang University, the director of the Social Computing Center, and the vice director of the Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations. His general area of research is data mining and complex networks, with special interests in social media analytics and quantitative financial analytics. He received the various national awards including NSFC Excellent Youth, MOE Changjiang Young Scholars, and MOE Excellent Doctoral Dissertation.



Tongliang Liu received the BE degree from the University of Science and Technology of China, Hefei, China, in 2012, and the PhD degree from the University of Technology Sydney, Sydney, Australia, in 2016. He is currently a lecturer in the Centre for Artificial Intelligence, School of Software, Faculty of Engineering and Information Technology, University of Technology Sydney. Previously, he was a visiting PhD student in the Barcelona Graduate School of Economics (Barcelona GSE) and in the Department of Economics, Pompeu Fabra University. His research interests include statistical learning theory, computer vision, and optimization. He has more than 20 research papers that have been published in leading journals and international conferences/workshops including the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Neural Networks and Learning Systems*, the *IEEE Transactions on Image Processing*, NECO, ICML, KDD, IJCAI, and AAAI. He won the Best Paper Award at the IEEE International Conference on Information Science and Technology 2014.

Dacheng Tao (F'15) is a professor of computer science in the School of Information Technologies, Faculty of Engineering and Information Technologies, University of Sydney. He was professor of computer science and director of Centre for Artificial Intelligence, University of Technology Sydney. He mainly applies statistics and mathematics to Artificial Intelligence and Data Science. His research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and more than 200 publications at prestigious journals and prominent conferences, such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Neural Networks and Learning Systems*, the *IEEE Transactions on Image Processing*, the *Journal of Machine Learning Research*, the *International Journal of Computer Vision*, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM, and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award at IEEE ICDM'07, the best student paper award at IEEE ICDM'13, and the 2014 ICDM 10-year highest-impact paper award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award, and the 2015 UTS Vice-Chancellor's Medal for Exceptional Research. He is a fellow of the IEEE, the OSA, the IAPR, and the SPIE.



Yun Fu (S'07-M'08-SM'11) received the BEng degree in information engineering and the MEng degree in pattern recognition and intelligence systems from Xi'an Jiaotong University, China, respectively, and the MS degree in statistics and the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, respectively. He is an interdisciplinary faculty member affiliated in the College of Engineering, College of Computer and Information Science, Northeastern University, since 2012. His research interests include machine learning, computational intelligence, big data mining, computer vision, pattern recognition, and cyber-physical systems. He has extensive publications in leading journals, books/book chapters, and international conferences/workshops. He serves as an associate editor, chairs, PC member, and reviewer of many top journals and international conferences/workshops. He received seven Prestigious Young Investigator Awards from NAE, ONR, ARO, IEEE, INNS, UIUC, Grainger Foundation; seven Best Paper Awards from IEEE, IAPR, SPIE, SIAM; and three major Industrial Research Awards from Google, Samsung, and Adobe, etc. He is currently an associate editor of the *IEEE Transactions on Neural Networks and Leaning Systems*. He is a fellow of the IAPR, a lifetime senior member of the ACM and the SPIE, lifetime member of the AAAI, OSA, and Institute of Mathematical Statistics, member of Global Young Academy (GYA), INNS, and Beckman Graduate Fellow during 2007-2008. He is a senior member of the IEEE.