Spectral Ensemble Clustering

Hongfu Liu[1], Tongliang Liu[2], Junjie Wu[3*], Dacheng Tao[2], Yun Fu[1]
[1] Department of Electrical & Computer Engineering, College of Engineering, Northeastern University, Boston. liu.hongf@husky.neu.edu, yunfu@ece.neu.edu.
[2] QCIS and FEIT, University of Technology, Sydney. tliang.liu@gmail.com (co-first author), dacheng.tao@uts.edu.au.
[3] School of Economics and Management, Beihang University, Beijing. wujj@buaa.edu.cn.
[*] corresponding author

ABSTRACT

Ensemble clustering, also known as consensus clustering, is emerging as a promising solution for multi-source and/or heterogeneous data clustering. The co-association matrix based method, which redefines the ensemble clustering problem as a classical graph partition problem, is a landmark method in this area. Nevertheless, the relatively high time and space complexity preclude it from real-life large-scale data clustering. We therefore propose SEC, an efficient Spectral Ensemble Clustering method based on co-association matrix. We show that SEC has theoretical equivalence to weighted K-means clustering and results in vastly reduced algorithmic complexity. We then derive the latent consensus function of SEC, which to our best knowledge is among the first to bridge co-association matrix based method to the methods with explicit object functions. The robustness and generalizability of SEC are then investigated to prove the superiority of SEC in theory. We finally extend SEC to meet the challenge rising from incomplete basic partitions, based on which a scheme for big data clustering can be formed. Experimental results on various real-world data sets demonstrate that SEC is an effective and efficient competitor to some state-of-the-art ensemble clustering methods and is also suitable for big data clustering.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Ming*

General Terms

Algorithm, Theory

Keywords

Ensemble Clustering, Co-association Matrix, K-means

1. INTRODUCTION

Ensemble clustering, also known as consensus clustering, is emerging as a promising solution for multi-source and/or heterogeneous data clustering and is attracting increasing academic attention. It aims to find a single partition that most agrees with multiple existing basic partitions [22]. Consensus clustering is of recognized benefit in generating robust partitions, finding bizarre clusters, handling noise and outliers, and integrating solutions from multiple sources [21].

Consensus clustering can be roughly divided into two categories: those with implicit or explicit objectives. Methods that utilize implicit objectives do not set objective functions, but instead directly adopt some heuristics to find approximate solutions. Representative methods include co-association matrix-based methods [27, 26, 10, 14], graph-based algorithms [22, 8], relabeling and voting methods [1, 9, 19], locally adaptive cluster-based methods [7], and genetic algorithm-based methods [32]. Methods with explicit objectives employ explicit objective functions for consensus clustering. For instance, [23] used K-means to find the solution based on quadratic entropy, which was then generalized in [28] as the paradigm of K-means-based consensus clustering. Other solutions for different objective functions include non-negative matrix factorization [12], EM algorithm [24], simulated annealing [15], combination regularization [30], and hill-climbing method [5]. Many other algorithms for consensus clustering can be found in the survey [13, 25, 20].

Of these consensus clustering methods, the co-association matrix-based method is a landmark. First, the information represented by basic partitions is summarized into a co-association matrix, which measures how many times a pair of instances appears in the same cluster; then a graph partition method can be used to obtain the final consensus clustering. The main contribution of the co-association method is the redefinition of the consensus clustering problem as a classical graph partition problem, so that agglomerative hierarchical clustering, spectral clustering, or other algorithms can directly run on the co-association matrix without much modification. However, the co-association matrix-based method also suffers from some limitations. For instance, the high time and space complexity prevents handling large-scale data clustering and no explicit objective function is used to supervise clustering.

In light of this, we propose Spectral Ensemble Clustering (SEC), which employs spectral clustering on the co-association matrix to find the consensus partition. SEC equivalently results in weighted K-means clustering, which decreases the time complexity from $\mathcal{O}(n^3)$ to roughly $\mathcal{O}(n)$ and decreases the space complexity from $\mathcal{O}(n^2)$ to roughly $\mathcal{O}(n)$ as well. We then derive the intrinsic consensus ob-

KDD '15 Sydney, Australia
ACM 978-1-4503-3664-2/15/08?$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2783287.

jective of SEC and provide a robustness and generalization analysis. Further we extend SEC to handle incomplete basic partitions. Experimental results on various real-world data sets demonstrate that SEC delivers efficient and high quality clustering compared to some state-of-the-art consensus clustering methods. SEC is also highly robust to incomplete basic partitions with many missing values. Finally, SEC is used to explore big data clustering of Weibo data.

2. SPECTRAL ENSEMBLE CLUSTERING

Let $\mathcal{X} \in \mathcal{R}^{n\times d}$ denote the data matrix with n instances and d dimensions, and given r basic crisp partitions of $\mathcal{X}$ (a basic partition is a partition of $\mathcal{X}$ given by running some clustering algorithm) in $\Pi = \{\pi_1, \pi_2, \cdots, \pi_r\}$, a co-association matrix [10] $\mathbf{S}$ is built as follows:

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

Similarity between a pair of instances simply counts the co-occurrence number in the same cluster in Π. Spectral Ensemble Clustering (SEC) applies spectral clustering on the co-association matrix $\mathbf{S}$ to obtain final clustering π.

Let $\mathbf{H} = [\mathbf{H}_1, \cdots, \mathbf{H}_K]$ be a $n \times K$ partition matrix, where K is the user-specific cluster number, and $\mathbf{D}$ is the diagonal matrix whose diagonal entry is the sum of rows of $\mathbf{S}$. The objective function of normalized cuts spectral clustering is the following trace maximization problem [33]:

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

where $\mathbf{Z} = \mathbf{D}^{-1/2}\mathbf{H}(\mathbf{H}^\top \mathbf{D}\mathbf{H})^{-1/2}$. A well-known solution to Eq. 1 is to obtain the top k eigenvectors of $\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ and run K-means to get the final consensus clustering π.

2.1 From SEC to Weighted K-means

Performing spectral clustering on co-association matrix also has high time complexity. We here propose a low time-cost algorithm for this purpose; that is, we transform SEC into weighted K-means clustering via a binary matrix.

Let $\mathbf{B} = \{b(x)\}$ be a binary data set derived from the set of r basic partitionings Π as follows:

$$b(x) = \langle b(x)_1, \cdots, b(x)_r \rangle, b(x)_i = \langle b(x)_{i1}, \cdots, b(x)_{iK_i} \rangle,$$
$$b(x)_{ij} = \begin{cases} 1, & \text{if } \pi_i(x) = j \\ 0, & \text{otherwise} \end{cases},$$

where "$\langle\ \rangle$" indicates a transverse vector. Therefore, $\mathbf{B}$ is an $n \times \sum_{i=1}^{r} K_i$ binary data matrix with $|b(x)_i| = 1$, K_i is the number of clusters of π_i, $\forall\, l, i$. After introducing the binary matrix, we obtain the theorem to connect SEC and classical weighted K-means clustering. All the proofs are given in the appendix for concision.

THEOREM 1. *Given a set of basic partitions Π, the co-association matrix-based method with spectral clustering has the equivalent objective function to classical weighted K-means clustering such that*

$$\max \frac{1}{K} \mathrm{tr}(\boldsymbol{Z}^\top \boldsymbol{D}^{-1/2} \boldsymbol{S} \boldsymbol{D}^{-1/2} \boldsymbol{Z}) \Leftrightarrow \sum_{i=1}^{n} f_{m_1,\ldots,m_K}(x_i), \qquad (2)$$

where $f_{m_1,\ldots,m_K}(x) = \min_k w_{b(x)} \| \frac{b(x)}{w_{b(x)}} - m_k \|^2$, $diag(w_{b(x)}) = \boldsymbol{D}$, and $m_k = \frac{\sum_{x\in C_k} b(x)}{\sum_{x\in C_k} w_{b(x)}}$.

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

REMARK 1. *Theorem 1 transforms SEC into weighted K-means clustering in an explicit theoretically equivalent way. Weighted K-means is conducted on an $n \times \sum_{i=1}^{r} K_i$ data matrix. Recall that there is only one non-zero element in $b(x)_i$, the time complexity is $\mathcal{O}(InrK)$, where I is the number of iterations. Thus, the time complexity decreases from $\mathcal{O}(n^3)$ to roughly $\mathcal{O}(n)$, and the space complexity decreases from $\mathcal{O}(n^2)$ to roughly $\mathcal{O}(n)$ as well.*

REMARK 2. *Unlike [6] which built the connection between spectral clustering and weighted kernel K-means, here we equivalently have figured out the mapping function of the kernel, which turns out to be the binary data dividing its corresponding weight. By this means, we transfer SEC to weighted K-means rather than weighted kernel K-means.*

2.2 Intrinsic Consensus Objective Function

To understand the objective function of SEC in the partition level, here we derive the intrinsic consensus objective function of SEC from weighted K-means. In Table 1, given two partitions π and π_i containing K and K_i clusters, respectively, let $n_{kj}^{(i)}$ denote the number of data objects belonging to both cluster $C_j^{(i)}$ in π_i and cluster C_k in π, $n_{k+} = \sum_{j=1}^{K_i} n_{kj}^{(i)}$, and $n_{+j}^{(i)} = \sum_{k=1}^{K} n_{kj}^{(i)}$, $1 \le j \le K_i$, $1 \le k \le K$. Let $p_{kj}^{(i)} = n_{kj}^{(i)}/n$, $p_{k+} = n_{k+}/n$, and $p_{+j}^{(i)} = n_{+j}^{(i)}/n$. We then have a normalized contingency matrix (NCM), based on which a wide range of utility functions can be accordingly defined. For instance, the well-known category utility function [17] can be computed as follows:

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

Based on NCM, we have the following theorem.

THEOREM 2. *If a utility function has the same or equivalent formation as follows,*

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

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

$$\max \frac{1}{K} tr(\boldsymbol{Z}^T \boldsymbol{D}^{-1/2} \boldsymbol{S} \boldsymbol{D}^{-1/2} \boldsymbol{Z}) \Leftrightarrow \max \sum_{i=1}^{r} U(\pi, \pi_i). \qquad (5)$$

REMARK 3. *The same or equivalent formulation of U in Eq. 4 can be used as a utility function to supervise the consensus process. It is worth noting that the category utility function is a special case of the SEC utility function U with all weights one. Recall that the co-association matrix measures similarity in instance level; by transforming SEC into*

classical weighted K-means, we derive the utility function to measure the similarity in partition level, i.e., the two kinds of similarity in different levels can be interconvertible.

3. ROBUSTNESS AND GENERALIZATION

3.1 Robustness

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

DEFINITION 1 (ROBUSTNESS [31]). *Let $\mathcal{X}^n$ be the training sample. 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 $s \in \mathcal{X}^n, \forall x_s \in s, \forall x \in \mathcal{X}, \forall i = 1, ..., K$: if $x_s, x \in C_i$, then $|\ell(h_s, x_s) - \ell(h_s, x)| \leq \epsilon(s)$.*

We then have Theorem 3 to measure SEC's robustness as follows.

THEOREM 3. *Let $\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_2)$ be a covering number of $\mathcal{X}$. For any $x, x' \in \mathcal{X}, \|x - x'\|_2 \leq \gamma$, we define $\|b(x)_i - b(x')_i\|_2 \leq \gamma_i$ and $|w^i_{b(x)} - w^i_{b(x')}| \leq \gamma^i_w, i = 1, \ldots, r$, where $w^i_{b(x)} = \sum_{l=1}^n \delta(\pi_i(x), \pi_i(x_l))$. Then, for any centroids m_1, ..., m_K learned by SEC, we have*

$$|f_{m_1,\ldots,m_K}(x) - f_{m_1,\ldots,m_K}(x')| \leq \frac{2\sum_{i=1}^r \gamma^i_w}{r} + \sqrt{\frac{\sum_{i=1}^r \gamma_i^2}{r}}. \tag{6}$$

Hence, SEC is $(\mathcal{N}(\gamma, \mathcal{X}, \|\cdot\|_2), \frac{2\sum_{i=1}^r \gamma^i_w}{r} + \sqrt{\frac{\sum_{i=1}^r \gamma_i^2}{r}})$-robust.

REMARK 4. *From Theorem 3, we can see that even if some instances are "poorly" clustered by some basic partitions and cause the corresponding margins $\{\gamma_i\}$ and $\{\gamma^i_w\}$ to be large, the good overall performance of SEC will be preserved, provided that these instances are "well" clustered by many other basic partitions. This means that SEC could benefit from the ensemble of basic partitions, namely that instances could be "well" clustered by the majority of the basic partitions.*

3.2 Generalizability

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

THEOREM 4. *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*

$$\begin{aligned} &E_x f_{m_1,\ldots,m_K}(x) - \frac{1}{n}\sum_{i=1}^n f_{m_1,\ldots,m_K}(x_i) \\ &\leq \frac{\sqrt{2\pi}rK}{n}(\sum_{i=1}^n w^{-2}_{b(x_i)})^{\frac{1}{2}} + \frac{\sqrt{8\pi}rK}{\sqrt{n}\min_{x\in\{x_1,\ldots,x_n\}} w_{b(x)}} \\ &+ \frac{\sqrt{2\pi}rK}{n\min_{x\in\{x_1,\ldots,x_n\}} w^2_{b(x)}}(\sum_{i=1}^n w^2_{b(x_i)})^{\frac{1}{2}} + (\frac{\ln(1/\delta)}{2n})^{\frac{1}{2}}. \end{aligned} \tag{7}$$

REMARK 5. *Theorem 4 shows that if the third term of the upper bound goes to zero as n goes to infinity, the empirical reconstruction error of SEC will go to its expected reconstruction error. So, the convergence of*

$$\frac{\sqrt{2\pi}rK}{n}(\sum_{i=1}^n w^2_{b(x_i)})^{\frac{1}{2}} \frac{1}{\min_{x\in\{x_1,\ldots,x_n\}} w^2_{b(x)}}$$

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

REMARK 6. *The consistency of the basic crisp partitions will make $w_{b(x_i)}/|C^i_k|$ diverge little, where $|C^i_k|$ denotes the cardinality of the cluster in which x_i belongs. 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 $\mathcal{O}(1/\sqrt{n^3})$. However, for classical K-means clustering, the fastest known convergence rate is $\mathcal{O}(1/\sqrt{n})$ [4, 3]. The fast convergence rate will result in the expected risk of the learned partitioning decreasing quickly to the expected risk of the target partitioning [4]. This verifies the efficiency of SEC.*

4. INCOMPLETE EVIDENCE

In practice, Incomplete Basic Partitions (IBP) are often obtained due to distributed systems or missing data. An incomplete basic partition π_i is obtained by clustering a data subset $\mathcal{X}_i \subseteq \mathcal{X}$, $1 \leq i \leq r$, with the constraint that $\bigcup_{i=1}^r \mathcal{X}_i = \mathcal{X}$. Here, the problem is how to cluster $\mathcal{X}$ into K crisp clusters using SEC given r IBPs in $\Pi = \{\pi_1, \cdots, \pi_r\}$. The co-association matrix built using incomplete basic partitions cannot continue to represent the similarity of points. To address this limitation, we start with the objective of weighted K-means clustering and exploit it for handling incomplete basic partitions. Obviously, missing elements in basic partitions provide no utility in the ensemble process. Consequently, these missing elements do not contribute to the centroid; thus, let $|\mathcal{X}_i| = n^{(i)}$, we have

THEOREM 5. *Given r incomplete basic partitions, we have*

$$\sum_{i=1}^n f_{m_1,\ldots,m_K}(x_i) \Leftrightarrow \max \sum_{i=1}^r p^{(i)} \sum_{k=1}^K \frac{n^{(i)}_{k+}}{w^{(i)}_{C_k}} p_{k+} \sum_{j=1}^{K_i} (\frac{p^{(i)}_{kj}}{p_{k+}})^2, \tag{8}$$

where $f_{m_1,\ldots,m_K}(x) = \min_k w_{b(x)} \| \frac{b(x)}{w_{b(x)}} - m_k \|^2$, with $m_k = \langle m_{k,1}, \cdots, m_{k,r} \rangle$, $m_{k,i} = \sum_{x\in C_k\cap\mathcal{X}_i} b(x)_i / \sum_{x\in C_k\cap\mathcal{X}_i} w_{b(x)}$, $p^{(i)} = n^{(i)}/n$, $n^{(i)}_{k+} = |C_k \cap \mathcal{X}_i|$, $w^{(i)}_{C_k} = \sum_{x\in C_k\cap\mathcal{X}_i} w_{b(x)i}$.

REMARK 7. *In Theorem 5, the utility function of SEC on IBPs has one more parameter $p^{(i)}$ than that on complete basic partitions, indicating that basic partitions with more elements are naturally assigned more importance in the ensemble process. However, the co-association matrix only reflects the summation information of basic partitions. On the surface, it seems that each point plays an equal role; in fact, by Theorem 5 we find that these points contribute differently depending on the missing rate of their basic partitions.*

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

THEOREM 6. *For the objective function in Theorem 5, SEC with IBPs is guaranteed to converge in finite two-phase iterations of weighted K-means clustering.*

Table 2: Experimental Data Sets

Data set	Source	#Instances	#Features	#Classes
breast_w	UCI	699	9	2
iris	UCI	150	4	3
wine	UCI	178	13	3
cacmcisi	CLUTO	4663	14409	2
classic	CLUTO	7094	41681	4
cranmed	CLUTO	2431	41681	2
hitech	CLUTO	2301	126321	6
k1b	CLUTO	2340	21839	6
la12	CLUTO	6279	31472	6
mm	CLUTO	2521	126373	2
re1	CLUTO	1657	3758	25
reviews	CLUTO	4069	126373	5
sports	CLUTO	8580	126373	7
tr11	CLUTO	414	6429	9
tr12	CLUTO	313	5804	8
tr41	CLUTO	878	7454	10
tr45	CLUTO	690	8261	10
letter	LIBSVM	20000	16	26
mnist	LIBSVM	70000	784	10

5. BIG DATA CLUSTERING USING SEC

Consensus clustering seems not a good first choice for big data clustering; because r basic partitions are first produced, then ensemble clustering is used to obtain the final clustering and this requires high time and space cost. In terms of large scale data, it is difficult to obtain basic partitions and perform the ensemble process. However, SEC with incomplete basic partitions enables big data processing.

In terms of large-scale data clustering, we propose a type of row-segmentation strategy. Generally speaking, we randomly select a certain percentage of data instances to obtain a data subset and run K-means to obtain the label from 1 to K; those unsampled data are labeled "0". This selection process is repeated r times to obtain the IBPs, prior to using the SEC algorithm to complete the clustering. The benefit of the row-segmentation strategy is two-fold: first, a big data set is decomposed into several smaller ones, which can be handled separately and independently; second the high dimensionality of the data is decreased to only r dimensions. The experimental results in the next section demonstrate that the row-segmentation strategy even outperforms directly clustering on the whole data.

6. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of SEC on various real-world data sets from different domains and compare it with state-of-the-art consensus clustering algorithms.

6.1 Experimental Setup

Data. Various real-world data sets with true cluster labels are used for experiments. Table 2 summarizes their important characteristics. Three data sets are used: the UCI Machine Learning Repository[1], CLUTO[2] and LIBSVM[3].

Tool. SEC is completely coded in Matlab. The *kmeans* function with square Euclidean distance (UCI and LIBSVM data sets) and cosine similarity (CLUTO data sets) are run 100 times to obtain $r = 100$ basic partitions by varying the cluster number from the true cluster number K to $\sqrt{n}$. The comparative methods included consensus clustering with category utility function (CCC) [28], graph-based consensus clustering (GCC) [22], and co-association matrix with agglomerative hierarchical clustering (HCC) [10]. Note that the cluster number for each algorithm is set to the true cluster number K and that each algorithm runs 10 times and returns the average result.

Validity. Since true cluster labels are available for all data sets, we employed external measures to assess cluster validity. Although accuracy is a measure used in classification, here it might not be suitable due to the unknown mapping relationship between the true cluster labels and clustering result. Thus, we choose the widely used R_n [29] for cluster validity. R_n is a positive measure and a larger value indicates a better performance.

6.2 Effectiveness and Efficiency of SEC

Here, we illustrate the performance of SEC with some well-established methods for consensus clustering. The clustering results of these methods are shown in Table 3, with the best results highlighted in bold and the last column showing the baseline K-means results.

For most data sets, consensus clustering algorithms enjoy superior performance to single K-means clustering. Among consensus clustering algorithms, SEC is a powerful competitor and achieves the best performance 14 times, and the second-best performance 4 times of 19 data sets. GP produces the best results of three graph-based consensus clustering algorithms (CSPA, HGPA, and MCLA) and performs particularly well on some balanced datasets, such as *iris* and *cranmed*. Although HCC obtains satisfactory results on *k1b* and *sports*, the performance on *cacmcisi* and *mm* is extremely poor, indicating that HCC, which has no utility function to supervise the combining process, is less robust than the other algorithms. CCC has similar utility function to SEC; however, by enforcing the weights of the instances in large clusters, SEC outperforms CCC in most cases. Note that the negative results, such as the *cacmcisi* via CCC and HCC, indicate that the clustering results are worse than random labeling.

Table 4 shows the average time of ten runs via these methods. In terms of efficiency, K-means-based consensus clustering has obvious advantages over other methods. SEC and CCC are tied, and HCC might be suitable for small datasets but struggles as the number of instances increases. Note that these experiments are all run on the same computer.

6.3 Stability and Scalability of SEC

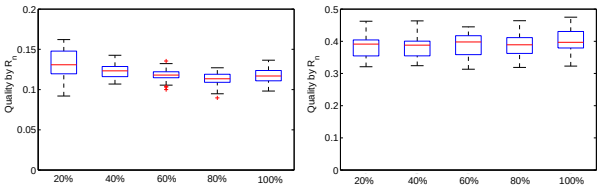
To further validate the performance on large-scale data, in this part we test the stability and scalability of SEC on *letter* and *mnist* with 20,000 and 70,000 instances, respectively. The basic partitioning generation strategy varying the cluster number from K to $\sqrt{n}$ seems not suitable for large-scale data due to the large n; thus we change the variation range of cluster number from 2 to $2K$ to generate 100 basic partitions. To test the scalability of SEC, we have the assumption that if an algorithm can scale down and work well on the sample data, the algorithm can also scale up and enjoy a good scalability. Here we conduct stratified sample according to the true cluster labels with the ratio from 20% to 100% with the 20% step and repeat 50 times on each sampling ratio. Note that each run only calls one time SEC to see the stability of SEC as well.

From Figure 1, the performance of SEC on each sampling ratio keeps still even on only 20% instances. This indicates that SEC has a good scalability and suits for big data clus-

[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/.

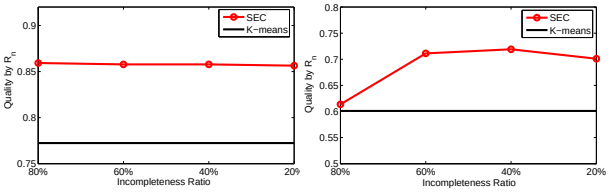
Table 3: Clustering Results (by R_n)

Data set	SEC	CCC	GCC	HCC	K-means
breast_w	**0.8230**	0.0556	0.6115	0.7809	0.8391
iris	**0.9222**	0.7352	**0.9222**	0.7323	0.6998
wine	**0.3272**	0.1448	0.1448	0.1490	0.1283
cacmcisi	**0.6388**	-0.0447	0.3177	-0.0298	-0.0310
classic	**0.7069**	0.4224	0.3826	0.3828	0.2928
cranmed	0.9512	**0.9560**	0.7482	0.9352	0.9496
hitech	**0.3047**	0.2227	0.1604	0.2676	0.2530
k1b	0.6236	0.5288	0.2677	**0.6351**	0.4469
la12	**0.5311**	0.3455	0.3571	0.3647	0.2151
mm	**0.6184**	0.5450	0.3753	-0.0006	0.7724
re1	**0.2851**	0.2630	0.2299	0.2788	0.2382
reviews	**0.5036**	0.3767	0.3942	0.4589	0.6011
sports	0.4652	0.3211	0.2911	**0.4840**	0.3345
tr11	**0.5926**	0.5217	0.3755	0.5896	0.4050
tr12	**0.4701**	0.4219	0.4673	0.4485	0.4073
tr41	**0.4578**	0.3839	0.3649	0.4280	0.3568
tr45	**0.4688**	0.3947	0.3805	0.4554	0.4096
letter	0.1202	0.1225	**0.1350**	0.1179	0.1185
mnist	0.4027	0.3795	0.3678	**0.4482**	0.3912



(a) *letter* (b) *mnist*

Figure 1: Stability and Scalability of SEC with Different Sampling Ratios



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

Figure 2: Performance of SEC with Different Incompleteness Ratios

tering. Beside we can also see that on *letter* the volatility even shrinks when then sampling ratio makes larger, which demonstrates that SEC enjoys good stability as well.

6.4 Performances of SEC with Incompleteness

Next we demonstrate the performance of SEC with IBPs. The row-segmentation strategy is used to generate IBPs. The instances are first randomly sampled from 20% to 80%, prior to calling *kmeans* on the data subset to generate the basic partitions. Unsampled instances are assigned the "0" label. The above process repeats 100 times to obtain 100 IBPs. Then SEC is used to ensemble these incomplete basic partitions and obtain the consensus clustering result.

The results of *mm* and *reviews* are shown in Figure 2. SEC obtains a stably high performance on *mm*, even at low percentage (20%), and in *reviews* with increasing percent random selection the performance of SEC also improves. The black line represents the K-means result with all instances, and it can be seen that SEC (with different random selection percentages) outperforms K-means. This indicates SEC appears to be highly robust to incomplete data and provides a way to handle big data clustering with incomplete evidence via the row-segmentation strategy.

Table 4: Run Time (by second)

Data set	SEC	CCC	GCC	HCC
breast_w	**0.10**	0.14	0.53	3.32
iris	0.04	**0.02**	3.87	0.10
wine	0.05	**0.03**	3.01	0.12
cacmcisi	**0.68**	1.20	54.87	632.14
classic	**1.87**	2.23	96.15	1908.10
cranmed	**0.33**	0.37	22.24	141.91
hitech	**0.70**	0.74	27.76	144.44
k1b	**0.48**	0.56	27.99	130.28
la12	**0.43**	0.45	72.49	1329.20
mm	0.16	**0.15**	16.58	141.08
re1	**0.64**	0.92	31.29	85.60
reviews	**0.27**	0.30	42.68	500.13
sports	**0.58**	0.71	108.85	1697.00
tr11	0.12	**0.11**	7.89	2.74
tr12	**0.09**	**0.09**	6.89	0.99
tr41	**0.19**	0.21	14.92	20.17
tr45	0.16	**0.15**	11.39	10.01
letter	**8.53**	9.07	348.32	1847.39
mnist	**12.95**	13.29	112.33	19995.68

Table 5: Representative Keywords of Weio Clusters

No.	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*

6.5 SEC Applied to Weibo Data Clustering

Sina Weibo[4], a Twitter-like service launched in China in 2009, has accumulated more than 500 million users in less than five years. The Weibo platform produces very large volumes of user generated content; for example, on September 1st 2013, there were 97,231,274 tweets published on the Weibo platform, which provides a valuable data source for commercial applications and academic research. However, conducting cluster analysis on such big data is a difficult data mining task.

Here we employ SEC to cluster the entire Weibo data published on September 1st, 2013 to discover hot events. First, we remove over 30 million advertising tweets and use SCWS[5] to conduct word segmentation; this produces a dataset with 61,212,950 instances and 10,000 features. Next, the row-segmentation strategy is carried out to acquire 100 data subsets with 10,000,000 instances, and 100 IBPs are performed on these data subsets using CLUTO (note that the maximum capacity for CLUTO is 10,000,000 instances). Finally, SEC fuses the partial information into integrated clustering via parallel computing. The cluster number was set to 100 for both basic partitions and final consensus clustering.

To handle such big data, we use a distributed computing cluster with 10 servers to conduct consensus clustering in parallel, and each iteration takes approximately 3 hours. The results of the representative keywords of some clusters are shown in Table 4. Cluster 3, 21, and 83 represent "term begins", "mid-autumn festival", and "travel" events, respectively; Cluster 40 identifies the conflict between China and Japan due to "the September 18th incident"; and Cluster 65 represents the event that Meng Ge, a famous singer in China, apologized for her son's crime. Although the basic partitions are highly incomplete, some interesting events can be found using the row-segmentation strategy. SEC appears to be a good choice for clustering big data.

[4]http://www.weibo.com/.
[5]http://www.xunsearch.com/scws/.

7. CONCLUSIONS

In this paper, we propose SEC and uncover an equivalent relationship with weighted K-means clustering that dramatically decreases the time and space complexity. Based on weighted K-means, the intrinsic consensus objective function of SEC is derived, then we investigate its robustness and generalizability, and extend it to incomplete basic partitionings. Experimental results demonstrate that SEC produces high quality, efficient clustering compared with other state-of-the-art methods, which is further illustrated in an application of big data clustering of Weibo data.

8. ACKNOWLEDGEMENTS

This work was partially supported by National Natural Science Foundation of China (71322104, 71171007, 71471009, 71490720), National Center for International Joint Research on E-Business Information Processing (2013B01035), National High Technology Research and Development Program of China (863 Program) (SS2014AA012303). Dr. Dacheng Tao's work was supported by Australian Research Council Projects (DP-120103730, DP-140102164, FT-130101457). Dr. Yun Fu's work was supported by NSF CNS award (1314484). We thank KDD anonymous reviewers and SPC for their constructive comments, which help to improve this work to a new level in the final revision.

9. REFERENCES

[1] H. Ayad and M. Kamel. Cumulative voting consensus method for partitions with variable number of clusters. *PAMI*, 30(1):160–173, 2008.

[2] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *JMLR*, 2005.

[3] P. Bartlett, T. L. T, and G. Lugosi. The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 1998.

[4] G. Biau, L. Devroye, and G. Lugosi. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory*, 2008.

[5] C. Carpineto and G. Romano. Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. *PAMI*, 2012.

[6] I. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of KDD*, 2004.

[7] C. Domeniconi and M. Al-Razgan. Weighted cluster ensembles: Methods and analysis. *TKDD*, 2009.

[8] X. Z. Fern and C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of ICML*, 2004.

[9] R. Fischer and J. Buhmann. Path-based clustering for grouping of smooth curves and texture segmentation. *PAMI*, 2003.

[10] A. Fred and A. Jain. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005.

[11] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer, 1991.

[12] T. Li, D. Chris, and I. Michael. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. *Proceedings of ICDM*, 2007.

[13] T. Li, M. Ogihara, and S. Ma. On combining multiple clusterings. In *Proceedings of CIKM*, 2004.

[14] A. Lourenco, S. Bulò, N. Rebagliati, A. Fred, M. Figueiredo, and M. Pelillo. Probabilistic consensus clustering using evidence accumulation. *Machine Learning*, 98(1-2):331–357, 2013.

[15] Z. Lu, Y. Peng, and J. Xiao. From comparing clusterings to combining clusterings. In *Proceedings of AAAI*. AAAI Press, 2008.

[16] S. Mendelson. *A few notes on statistical learning theory*. Advanced Lectures on Machine Learning, 2003.

[17] B. Mirkin. Reinterpreting the category utility function. *Machine Learning*, 2001.

[18] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT Press, 2012.

[19] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 2003.

[20] M. Naldi, A. Carvalho, and R. Campello. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, 27(2), 2013.

[21] N. Nguyen and R. Caruana. Consensus clusterings. In *Proceedings of ICDM*, 2007.

[22] A. Strehl and J. Ghosh. Cluster ensembles — a knowledge reuse framework for combining partitions. *Journal of Machine Learning Research*, 2002.

[23] A. Topchy, A. Jain, and W. Punch. Combining multiple weak clusterings. In *Proceedings of ICDM*, 2003.

[24] A. Topchy, A. Jain, and W. Punch. A mixture model for clustering ensembles. In *Proceedings of SDM*, 2004.

[25] S. Vega-Pons and J. Ruiz-Shulcloper. A survey of clustering ensemble algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*, 2011.

[26] F. Wang, X. Wang, and T. Li. Generalized cluster aggregation. In *Proceedings of IJCAI*, 2009.

[27] X. Wang, C. Yang, and J. Zhou. Clustering aggregation by probability accumulation. *Pattern Recognition*, 2009.

[28] J. Wu, H. Liu, H. Xiong, and J. Cao. A theoretic framework of k-means-based consensus clustering. In *Proceedings of IJCAI*, 2013.

[29] J. Wu, H. Xiong, and J. Chen. Adapting the right measures for k-means clustering. In *Proceedings of KDD*, 2009.

[30] S. Xie, J. Gao, W. Fan, D. Turaga, and P. Yu. Class-distribution regularized consensus maximization for alleviating overfitting in model combination. In *Proceedings of KDD*, 2014.

[31] H. Xu and S. Mannor. Robustness and generalization. *Machine learning*, 2012.

[32] H. Yoon, S. Ahn, S. Lee, S. Cho, and J. Kim. Heterogeneous clustering ensemble method for combining different cluster results. *Data Mining for Biomedical Applications*, 2006.

[33] S. Yu and J. Shi. Multiclass spectral clustering. In *Proceedings of ICCV*, 2003.

APPENDIX

For the concision of math description, we let Z_K denote $\{1,\ldots,K\}$ and let X_n denote $\{x_1,\ldots,x_n\}$ in the 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 [6], we have

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

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

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

where $\mathbf{G} = diag(\frac{\mathbf{W}_1^{1/2}\mathbf{e}}{\sqrt{s_1}}, \cdots, \frac{\mathbf{W}_k^{1/2}\mathbf{e}}{\sqrt{s_k}})$. 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. ■

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:

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

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, we get the utility function of SEC. ■

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 $f_{m_1,\ldots,m_K}(x) \le \max\{\frac{\|b(x)\|^2}{w_{b(x)}}, w_{b(x)}\|m_k\|^2\}$. We have

$$\frac{\|b(x)\|^2}{w_{b(x)}} \le \frac{r}{r} = 1,$$

$$w_{b(x)}\|m_k\|^2 \quad = \frac{w_{b(x)}\|\sum_{b_l\in C_k} b_l\|^2}{(\sum_{b_l\in C_k} w_{b_l})^2} \le 1. \tag{9}$$

This concludes the proof.

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

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

The first inequality holds due to the triangle inequality. ■

Now we begins the proof of Theorem 3.

Proof. We have

$$\begin{aligned}
&|f_{m_1,\ldots,m_K}(x) - f_{m_1,\ldots,m_K}(x')| \\
&= |\min_{k\in Z_K} w_{b(x)}\|\frac{b(x)}{w_{b(x)}} - m_k\| - \min_{k\in Z_K} w_{b(x')}\|\frac{b(x')}{w_{b(x')}} - m_k\|| \\
&\le \max_{k\in Z_K} |w_{b(x)}\|\frac{b(x)}{w_{b(x)}} - m_k\| - w_{b(x')}\|\frac{b(x')}{w_{b(x')}} - m_k\||
\end{aligned}$$

$$
\begin{aligned}
&= \max_{k\in Z_K} |\frac{r}{w_{b(x)}} - \langle b(x), m_k\rangle + w_{b(x)}\|m_k\|^2 - \frac{r}{w_{b(x')}} \\
&\quad + \langle b(x'), m_k\rangle - w_{b(x')}\|m_k\|^2| \\
&\le \max_{k\in Z_K} (|\frac{r}{w_{b(x)}} - \frac{r}{w_{b(x')}}| + |\langle b(x) - b(x'), m_k\rangle| \\
&\quad + \|m_k\|^2 |w_{b(x)} - w_{b(x')}|) \\
&\le \max_{k\in Z_K} (|\frac{r}{w_{b(x)}} - \frac{r}{w_{b(x')}}| + \|b(x) - b(x')\|\|m_k\| \\
&\quad + \|m_k\|^2 |w_{b(x)} - w_{b(x')}|).
\end{aligned}
$$

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

$$
\begin{aligned}
&|f_{m_1,\ldots,m_K}(x) - f_{m_1,\ldots,m_K}(x')| \\
&\le \max_{k\in Z_K} (|\frac{r}{w_{b(x)}} - \frac{r}{w_{b(x')}}| + \|b(x) - b(x')\|\|m_k\| \\
&\quad + \|m_k\|^2 |w_{b(x)} - w_{b(x')}|) \\
&\le \max_{k\in Z_K} (\frac{r}{\min_{x\in X_n}(w_{b(x)})^2} + \|m_k\|^2) |w_{b(x)} - w_{b(x')}| \\
&\quad + \|b(x) - b(x')\|\|m_k\| \\
&\le \frac{r + \min_{x\in X_n} w_{b(x)}}{\min_{x\in X_n}(w_{b(x)})^2} \sum_{i=1}^{r} \gamma_w^i + (\frac{\sum_{i=1}^r \gamma_i^2}{\min_{x\in X_n} w_{b(x)}})^{\frac{1}{2}} \\
&\le \frac{2\sum_{i=1}^r \gamma_w^i}{r} + (\frac{\sum_{i=1}^r \gamma_i^2}{r})^{\frac{1}{2}}.
\end{aligned}
$$

This completes the proof. ■

D. PROOF OF THEOREM 4

The Glivenko-Cantelli theorem [16] 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:*

$$
\mathfrak{R}_n(F) = E_\sigma \sup_{f\in F} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i)
$$

and

$$
\mathcal{G}_n(F) = E_\gamma \sup_{f\in F} \frac{1}{n} \sum_{i=1}^{n} \gamma_i f(x_i),
$$

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

$$
\mathfrak{R}(F) = E_x \mathfrak{R}_n(F)
$$

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*

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

Then,

$$
E_x \Phi(X) \le 2\mathfrak{R}(F).
$$

The following theorem [18], 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*

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

Combining Theorem A 2 and Lemma 1, we have

THEOREM A 3. *Let π be any partitioning 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*

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

We use Lemmas 2 and 3 (see proofs in [11]) 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*

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

Then

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

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

LEMMA 3.

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

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

LEMMA 4.

$$
\begin{aligned}
\mathfrak{R}(F_{\Pi_K}) \le \frac{\sqrt{\pi/2} r K}{n} &((\sum_{i=1}^{n} \frac{1}{(w_{b(x_i)})^2})^{\frac{1}{2}} + \frac{2\sqrt{n}}{\min_{x\in X_n} w_{b(x)}} \\
&+ (\sum_{i=1}^{n} (w_{b(x_i)})^2)^{\frac{1}{2}} \frac{1}{\min_{x\in X_n}(w_{b(x)})^2}).
\end{aligned}
$$

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

$$\Omega_{\boldsymbol{M}} = \sum_{i=1}^{n} \gamma_i \min_{k \in Z_K} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2$$

and

$$\Xi_{\boldsymbol{M}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2,$$

where γ_i and γ_{ik} are independent Gaussian random variables indexed by i and k. And e_k are the natural bases indexed by k.

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

$$\begin{aligned}
&E(\Omega_{\boldsymbol{M}} - \Omega_{\boldsymbol{M}'})^2 \\
&= \sum_{i=1}^{n} (\min_{k \in Z_K} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2 \\
&\quad - \min_{k \in Z_K} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M}' e_k \|^2)^2 \\
&\leq \sum_{i=1}^{n} \max_{k \in Z_K} (w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2 \\
&\quad - w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M}' e_k \|^2)^2 \\
&\leq \sum_{i=1}^{n} \sum_{k=1}^{K} (w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2 \\
&\quad - w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M}' e_k \|^2)^2 \\
&= E(\Xi_{\boldsymbol{M}} - \Xi_{\boldsymbol{M}'})^2.
\end{aligned}$$

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

Using Slepian's Lemma and Lemma 3, we have

$$\begin{aligned}
&\mathfrak{R}(F_{\Pi_K}) \\
&= E_\sigma \sup_{f_{m_1,\ldots,m_K} \in F_{\Pi_K}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f_{m_1,\ldots,m_K}(x_i) \\
&= E_\sigma \sup_{\boldsymbol{M}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \min_{k \in Z_K} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2 \\
&\leq E_\gamma \frac{\sqrt{\pi/2}}{n} \sup_{\boldsymbol{M}} \sum_{i=1}^{n} \gamma_i \min_{k \in Z_K} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2 \\
&= E_\gamma \frac{\sqrt{\pi/2}}{n} \sup_{\boldsymbol{M}} \Omega_{\boldsymbol{M}} \\
&\leq E_\gamma \frac{\sqrt{\pi/2}}{n} \sup_{\boldsymbol{M}} \Xi_{\boldsymbol{M}} \\
&= E_\gamma \frac{\sqrt{\pi/2}}{n} \sup_{\boldsymbol{M}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} w_{b(x_i)} \| \frac{b(x_i)}{w_{b(x_i)}} - \boldsymbol{M} e_k \|^2 \\
&= \frac{\sqrt{\pi/2}}{n} E_\gamma (\sup_{\boldsymbol{M}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} w_{b(x_i)} (\frac{\|b(x_i)\|^2}{(w_{b(x_i)})^2} \\
&\quad - 2 \left\langle \frac{b(x_i)}{w_{b(x_i)}}, \boldsymbol{M} e_k \right\rangle + \|\boldsymbol{M} e_k\|^2)) \\
&\leq \frac{\sqrt{\pi/2}}{n} (E_\gamma \sup_{\boldsymbol{M}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \frac{r}{w_{b(x_i)}} \\
&\quad + 2 E_\gamma \sup_{\boldsymbol{M}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \langle b(x_i), \boldsymbol{M} e_k \rangle \\
&\quad + E_\gamma \sup_{\boldsymbol{M}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} w_{b(x_i)} \|\boldsymbol{M} e_k\|^2).
\end{aligned}$$

We give upper bounds to the three terms respectively.

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

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

$$\begin{aligned}
&2 E_\gamma \sup_{\boldsymbol{M}} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \langle b(x_i), \boldsymbol{M} e_k \rangle \\
&= 2 E_\gamma \sup_{\boldsymbol{M}} \sum_{k=1}^{K} \left\langle \sum_{i=1}^{n} \gamma_{ik} b(x_i), \boldsymbol{M} e_k \right\rangle \\
&\quad \text{(note: by Cauchy-Schwartz inequality)} \\
&\leq 2 E_\gamma \sup_{\boldsymbol{M}} \sum_{k=1}^{K} \| \sum_{i=1}^{n} \gamma_{ik} b(x_i) \| \|\boldsymbol{M} e_k\| \\
&\leq 2 E_\gamma \sum_{k=1}^{K} \| \sum_{i=1}^{n} \gamma_{ik} b(x_i) \| \frac{\sqrt{r}}{\min_{x \in X_n} w_{b(x)}} \\
&\quad \text{(note: by Jensen's inequality and the orthogaussian property of the Gaussian random variable)} \\
&\leq 2 \sum_{k=1}^{K} (\sum_{i=1}^{n} \|b(x_i)\|^2)^{\frac{1}{2}} \frac{\sqrt{r}}{\min_{x \in\ X_n} w_{b(x)}} \\
&= \frac{2\sqrt{n} r K}{\min_{x \in X_n} w_{b(x)}}.
\end{aligned}$$

The second inequality holds because

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

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

$$\begin{aligned}&E_\gamma \sup_{\boldsymbol{M}} \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} w_{b(x_i)} \|\boldsymbol{M}e_k\|^2 \\ &\leq E_\gamma \sum_{k=1}^K |\sum_{i=1}^n \gamma_{ik} w_{b(x_i)}| (\frac{\sqrt{r}}{\min_{x\in X_n}(w_{b(x)})^2})^2 \\ &\leq \sum_{k=1}^K (\sum_{i=1}^n w_{b(x_i)}^2)^{\frac{1}{2}} (\frac{\sqrt{r}}{\min_{x\in X_n}(w_{b(x)})^2})^2 \\ &= rK(\sum_{i=1}^n w_{b(x_i)}^2)^{\frac{1}{2}} \frac{1}{\min_{x\in X_n}(w_{b(x)})^2}.\end{aligned}$$

Thus, we have

$$\begin{aligned}&\mathfrak{R}(F_{\Pi_K}) \\ &\leq \frac{\sqrt{\pi/2}}{n}(rK(\sum_{i=1}^n \frac{1}{w_{b(x_i)}^2})^{\frac{1}{2}} + \frac{2\sqrt{n}rK}{\min_{x\in X_n} w_{b(x)}} \\ &\quad + rK(\sum_{i=1}^n w_{b(x_i)}^2)^{\frac{1}{2}} \frac{1}{\min_{x\in X_n} w_{b(x)}^2}) \\ &= \frac{\sqrt{\pi/2}rK}{n}(\sum_{i=1}^n \frac{1}{w_{b(x_i)}^2})^{\frac{1}{2}} + \frac{2\sqrt{n}}{\min_{x\in X_n} w_{b(x)}} \\ &\quad + (\sum_{i=1}^n w_{b(x_i)}^2)^{\frac{1}{2}} \frac{1}{\min_{x\in X_n}(w_{b(x)})^2}).\end{aligned}$$

This concludes the proof of Lemma 4. ■

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

E. PROOF OF THEOREM 5

Proof. The proof of Theorem 5 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:

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

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

F. PROOF OF THEOREM 6

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 $g = \langle g_1, \cdots, g_k\rangle$, $g_k = \langle g_{k,i}, \cdots, g_{k,r}\rangle$, and $g_k \neq m_k$,

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

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

$$\begin{aligned}\Delta &= \sum_{i=1}^r \sum_{k=1}^K \sum_{x\in C_k\cap\mathcal{X}_i} w_{b(x)} [\phi(b(x)_i) - \phi(g_{k,i}) \\ &\quad + (b(x)_i - g_{k,i})^\top \nabla(g_{k,i}) - \phi(b(x)_i) + \phi(m_{k,i}) \\ &\quad - (b(x)_i - m_{k,i})^\top \nabla(m_{k,i})] \\ &= \sum_{i=1}^r \sum_{k=1}^K \sum_{x\in C_k\cap\mathcal{X}_i} w_{b(x)} [\phi(m_{k,i}) - \phi(g_{k,i}) \\ &\quad + (b(x)_i - g_{k,i})^T \nabla(g_{k,i})] \\ &= \sum_{i=1}^r \sum_{k=1}^K w_{C_k}^{(i)} ||m_{k,i} - g_{k,i}||^2 > 0.\end{aligned} \tag{11}$$

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. ■