Hindawi
Computational Intelligence and Neuroscience
Volume 2017, Article ID 2658707, 14 pages
https://doi.org/10.1155/2017/2658707

Research Article

Fast Constrained Spectral Clustering and Cluster Ensemble with Random Projection

Wenfen Liu,[1,2,3] Mao Ye,[4] Jianghong Wei,[3] and Xuexian Hu[3]

[1] *Guangxi Key Laboratory of Cryptogpraphy and Information Security, School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China*
[2] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3] *State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450002, China*
[4] *National University of Defense Technology, Nanjing 210012, China*

Correspondence should be addressed to Mao Ye; yemaoxxgc@163.com

Received 7 March 2017; Accepted 1 August 2017; Published 25 September 2017

Academic Editor: Diego Andina

Constrained spectral clustering (CSC) method can greatly improve the clustering accuracy with the incorporation of constraint information into spectral clustering and thus has been paid academic attention widely. In this paper, we propose a fast CSC algorithm via encoding landmark-based graph construction into a new CSC model and applying random sampling to decrease the data size after spectral embedding. Compared with the original model, the new algorithm has the similar results with the increase of its model size asymptotically; compared with the most efficient CSC algorithm known, the new algorithm runs faster and has a wider range of suitable data sets. Meanwhile, a scalable semisupervised cluster ensemble algorithm is also proposed via the combination of our fast CSC algorithm and dimensionality reduction with random projection in the process of spectral ensemble clustering. We demonstrate by presenting theoretical analysis and empirical results that the new cluster ensemble algorithm has advantages in terms of efficiency and effectiveness. Furthermore, the approximate preservation of random projection in clustering accuracy proved in the stage of consensus clustering is also suitable for the weighted k-means clustering and thus gives the theoretical guarantee to this special kind of k-means clustering where each point has its corresponding weight.

1. Introduction

With the arrival of the big data era, data has become an important asset. How to analyse the large scale data efficiently is becoming a big challenge [1, 2]. As an underlying method for data analysis, clustering can partition a data set into several subsets according to the similarities of points [3], and it has become a basic tool for image analysis [4, 5], community detection [6, 7], disease diagnosis [8], and so on. Therefore, more and more attention has been paid to the design of efficient and effective clustering algorithms.

Constrained clustering can improve the accuracy of the clustering result via encoding constraint information into unsupervised clustering. As an important area of clustering, many constrained clustering algorithms [9–17] have been proposed. Since spectral clustering often has high clustering accuracy and the suitability for a wide range of geometries [18, 19], constrained spectral clustering (CSC) [11–17] can usually have better performance than other constrained clustering algorithms. However, the $O(n^2)$ space complexity and $O(n^3)$ time complexity of many CSC algorithms [11–15] restrict their applications over large scale data sets, where n is the number of data points. The most efficient CSC algorithm known is SCACS algorithm [16], which reduces the space and time complexities to be linear with n through incorporating the landmark-based graph construction [20, 21] with the constrained normalized cuts problem [15]. What is needed to be noticed is that the constrained normalized cuts problem [15] makes SCACS algorithm solve the generalized eigenvector problem twice. In 2016, Cucuringu et al. [17] proposed a new CSC algorithm with better accuracy and shorter running time empirically than constrained normalized cuts

problem. Taking a new encoding technique of constraint information, the new CSC model just needs the computation of eigenvectors once.

By means of integrating many basic partitions into a unified partition, ensemble clustering has many excellent properties such as the improvement of clustering quality, the robustness and stability of clustering results, the handling of noise, the reuse of knowledge [3], and the suitability to multisource and heterogeneous data [22]. Researchers have proposed many ensemble clustering algorithms [22–29]. Since there are different notations in different literatures, we call the integration of basic partitions as ensemble clustering or consensus clustering and call the union of the stages of basic clustering and ensemble clustering as cluster ensemble in the following. Among different ensemble clustering methods, the method based on coassociation matrix has become a landmark [22]. Specifically, the coassociation matrix is constructed to represent the similarities of pairs of points from the basic partitions and the final partition result is computed via the graph partition method on the matrix. Thus, this kind of method suffers from the high space and time complexity. Recently, Liu et al. [22] transformed spectral clustering on coassociation matrix to weighted k-means clustering over specific binary matrix equivalently, which decreased the space and time complexities vastly. However, when the number of basic partitions or clusters is large, the corresponding binary matrix will be high dimensional.

As the seminal work, Johnson and Lindenstrauss [30] pointed out that the random projection produced by random orthogonal matrix could preserve the pairwise distances of data sets approximately with reduced dimensions. Subsequently, a lot of researches constructed more matrices with the above properties: random Gaussian matrix [31], random sign matrix [32], random matrix based on randomized Hadamard transform [33], random matrix based on block random hashing [34], and so on. In addition, dimensionality reduction with random projection has also been widely applied to data mining methods such as classification [35], clustering [36–38], and anomaly detection [39]. In terms of object function, there are several works [36–38] to prove that random projection can maintain the accuracy of k-means clustering approximately. Since its objective function is different from that of k-means clustering, the theoretical analysis of the influence of random projection on weighted k-means clustering is still scarce.

Our Contribution. In this paper, our contributions can be divided into three parts: the first part is the proposition of a fast CSC algorithm which is suitable for a wide range of data sets; the second part is the analysis of the effect of random projection on the spectral ensemble clustering; the third part is the proposition of a scalable semisupervised cluster ensemble algorithm. More specifically, the contributions are as follows:

(i) We propose a fast CSC algorithm whose space and time complexities are linear with the size of a data set: we compress the size of the original model proposed by Cucuringu et al. [17] by the encoding of landmark-based graph construction and improve the efficiency further via random sampling in the process of k-means clustering. Besides, we prove that the new CSC algorithm will have the comparable clustering result of the original model asymptotically. Experimental results show that the new algorithm not only can utilize the constraint information effectively, but also costs less running time and fits a wider range of data sets compared to the state of the art SCACS method.

(ii) With respect to the difference of objective function caused by random projection, we give a detailed proof that random projection can keep the clustering quality of spectral ensemble clustering within a small factor. Based on this theoretical analysis, we design a spectral ensemble clustering algorithm with reduced dimensions caused by sparse random projection. Experiments over different data sets also verify the correctness of our theoretical results. Moreover, since the theoretical analysis is also suitable for the ordinary weighted k-means clustering, the influence of random projection on weighted k-means clustering is also obtained.

(iii) We propose a scalable semisupervised cluster ensemble algorithm through the combination of the fast CSC algorithm and spectral ensemble clustering algorithm with random projection. The efficiency and effectiveness of the new cluster ensemble algorithm are also demonstrated theoretically and empirically.

The remainder of our paper is organized as follows. In Section 2, we introduce the CSC model of Cucuringu et al. [17], landmark-based graph construction, and two related components in our cluster ensemble algorithm: spectral ensemble clustering and random projection. In Section 3, we present our fast CSC algorithm and give its asymptotic property. Then, the algorithm formulation and theoretical analysis of spectral ensemble clustering with random projection are displayed in Section 4. In Section 5, we show the experiment results of our algorithms. Finally, we draw the conclusions of the article and put forward the future directions in Section 6.

2. Preliminaries

In this section, we present the CSC algorithm proposed by Cucuringu et al. [17] and introduce landmark-based graph construction [20, 21] which will be applied to our fast CSC algorithm. In addition, we also introduce spectral ensemble clustering algorithm [22] and sparse random projection [34] which can be used to speed up the spectral ensemble clustering.

2.1. Constrained Spectral Clustering. Here, we first introduce the notion of undirected graph which is very important in constrained spectral clustering and then show the CSC model proposed by Cucuringu et al. [17].

Let $G = (V, E, W)$ be an undirected graph, where $V = \{v_1, v_2, \ldots, v_n\}$ is the vertex set, E is the edge set, and W is the weight set with respect to the edges. $w_{ij} = w_{ji}$ is specially the nonnegative weight of the edge between the vertices v_i and v_j, indicating the level of "affinity" between v_i and v_j. If

$w_{ij} = 0$, there is no edge between the vertices v_i and v_j. We denote $\mathbf{L}_G = \mathbf{D} - \mathbf{W}$ as the Laplacian matrix of G, where the diagonal entry of diagonal matrix $\mathbf{D}$ is $\mathbf{D}(i,i) = \sum_{j \neq i} w_{ij}$; $\mathbf{W}$ is an adjacency matrix with $\mathbf{W}(i,j) = \mathbf{W}(j,i) = w_{ij}$.

The constrained spectral clustering has three undirected graphs: one data graph G_D and two knowledge graphs G_{ML} and G_{CL}. In data graph $G_D = (V, E_D, W_D)$, each weight indicates the similarity level of vertices in the corresponding edge. The "must link" (ML) graph $G_{\mathrm{ML}} = (V, E_{\mathrm{ML}}, W_{\mathrm{ML}})$ gives the "must link" information of vertices: each edge in G_{ML} indicates that the corresponding vertices should be in the same group and the level of "must link" belief is described by the weight. The "cannot-link" (CL) graph $G_{\mathrm{CL}} = (V, E_{\mathrm{CL}}, W_{\mathrm{CL}})$ has analogous components to G_{ML}. The values of weights in the two knowledge graphs are both nonnegative and set according to the constraint information such as prior knowledge. For example, assuming that the range of value of weight is set from 0 to 1, if we have known that points v_1, v_2 are in the same group, their corresponding weight $w_{\mathrm{ML},12} = 1$. If we only have 40% confidence in the constraint information that the two points are in the same group, the weight $w_{\mathrm{ML},12} = 0.4$, and if we have no constraint information about these two points, $w_{\mathrm{ML},12} = w_{\mathrm{CL},12} = 0$.

Viewing pairwise similarities of vertices as the implicit ML constraints declaration, Cucuringu et al. [17] defined a generalized ML graph $\widetilde{G}_D[\alpha] = (V, E_D \cup E_{\mathrm{ML}}, W_D + \alpha * W_{\mathrm{ML}})$ where α is the level of trust for ML constrains. Let k be the number of clusters and $\mathbf{x}_{C_i}$ be the indicator vector of cluster C_i such that $\mathbf{x}_{C_i}(j) = 1$ if the jth data point belongs to cluster C_i and $\mathbf{x}_{C_i}(j) = 0$ otherwise. In order to violate as few ML constraints as possible and meet as many CL constraints as possible, the constrained k way cuts problem [17] can be described as

$$\begin{aligned} \underset{\{\{\mathbf{x}_{C_1}, \mathbf{x}_{C_2}, \ldots, \mathbf{x}_{C_k}\}\}}{\arg\min} \quad & \max_{\mathbf{x} \in \{\mathbf{x}_{C_1}, \mathbf{x}_{C_2}, \ldots, \mathbf{x}_{C_k}\}} \frac{\mathbf{x}^T \mathbf{L}_{\widetilde{G}_D} \mathbf{x}}{\mathbf{x}^T \mathbf{L}_{G_{\mathrm{CL}}} \mathbf{x}} \\ \text{s.t.} \quad & \sum_{i=1}^{k} \mathbf{x}_{C_i} = \{1\}^n, \quad \mathbf{x}_{C_i} \in \{0,1\}^n. \end{aligned} \tag{1}$$

To solve the problem in (1) approximately, Cucuringu et al. [17] relaxed the condition "$\mathbf{x}_i \in \{0,1\}^n$, $\sum_{i=1}^{k} \mathbf{x}_i = \{1\}^n$" to be the real vectors. Thus, the solution vectors of the relaxed problem are the first k nontrivial generalized eigenvectors of the problem

$$\mathbf{L}_{\widetilde{G}_D} \mathbf{x} = \lambda \mathbf{L}_{G_{\mathrm{CL}}} \mathbf{x}. \tag{2}$$

After getting the generalized eigenvectors, an additional embedding phase embeds the row vectors of eigenvectors matrix onto the k-dimensional sphere and gives the theoretical guarantees of clustering results. The detailed embedding procedures can be accessed in [17]. However, the construction cost and storage cost of data graphs for large scale data sets are both huge ($O(n^2)$). What is more, if the number of iterations in the process of k-means clustering on the embedded eigenvectors matrix is great, the process will also be time-consuming over large scale data sets.

2.2. Landmark-Based Graph Construction. Based on sparse coding theory [40], the landmark-based graph construction [20, 21] scales linearly with the number of data points and can suit large scale data sets very well.

Let data set be $\mathbf{A} \in \mathbb{R}^{n \times d}$ and the row vector $\mathbf{a}_i$ of $\mathbf{A}$ be data points; sparse coding problem is defined as follows:

$$\begin{aligned} \min_{\mathbf{U},\mathbf{Z}} \quad & \left\| \mathbf{A}^T - \mathbf{U}\mathbf{Z} \right\|^2 \\ \text{s.t.} \quad & \mathbf{Z} \text{ is sparse,} \end{aligned} \tag{3}$$

where each column vector of $\mathbf{U} \in \mathbb{R}^{d \times p}$ is the basis vector, column vectors of $\mathbf{Z} \in \mathbb{R}^{p \times n}$ are the representations of data points over $\mathbf{U}$ and p is the number of basis vectors. To avoid the high time complexity of solving sparse coding problem, landmark-based graph construction just samples points randomly from input data $\mathbf{A}$ as basis vectors. In the process of computing $\mathbf{Z}$, if $\mathbf{u}_j$ is among the r nearest basis vectors of data points $\mathbf{a}_i$, $\mathbf{Z}(j,i)$ can be computed as

$$\mathbf{Z}(j,i) = \frac{K_\sigma(\mathbf{a}_i, \mathbf{u}_j)}{\sum_{j' \in U(i,r)} K_\sigma(\mathbf{a}_i, \mathbf{u}_{j'})}, \tag{4}$$

where $U(i,r)$ is the indices set of the r nearest basis vectors of $\mathbf{a}_i$ and $K_\sigma(\cdot)$ is Gaussian kernel function with bandwidth σ; otherwise $\mathbf{Z}(j,i) = 0$.

After obtaining the sparse representation $\mathbf{Z} \in \mathbb{R}^{p \times n}$, graph affinity matrix is constructed as follows:

$$\mathbf{W} = \widehat{\mathbf{Z}}^T \widehat{\mathbf{Z}}, \tag{5}$$

where $\widehat{\mathbf{Z}} = \mathbf{D}^{-1/2}\mathbf{Z}$ and $\mathbf{D}$ is a diagonal matrix with diagonal entry $\mathbf{D}(i,i) = \sum_j \mathbf{Z}(i,j)$. Since Chen and Cai [20, 21] have pointed out that $\mathbf{W}$ was automatically normalized, the normalized graph Laplacian matrix for $\mathbf{A}$ is $\mathbf{I} - \widehat{\mathbf{Z}}^T \widehat{\mathbf{Z}}$. Considering $p \ll n$, the $O(npd)$ time of computing $\widehat{\mathbf{Z}}$ is much less than the $O(n^2 d)$ time of the nearest neighbors graph construction.

2.3. Spectral Ensemble Clustering. To gain the unified results from different basic partitions, spectral ensemble clustering applies spectral clustering to the coassociation matrix [24] derived from basic partitions. In 2015, Liu et al. [22] transformed spectral ensemble clustering into weighted $k$means clustering over specific binary matrix. This transformation decreased the time and space complexities effectively and our new ensemble clustering method is based on this nice transformation.

Given g basic clustering results $\Pi = \{\pi_1, \pi_2, \ldots, \pi_g\}$ of data set $\mathbf{A} \in \mathbb{R}^{n \times d}$; the coassociation matrix $\mathbf{C}$ is constructed in the following way:

$$\mathbf{C}(j,k) = \sum_{i=1}^{g} \eta\left(\pi_i(\mathbf{a}_j), \pi_i(\mathbf{a}_k)\right), \tag{6}$$

where $\pi_i(\mathbf{a}_j)$ is the label of $\mathbf{a}_j$ in the ith clustering result π_i, and

$$\eta(a,b) = \begin{cases} 1, & \text{if } a = b \\ 0, & \text{if } a \neq b. \end{cases} \tag{7}$$

Viewing this coassociation matrix as adjacency matrix, spectral ensemble clustering uses spectral clustering to get final clustering result. In the process of the transformation from spectral clustering to weighted k-means clustering, binary matrix $\mathbf{B} = \{\mathbf{b}(\mathbf{a})\}$ [22] is built as follows:

$$\mathbf{b}(\mathbf{a}) = \left[\mathbf{b}(\mathbf{a})_1, \ldots, \mathbf{b}(\mathbf{a})_g\right], \tag{8}$$

where $\mathbf{b}(\mathbf{a})_i = [b(\mathbf{a})_{i1}, \ldots, b(\mathbf{a})_{ik_i}]$, $b(\mathbf{a})_{ij} = 1$ if $\pi_i(\mathbf{a}) = j$, and $b(\mathbf{a})_{ij} = 0$ otherwise; "[]" indicates a row vector. The following lemma [22] presents the connection between spectral ensemble clustering and weighted k-means clustering.

Lemma 1 (see [22]). *Given a basic partitions set Π, let the corresponding coassociation matrix be $\mathbf{C}$, the diagonal matrix whose diagonal elements are sums of rows of $\mathbf{C}$ be $\mathbf{D}1$, and the diagonal element set of $\mathbf{D}1$ be $\{w_{\mathbf{b}(\mathbf{a})}\}$. Then normalized cuts spectral clustering on coassociation matrix $\mathbf{C}$ has equivalent objective function to weighted k-means clustering on data sets $\{\mathbf{b}(\mathbf{a})/w_{\mathbf{b}(\mathbf{a})}\}$ with weight set $\{w_{\mathbf{b}(\mathbf{a})}\}$.*

Through Lemma 1, the space and time complexities of spectral ensemble clustering can be decreased dramatically. However, when the number of basic partitions and cluster number are large, the binary matrix $\mathbf{B}$ will be a high dimensional data set, resulting in long running time for weighted k-means clustering.

2.4. Random Projection. Recently, random projection has become a common technique of dimensionality reduction [36–39, 41]. Random projection often has low computing complexity and can preserve the structure of original data approximately. In this paper, we use the sparse random projection proposed by Kane and Nelson [34]. When most of the elements of data are zero, the sparse random projection can utilize the sparsity of data effectively and speed up the process of dimensionality reduction.

Lemma 2 (see [34]). *For any $0 < \delta, \varepsilon < 1/2$, $d > 0$, there exists an $d \times (av)$ sparse random matrix $\mathbf{R}$, where $a = \Theta(\varepsilon^{-1}\log(1/\delta))$ and $v = \Theta(\varepsilon^{-1})$, such that for any fixed $\mathbf{x} \in \mathbb{R}^d$*

$$\Pr\left\{(1-\varepsilon)\|\mathbf{x}\|_2^2 \leq \left\|\mathbf{R}^T\mathbf{x}\right\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}\|_2^2\right\} > 1-\delta. \tag{9}$$

And the random matrix $\mathbf{R}$ can be constructed as follows:

$$\mathbf{R}^T = \begin{bmatrix} \sqrt{\frac{1}{a}} \cdot \Phi_1 \cdot \mathbf{D}_1 \\ \vdots \\ \sqrt{\frac{1}{a}} \cdot \Phi_a \cdot \mathbf{D}_a \end{bmatrix}, \tag{10}$$

where matrix Φ_l ($l \in [1,a]$) is a $v \times d$ sparse matrix with nonzero elements $\Phi(h(i),i) = 1$, $h : \{1,\ldots,d\} \to \{1,\ldots,v\}$ is a random hashing such that $\Pr\{h(i) = j\} = 1/v$ for $i \in \{1,\ldots,d\}$, $j \in \{1,\ldots,v\}$, and matrix $\mathbf{D}_l$ is a $d \times d$ diagonal matrix with $\Pr\{\mathbf{D}_l(i,i) = \pm 1\} = 0.5$.

The number of nonzero (nnz) elements of sparse random matrix $\mathbf{R}$ is ad, and the time complexity of $\mathbf{AR}$ is $\mathrm{nnz}(\mathbf{A})a$. Lemma 2 implies that the sparse random projection can preserve the length of data points approximately. Thus, for n data points, since there are $n(n-1)/2$ pairwise distances, we can conclude that the pairwise distances squares can be preserved within a factor of $1 \pm \varepsilon$ with $a = \Theta(2\varepsilon^{-1}\log(n/\delta))$.

3. Fast Constrained Spectral Clustering Framework

In this section, we introduce our fast CSC framework for large scale data sets. Inspired by [20, 21], we also try to compute the sparse representation $\widehat{\mathbf{Z}}$ and obtain the approximate adjacency matrix $\mathbf{W} = \widehat{\mathbf{Z}}^T\widehat{\mathbf{Z}}$, where $\widehat{\mathbf{Z}} \in \mathbb{R}^{p\times n}$, and $p \ll n$. Then, our fast framework decreases the size of graph Laplacian through the above approximate graph reconstruction. At last, we analyse the asymptotic property of our new CSC algorithm.

3.1. Framework Formulation. To get the generalized eigenvector $\mathbf{x}$ approximately, we can let $\mathbf{x} = \widehat{\mathbf{Z}}^T\mathbf{y}$, where $\widehat{\mathbf{Z}} \in \mathbb{R}^{p\times n}$ is the sparse representation in (5) and $\mathbf{y} \in \mathbb{R}^p$. Thus, bringing the $\mathbf{x}$ back to (1) can decrease the size of problem apparently if $p \ll n$.

Specifically, we use $\mathbf{Q}$ to denote constraint matrix, where $\mathbf{Q}(i,j) = 1$ if edge $(\mathbf{v}_i,\mathbf{v}_j) \in E_{\mathrm{ML}}$, $\mathbf{Q}(i,j) = -1$ if edge $(\mathbf{v}_i,\mathbf{v}_j) \in E_{\mathrm{CL}}$, and $\mathbf{Q}(i,j) = 0$ otherwise. Let adjacency matrix be computed approximately by $\mathbf{W} = \widehat{\mathbf{Z}}^T\widehat{\mathbf{Z}}$. Next, bring $\mathbf{x} = \widehat{\mathbf{Z}}^T\mathbf{y}$ into (1) and relax their solution over real vectors. Thus, we reformulate the original problem as the following problem.

Problem 3. One has

$$\begin{aligned} \underset{\{\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k\}\}}{\arg\min} \quad & \max_{\mathbf{y}\in\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_k\}} \frac{\mathbf{y}^T\widehat{\mathbf{Z}}\mathbf{L}_{\widetilde{G}_D}\widehat{\mathbf{Z}}^T\mathbf{y}}{\mathbf{y}^T\widehat{\mathbf{Z}}\mathbf{L}_{G_{\mathrm{CL}}}\widehat{\mathbf{Z}}^T\mathbf{y}} \\ \text{s.t.} \quad & \mathbf{y}_i \in \mathbb{R}^p \quad \text{for any } i \in [1,k]. \end{aligned} \tag{11}$$

To obtain shorthand notations, we denote $\widehat{\mathbf{Z}}\mathbf{L}_{\widetilde{G}_D}\widehat{\mathbf{Z}}^T$ by $\mathbf{L}_{\mathrm{CGD}}$ and denote $\widehat{\mathbf{Z}}\mathbf{L}_{G_{\mathrm{CL}}}\widehat{\mathbf{Z}}^T$ by $\mathbf{L}_{\mathrm{CCL}}$. Thus, the first k nontrivial generalized eigenvectors of the problem

$$\mathbf{L}_{\mathrm{CGD}}\mathbf{y} = \lambda\mathbf{L}_{\mathrm{CCL}}\mathbf{y} \tag{12}$$

are the solution vectors of (11).

In order to speed up the k-means clustering on the embedded eigenvector matrix, we sample row vectors of eigenvectors matrix randomly and get k centers through k-means clustering over the selected row vectors. According to

Input: data set $\mathbf{A} \in \mathbb{R}^{n\times d}$, the number of landmark points p, constraint matrix $\mathbf{Q}$, cluster number k, confidence parameter α, sample rate s;
Output: the grouping result.
(1) Compute the sparse representation $\widehat{\mathbf{Z}} \in \mathbb{R}^{\mathbf{p}\times\mathbf{n}}$ in Equation (5);
(2) Compute Laplacian $\mathbf{L}_{\mathrm{CGD}} = \widehat{\mathbf{Z}}\mathbf{L}_{\widetilde{G}_D}\widehat{\mathbf{Z}}^T$ and $\mathbf{L}_{\mathrm{CCL}} = \widehat{\mathbf{Z}}\mathbf{L}_{G_{\mathrm{CL}}}\widehat{\mathbf{Z}}^T$, where $\mathbf{L}_{\widetilde{G}_D}$ is the Laplacian matrix of $\widetilde{G}_D$, $\mathbf{L}_{G_{\mathrm{CL}}}$ is the Laplacian matrix of G_{CL};
(3) Solve the first k non-trivial generalized eigenvectors $\mathbf{Y}$ of Equation (12);
(4) Compute $\mathbf{X} = \widehat{\mathbf{Z}}^T\mathbf{Y}$;
(5) Embed $\mathbf{X}$ into a k-dimensional sphere $\widehat{\mathbf{X}}$ using the embedding process in [17];
(6) Sample $n \times s$ row vectors of $\widehat{\mathbf{X}}$ randomly and run k-means clustering on the sampled row vectors;
(7) Get the clustering result utilizing distances between centers of k-means clustering and row vectors of $\widehat{\mathbf{X}}$.

Algorithm 1: Fast constrained spectral clustering.

the distances between centers and row vectors, we can partition all the row vectors into different clusters. Cucuringu et al. [17] have pointed out that the specific embedding process after getting the generalized eigenvectors can concentrate the row vectors of eigenvector matrix onto the k-dimensional sphere and a simple partition algorithm such as k-means clustering can be applied to get the final clustering result. Since random sampling is a popular scalability method for k-means clustering [42], we will take it to improve the efficiency of the clustering on the row vectors of eigenvector matrix. The experimental results in Section 5 also show that random sampling has little influence on the clustering results and makes the algorithm more efficient than the original one.

Our fast CSC framework is shown in Algorithm 1. In our new algorithm, parameter α (in $\mathbf{L}_{\widetilde{G}_D}$ of Step (2)) stands for the trust level on constraint information. Since the α of the original problem (see (2)) has been taken to a constant in the previous work [17], we also set α as a constant.

The complexity analysis of Algorithm 1 is presented as follows. The time of computing $\widehat{\mathbf{Z}}$ is $O(\mathrm{npd})$. In Step (2), the $\mathbf{L}_{\mathrm{CGD}}$ is computed as follows:

$$\begin{aligned}\mathbf{L}_{\mathrm{CGD}} &= \widehat{\mathbf{Z}}\mathbf{L}_{\widetilde{G}_D}\widehat{\mathbf{Z}}^T = \widehat{\mathbf{Z}}\left(\mathbf{I} - \widehat{\mathbf{Z}}^T\widehat{\mathbf{Z}} + \alpha\mathbf{L}_{G_{\mathrm{ML}}}\right)\widehat{\mathbf{Z}}^T \\ &= \widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T - \left(\widehat{\mathbf{Z}}\widehat{\mathbf{Z}}^T\right)^2 + \alpha\widehat{\mathbf{Z}}\mathbf{L}_{G_{\mathrm{ML}}}\widehat{\mathbf{Z}}^T.\end{aligned} \tag{13}$$

Let the number of data points with constraint information be c; then the time cost for computing $\alpha\widehat{\mathbf{Z}}\mathbf{L}_{G_{\mathrm{ML}}}\widehat{\mathbf{Z}}^T$ is $O(p^2c + pc^2)$. Hence, the time cost of Steps (1) and (2) is $O(p^2n + p^3) + O(p^2c + pc^2) = O(p^2n + p^3 + p^2c + pc^2)$. Besides, the time complexity of Step (3) is $O(p^3)$, that of Step (4) is $O(kpn)$, and that of Step (5) is $O(kn)$. Thus, the time cost of the first 5 steps is $O(p^2n)$ considering $p, c \ll n$ and $k \ll p, c$. Assuming the iteration numbers of k-means clustering are l, the time cost of Steps (6) and (7) is $O((ns)k^2l + nk^2)$, which is much less than the time cost $O(nlk^2)$ of k-means clustering on $\widehat{\mathbf{X}}$ with $(ns) \ll n$. Hence, the time complexity of our algorithm is

$$O\left(np^2 + nk^2 + npd\right). \tag{14}$$

Since three matrices $\widehat{\mathbf{Z}}$, $\mathbf{L}_{\mathrm{CGD}}$, and $\mathbf{L}_{\mathrm{CCL}}$ are stored, the memory complexity is

$$O\left(np + p^2\right). \tag{15}$$

3.2. Asymptotic Property of the Framework. In this subsection, we show that the partition result of our fast CSC algorithm could be comparable to that of the original model [17] as p converges to n.

Theorem 4. *Assuming the adjacency matrix* $\mathbf{W}$ *in the original model is full rank, the result of Step* (4) *in Algorithm 1 will converge to the generalized eigenvectors of* (2) *as* p *converges to* n.

Proof. From the construction of sparse representation $\widehat{\mathbf{Z}}$, we can get that

$$\lim_{p\to n}\widehat{\mathbf{Z}} = \widehat{\mathbf{W}}, \tag{16}$$

where $\widehat{\mathbf{W}}$ is the normalized adjacency matrix. Equation (12) can be rewritten as

$$\widehat{\mathbf{Z}}\left[\mathbf{I} - \widehat{\mathbf{W}} + \alpha\mathbf{L}_{G_{\mathrm{ML}}}\right]\widehat{\mathbf{Z}}^T\mathbf{y} = \lambda\widehat{\mathbf{Z}}\mathbf{L}_{G_{\mathrm{CL}}}\widehat{\mathbf{Z}}^T\mathbf{y}. \tag{17}$$

Equally, we have that

$$\widehat{\mathbf{Z}}\left[\mathbf{I} - \widehat{\mathbf{W}} + \alpha\mathbf{L}_{G_{\mathrm{ML}}} - \lambda\mathbf{L}_{G_{\mathrm{CL}}}\right]\widehat{\mathbf{Z}}^T\mathbf{y} = \mathbf{0}. \tag{18}$$

Since the rank of $\widehat{\mathbf{Z}}$ will be equal to n, $\widehat{\mathbf{Z}}$ can be removed. Thus the equation will be

$$\left[\mathbf{I} - \widehat{\mathbf{W}} + \alpha\mathbf{L}_{G_{\mathrm{ML}}}\right]\widehat{\mathbf{Z}}^T\mathbf{y} = \lambda\mathbf{L}_{G_{\mathrm{CL}}}\widehat{\mathbf{Z}}^T\mathbf{y}. \tag{19}$$

This equation shows that $\widehat{\mathbf{Z}}^T\mathbf{y}$ and λ in Step (4) of Algorithm 1 are indeed the eigenvector and eigenvalue of (2), respectively. Moreover, the number of eigenvectors of (19) will converge to n as p converges n. Hence Algorithm 2 could also get all the eigenvectors of (2) asymptotically. □

Since the eigenvectors of our framework will converge to that of original CSC model [17] and the random sampling

Input: binary matrix $\mathbf{B} \in \mathbb{R}^{n \times d'}$, weights set $\{w_{\mathbf{b}(\mathbf{x})}\}$, cluster number k.
Output: the final partition result.
(1) Generate a $d' \times (va)$ sparse random matrix $\mathbf{R}$ meeting the requirements of Lemma 2, where $a = \Theta(2\varepsilon^{-1}\log(n/\delta))$, $v = \Theta(\varepsilon^{-1})$, $0 < \delta, \varepsilon < 1/2$, $va < d'$;
(2) Compute $\widetilde{\mathbf{B}} = \mathbf{W}_{\mathbf{B}}^{-1}\mathbf{B}$, where $\mathbf{W}_{\mathbf{B}}$ is a diagonal matrix with diagonal entries $\{w_{\mathbf{b}(\mathbf{x})}\}$;
(3) Compute $\widehat{\mathbf{B}} = \widetilde{\mathbf{B}}\mathbf{R}$;
(4) Run weighted k-means clustering on $\widehat{\mathbf{B}}$ with weight set $\{w_{\mathbf{b}(\mathbf{x})}\}$ to obtain the final clustering result.

Algorithm 2: Spectral ensemble clustering with random projection.

has little influence on the clustering result of embedded eigenvectors matrix, our new CSC algorithm will generate the partition result which is comparable to that of original framework. In addition, the reason why we give the assumption of Theorem 4 is that each row vector of adjacency matrix is the similarity representation of certain point over the whole data set, and those representations are often linearly independent. In the experiments, we have demonstrated this theory empirically on the 30 nearest neighbors adjacency matrices of three data sets.

4. Spectral Ensemble Clustering with Random Projection

In this section, we propose an improved spectral ensemble clustering algorithm with random projection. The new ensemble clustering not only improves the efficiency of spectral ensemble clustering algorithm designed by Liu et al. [22], but also can theoretically preserve the approximate clustering result.

4.1. Algorithm Formulation. In this subsection, we give the detailed procedure of our new spectral ensemble clustering algorithm. We denote the original spectral ensemble clustering [22] by SEC and our improved spectral ensemble clustering with random projection by SECRP.

From the description of Section 2.3, we can know that the SEC algorithm transforms the spectral clustering on the coassociation matrix into weighted k-means clustering on the specific binary matrix $\mathbf{B}$. The dimension of binary matrix $\mathbf{B}$ is $\sum_{i=1}^{g} k_i$, where k_i is the cluster number of basic partition π_i. When the number of clusters and/or basic partitions is big, $\mathbf{B}$ is probably a high dimensional matrix on which the weighted k-means clustering runs slowly.

To avoid the high dimensions of $\mathbf{B}$, we design an improved SEC algorithm with random projection for dimensionality reduction. The new algorithm SECRP is showed in Algorithm 2.

The complexity analysis of the new algorithm is as follows. Obviously, the running time of Steps (1) and (2) is very short (compared with that of Step (3)). The time of Step (3) is $O(nnz(\mathbf{B})a) = O(nga)$, where g is the number of basic partitions; $nnz()$ denotes the number of nonzero entries. Another common method of dimensionality reduction is singular value decomposition (SVD). The time of running SVD on binary matrix $\mathbf{B}$ is $O((d')^3 + n(d')^2)$, and that of the product between eigenvectors and $\mathbf{B}$ is $O(nd'va)$. Since $g \approx d'/k$, random projection with sparse random matrix is a cost-effective method of dimensionality reduction. With respect to the weighted k-means clustering, dimensionality reduction of random projection can decrease the running time of each iteration from $O(nkd')$ to $O(nkva)$.

As a basic module, Algorithm 2 can be combined with different basic partition methods to produce different cluster ensemble algorithms. Thus, taking Algorithm 1 as the basic partition algorithm for Algorithm 2 could generate an efficient constrained cluster ensemble method with high accuracy (both basic partitions and final clustering are spectral clustering). Moreover, the last two steps of Algorithm 2 are just weighted k-means clustering with sparse random projection, which is also suitable for any other applications of weighted k-means clustering.

4.2. Theoretical Analysis of New Ensemble Algorithm. In this subsection, we demonstrate that our new algorithm SECRP can maintain the clustering result of SEC approximately.

For the theoretical analysis, we give the formal definition of weighted k-means clustering problem with matrix notation:

Definition 5 (weighted K-means clustering problem). Given an n points set $\mathbf{B}$ (each row is a data point), diagonal matrix $\mathbf{W}_{\mathbf{B}}$ whose diagonal entries set $\{w_{\mathbf{b}}\}$ is weights set and clusters number k find an $n \times k$ indicator matrix $\mathbf{X}_{\text{opt}}$ such that

$$\mathbf{X}_{\text{opt}} = \arg\min_{\mathbf{X}} \left\| \mathbf{W}_{\mathbf{B}}^{1/2} \left(\mathbf{B} - \mathbf{X}\mathbf{X}^T \mathbf{W}_{\mathbf{B}} \mathbf{B} \right) \right\|_F^2, \tag{20}$$

where $\|\cdot\|_F^2$ denotes the square of Frobenius norm; $\mathbf{X}$ is selected from the set of all indicator matrices. An indicator matrix has one nonzero element on each row. Specifically, if the ith point belongs to the jth cluster, $\mathbf{X}(i, j) = 1/\sqrt{w_{Cj}}$, where w_{Cj} denotes the sum of weights points in cluster C_j.

Since computing $\mathbf{X}_{\text{opt}}$ is an NP-hard problem, we focus on the approximate algorithm for weighted k-means clustering. The corresponding definition is as follows.

Definition 6 (weighted K-means approximation algorithm). An algorithm is called the "γ-approximation" for weighted k-means clustering problem, if the algorithm takes $\mathbf{B}$, k, and $\mathbf{W}_{\mathbf{B}}$ as input and outputs an indicator matrix $\mathbf{X}_\gamma$ such that

$$\Pr\left\{\left\|\mathbf{W}_{\mathbf{B}}^{1/2}\left(\mathbf{B}-\mathbf{X}_{\gamma}\mathbf{X}_{\gamma}^{T}\mathbf{W}_{\mathbf{B}}\mathbf{B}\right)\right\|_{F}^{2}\right.$$
$$\left.\leq \gamma \min_{\mathbf{X}}\left\|\mathbf{W}_{\mathbf{B}}^{1/2}\left(\mathbf{B}-\mathbf{X}\mathbf{X}^{T}\mathbf{W}_{\mathbf{B}}\mathbf{B}\right)\right\|_{F}^{2}\right\}\geq 1-\delta_{\gamma}, \tag{21}$$

where γ is the approximation factor and δ_{γ} is the failure probability of the "γ-approximation" weighted k-means clustering algorithm.

Though there is the γ-approximation k-means clustering algorithm such as [43], it is unclear whether the γ-approximation weighted k-means clustering algorithm exists or not. To facilitate the proof of our theory, we assume that the approximation algorithm exists and utilize the definition of approximation algorithm in the process of proof. And we will take the weighted version of the classical k-means clustering algorithm [44] as the weighted k-means clustering to verify our theoretical results in the following experiments.

Theorem 7. *Let $n\times d'$ matrix $\mathbf{B}$, weight set $\{\omega_{\mathbf{b}(\mathbf{a})}\}$, and cluster number k be the inputs of Algorithm 2. Let $\varepsilon \in (0, 1/3)$. Assuming that a γ-approximation weighted k-means clustering algorithm exists, then the output $\mathbf{X}_{\widehat{\gamma}}$ of Algorithm 2 satisfies with probability of at least $0.97-\delta_{\gamma}$:*

$$\left\|\mathbf{W}_{\mathbf{B}}^{1/2}\left(\widetilde{\mathbf{B}}-\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}\widetilde{\mathbf{B}}\right)\right\|_{F}^{2}$$
$$\leq\left(1+(1+\varepsilon)\gamma\right)\left\|\mathbf{W}_{\mathbf{B}}^{1/2}\left(\widetilde{\mathbf{B}}-\mathbf{X}_{opt}\mathbf{X}_{opt}^{T}\mathbf{W}_{\mathbf{B}}\widetilde{\mathbf{B}}\right)\right\|_{F}^{2}. \tag{22}$$

In the above, $\widetilde{\mathbf{B}}=\mathbf{W}_{\mathbf{B}}^{-1}\mathbf{B}$ is the computing result of Step (2) in Algorithm 2; $\mathbf{X}_{\mathrm{opt}}$ is the optimal solution of weighted k-means clustering on $\widetilde{\mathbf{B}}$.

This theorem reveals that random projection not only can be used to improve the efficiency of spectral ensemble clustering with lower dimensions, but also maintains its final result approximately.

In the following, we present a useful lemma which is needed in the proof of Theorem 7. The results of the lemma are based on the results of [36] and Lemma 2.

Lemma 8. *Let $\widetilde{\mathbf{B}}$, $\mathbf{R}$, $\mathbf{W}_{\mathbf{B}}$, k, and ε be the same as those in Theorem 7; denote $\mathbf{W}_{\mathbf{B}}^{1/2}\widetilde{\mathbf{B}}$ by $\mathbf{H}$, the product of top k singular vectors (left and right) and singular values of $\mathbf{H}$ by $\mathbf{H}_k$.*

(1) (Lemma 5 of [36]) Let the SVD of $\mathbf{H}_k$ be $\mathbf{H}_k=\mathbf{U}_k\Sigma_k\mathbf{V}_k^T$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ are the left and right singular vector matrices; Σ_k is a diagonal matrix whose diagonal elements are the k singular values. With probability of at least 0.97,

$$\mathbf{H}_k=\mathbf{H}\mathbf{R}\left(\mathbf{V}_k^T\mathbf{R}\right)^{\dagger}\mathbf{V}_k^T+\mathbf{E}, \tag{23}$$

where $(\,)^{\dagger}$ is the pseudoinverse of matrix; $\mathbf{E}$ is an $n\times d'$ matrix with $\|\mathbf{E}\|_F\leq 4\varepsilon\|\mathbf{H}-\mathbf{H}_k\|_F$.

(2) (Lemma 4 of [36]) For any $n\times d'$ matrix $\mathbf{G}$, with probability of at least 0.99,

$$\|\mathbf{G}\mathbf{R}\|_F\leq\sqrt{1+\varepsilon}\,\|\mathbf{G}\|_F. \tag{24}$$

(3) (Combination of Lemmas 2 and 3 of [36]) With probability of at least 0.99,

$$\left\|\left(\mathbf{V}_k^T\mathbf{R}\right)^{\dagger}\right\|_2\leq\frac{1}{1-\varepsilon}. \tag{25}$$

These conclusions are all about the influences of random matrix $\mathbf{R}$ on the norms of different matrices, which are useful for bounding the norms of the matrices in Theorem 7. In the following proof of Theorem 7, we start by decomposing the term $\|\mathbf{W}_{\mathbf{B}}^{1/2}(\widetilde{\mathbf{B}}-\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}\widetilde{\mathbf{B}})\|_F^2$ in (22). Then, based on the influences of random matrix in Lemma 8, we manipulate the norms of the different terms in the decomposition result.

Proof. Using the notation of Lemma 8, (22) can be decomposed into

$$\begin{aligned}
&\left\|\mathbf{W}_{\mathbf{B}}^{1/2}\left(\widetilde{\mathbf{B}}-\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}\widetilde{\mathbf{B}}\right)\right\|_{F}^{2}\\
&=\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{W}_{\mathbf{B}}^{1/2}\widetilde{\mathbf{B}}\right\|_{F}^{2}\\
&=\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|_{F}^{2}\\
&=\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}_k\right\|_{F}^{2}\\
&\quad+\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}_{\rho-k}\right\|_{F}^{2},
\end{aligned} \tag{26}$$

where $\mathbf{H}_{\rho-k}=\mathbf{H}-\mathbf{H}_k$. The last equation is based on the orthogonality of $\mathbf{H}_k$ and $\mathbf{H}_{\rho-k}$.

We first give the bound of the second term of (26). According to our definition of indicator matrix, $\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}=\mathbf{I}_k$. Thus, $\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}$ is a projector matrix; namely, its l_2 norm is 1. As a result, we get

$$\begin{aligned}
\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}_{\rho-k}\right\|_{F}^{2}&\leq\left\|\mathbf{H}_{\rho-k}\right\|_{F}^{2}\\
&\leq\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|_{F}^{2},
\end{aligned} \tag{27}$$

where the second inequality is caused by the fact that $\mathrm{rank}(\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\mathrm{opt}}\mathbf{X}_{\mathrm{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2})\leq k$ and the optimality of SVD.

We next bound the first term of (26). From the first statement of Lemma 8, we get

$$\begin{aligned}
&\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}_k\right\|_{F}\\
&\leq\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\mathbf{R}\left(\mathbf{V}_k^T\mathbf{R}\right)^{\dagger}\mathbf{V}_k^T\right\|_{F}\\
&\quad+\|\mathbf{E}\|_{F}\\
&\leq\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\mathbf{R}\right\|_{F}\left\|\left(\mathbf{V}_k^T\mathbf{R}\right)^{\dagger}\right\|_{F}\\
&\quad+\|\mathbf{E}\|_{F}.
\end{aligned} \tag{28}$$

From Definition 6 and the meaning of $\mathbf{X}_{\text{opt}}$ of Theorem 7, we get

$$\begin{aligned}&\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{HR}\right\|_{F}\\&\quad\leq\sqrt{\gamma}\min_{\mathbf{X}}\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}\mathbf{X}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{HR}\right\|_{F}\\&\quad\leq\sqrt{\gamma}\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\text{opt}}\mathbf{X}_{\text{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{HR}\right\|_{F}.\end{aligned}\tag{29}$$

Using the statement 2 of Lemma 8, (29) can be transformed to

$$\begin{aligned}&\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{HR}\right\|_{F}\\&\quad\leq\sqrt{\gamma\left(1+\varepsilon\right)}\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\text{opt}}\mathbf{X}_{\text{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|.\end{aligned}\tag{30}$$

Combining the statement 3 of Lemma 8 and (30), we get

$$\begin{aligned}&\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{HR}\right\|_{F}\left\|\left(\mathbf{V}_{k}^{T}\mathbf{R}\right)^{\dagger}\right\|_{F}+\|\mathbf{E}\|_{F}\\&\quad\leq\sqrt{\gamma\left(1+\varepsilon\right)}\cdot\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\text{opt}}\mathbf{X}_{\text{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|_{F}\\&\quad\cdot\left\|\left(\mathbf{V}_{k}^{T}\mathbf{R}\right)^{\dagger}\right\|_{F}+\|\mathbf{E}\|_{F}\\&\quad\leq\frac{\sqrt{\gamma\left(1+\varepsilon\right)}}{1-\varepsilon}\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\text{opt}}\mathbf{X}_{\text{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|_{F}\\&\quad+\|\mathbf{E}\|_{F}\leq\sqrt{\gamma}\left(\frac{\sqrt{1+\varepsilon}}{1-\varepsilon}+4\varepsilon\right)\\&\quad\cdot\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\text{opt}}\mathbf{X}_{\text{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|_{F}.\end{aligned}\tag{31}$$

From (28) and (31), and rescaling ε, we can get

$$\begin{aligned}&\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\widehat{\gamma}}\mathbf{X}_{\widehat{\gamma}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}_{k}\right\|_{F}\\&\quad\leq\sqrt{\gamma\left(1+\varepsilon\right)}\left\|\left(\mathbf{I}-\mathbf{W}_{\mathbf{B}}^{1/2}\mathbf{X}_{\text{opt}}\mathbf{X}_{\text{opt}}^{T}\mathbf{W}_{\mathbf{B}}^{1/2}\right)\mathbf{H}\right\|_{F}.\end{aligned}\tag{32}$$

Finally, combining (27) and (32) concludes the proof. □

It is easy to check that the above theoretical analysis can be also applied to ordinary weighted k-means clustering, indicating that the method of dimensionality reduction with random projection can preserve the clustering quality of weighted k means clustering approximately. Furthermore, the integration of Theorems 4 and 7 means that the new semisupervised cluster ensemble method (combination of Algorithms 1 and 2) can have an encouraging clustering result.

5. Experiments

In this section, we present the experimental results of our new algorithms in Sections 3 and 4. We implemented all the related algorithms in Matlab and conducted our experiments on a Windows machine with the Intel Core 3.6 GHz processor and 16 GB of RAM.

Table 1: Data sets information.

Data set	#instances	#attributes	#classes
Letter recognition	20,000	16	26
MNIST	70,000	784	10
CoverType	581,012	54	7

5.1. Data Sets and Experimental Settings. In order to facilitate the comparison, we performed experiments on three data sets which can be achieved from public web sites (http://archive.ics.uci.edu/ml/), (http://www.cad.zju.edu.cn/home/dengcai/). Table 1 summarizes their basic information.

The constraint information is generated from the real labels of data sets. In our experiments, we sample the labeled points randomly from data sets. The constraint matrix $\mathbf{Q}$ is constructed as

$$\mathbf{Q}\left(i,j\right)=\begin{cases}1 & \mathbf{x}_{i},\mathbf{x}_{j}\text{ have the same label}\\-1 & \mathbf{x}_{i},\mathbf{x}_{j}\text{ have different labels}\\0 & \text{no constraint.}\end{cases}\tag{33}$$

The validation measures of the partition result used in our experiments are cluster accuracy (CA) [45] and normalized mutual information (NMI) [25]. The CA is computed as

$$\text{CA}=\sum_{i=1}^{k}\frac{\max\left(\text{cluster}_{i}\mid\text{label}\right)}{n},\tag{34}$$

where k is the cluster number of clustering result, n is the number of data points, $\max(\text{cluster}_i \mid \text{label})$ is the maximum number of points with the same true label in the ith cluster. For computing the NMI, we construct two random variables C and L from the clustering result and true label, respectively. The probability distributions of random variables are the proportions of different clusters (or classes) over the whole data set. The NMI is computed as follows:

$$\begin{aligned}\text{NMI}&=\frac{\text{MI}\left(C,L\right)}{\sqrt{H\left(C\right)\cdot H\left(L\right)}}\\&=\frac{\sum_{c,l}n_{c,l}\log\left(\left(n\cdot n_{c,l}\right)/\left(n_{c}\cdot n_{l}\right)\right)}{\sqrt{\left(\sum_{c}n_{c}\log\left(n_{c}/n\right)\right)\left(\sum_{l}n_{l}\log\left(n_{l}/n\right)\right)}},\end{aligned}\tag{35}$$

where $\text{MI}(C, L)$ denotes the mutual information of random variables C and L, $H(\cdot)$ denotes the entropy of a random variable, n is the number of data points, $n_{c,l}$ is the number of points in both cluster c and class l, n_c is the points number of cluster c, and n_l is the points number of class l. The values of CA and NMI both vary from 0 and 1, and the higher value means better clustering solution.

5.2. Comparisons of Different Constrained Spectral Clustering. In this subsection, we compare our fast CSC (constrained spectral clustering) algorithm with other spectral clustering algorithms. Following is the list of information of different algorithms in comparison:

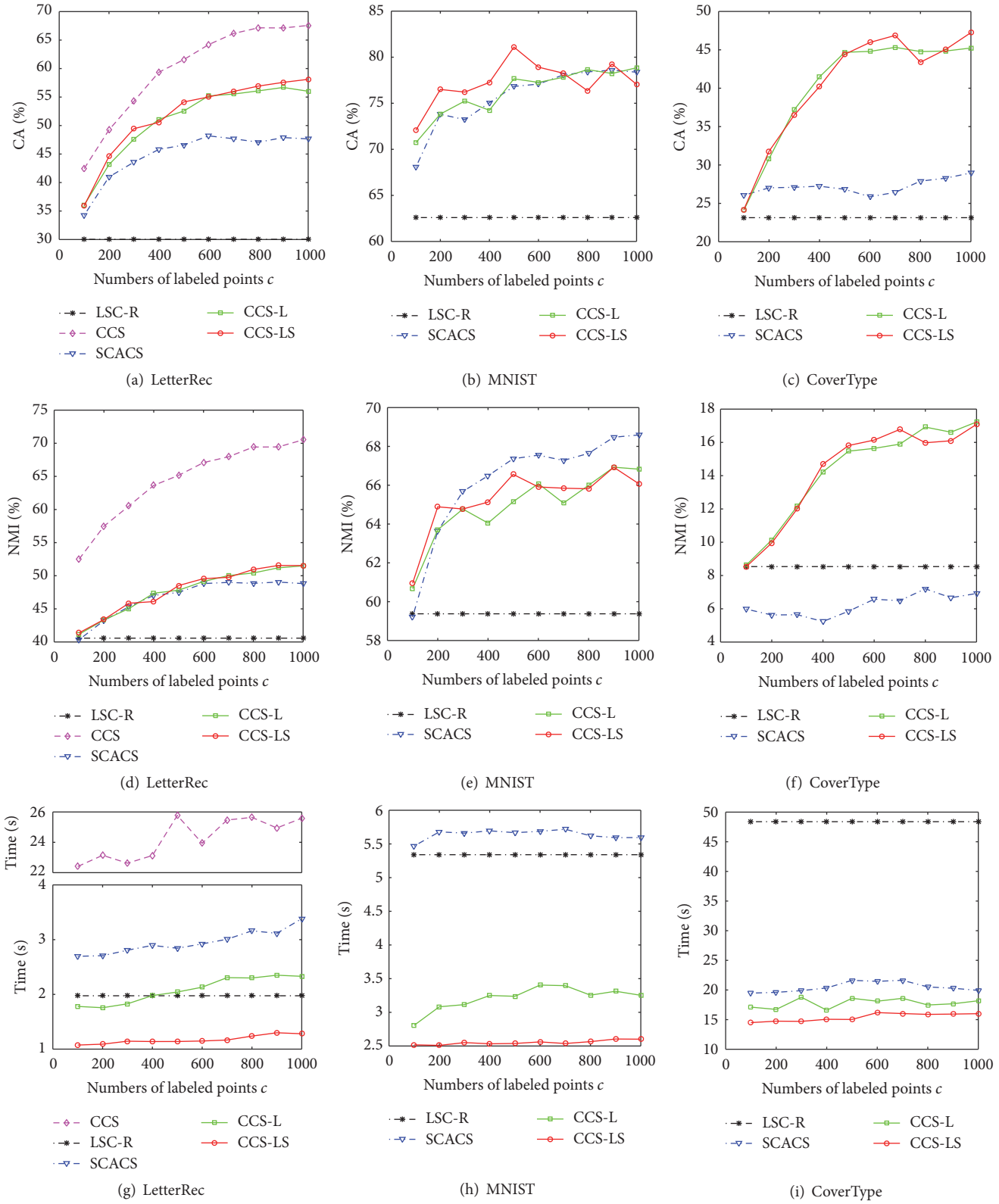

Figure 1: Performance of clustering algorithms with different constraint information.

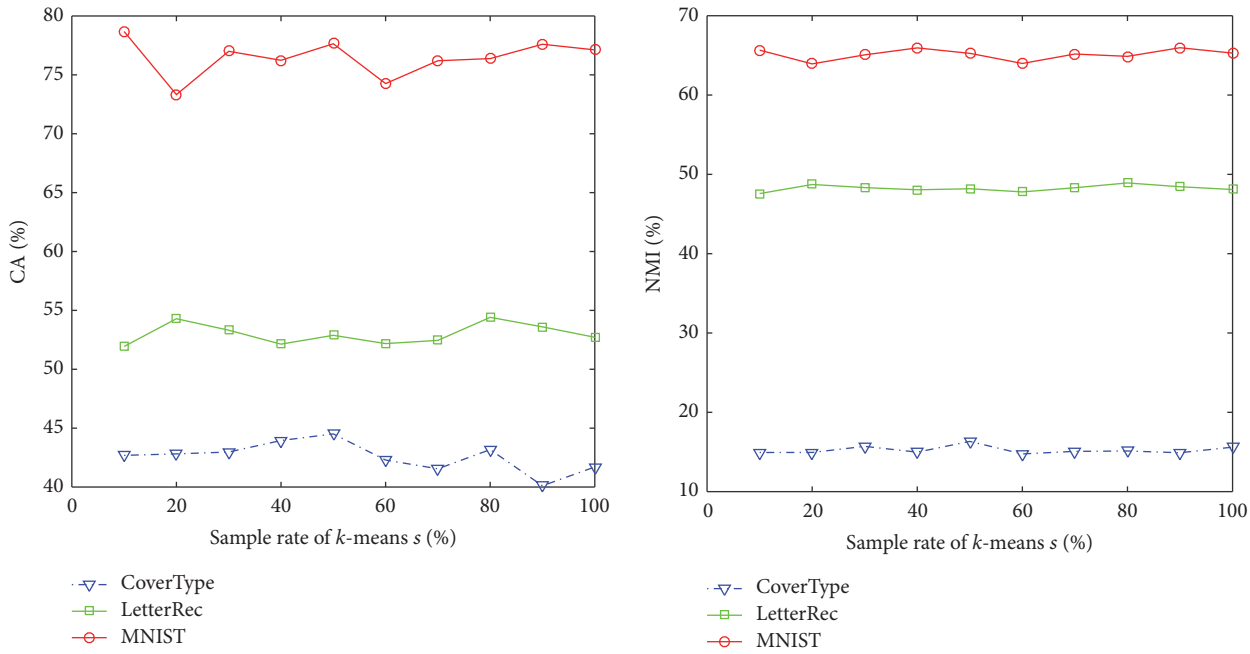

Figure 2: Influence of sample rates on proposed algorithms.

(i) *LSC-R* [20, 21]: the unsupervised spectral clustering baseline with landmark-based graph construction.

(ii) *SCACS* [16]: the most efficient CSC algorithm known and be set as the CSC baseline over MNIST and CoverType data sets.

(iii) *CCS* [17]: the original CSC model proposed in [17], set as the CSC baseline over LetterRec data set. (Since the constructions of the nearest neighbors graphs are both time-consuming on MNIST and CoverType data sets, we do not run CCS algorithm on these two data sets.)

(iv) *CCS-L*: our improved CCS algorithm with landmark-based graph construction.

(v) *CCS-LS*: our improved CCS algorithm with landmark-based graph construction and random sampling.

In the process of the landmark-based graph construction, we fix the number of landmark points $p = 500$ and the number of nearest neighbors $r = 3$. The parameters in SCACS algorithm that we used are $\beta_0 = 0.1$, which is the same as those in [16]. Since in the original model CCS [17] it has been pointed out that α could be a constant number and α was set to 5 in their implementation code, we also set $\alpha = 5$ in CCS, CCS-L, and CCS-LS.

First, we investigate the influence of the number of labeled points c on the performance of algorithms. We vary the value of c from 100 to 1000 with step size 100. For each value of c, we select the c labeled points randomly to produce constraint information and repeat 20 trials with different labeled points sets. The corresponding experimental results are presented in Figure 1. Figures 1(a), 1(b), and 1(c) are related to CA of clustering results, Figures 1(d), 1(e), and 1(f) are related to NMI, and Figures 1(g), 1(h), and 1(i) are related to running time. We can see that our algorithm CCS-LS outperforms LSC-R on all data sets and the values of CA and NMI increase with the growth of constraint information. Those indicate that our algorithm can employ the constraint information appropriately. Compared with SCACS, our algorithm has the similar performances on LetterRec and MNIST data sets and superior performances on CoverType data set, indicating that our algorithm adapts a wider range of geometries. Over the three data sets, the performances of CCS-LS are all close to CCS-L. What is more, our algorithm runs fastest among these algorithms.

Next, we study the influence of random sampling (Step (5) of Algorithm 1) which can be seen in Figure 2. In the experiments, we fix $c = 500$ and change the sample rate from 0.1 to 1 by a step size 0.1. We still run 20 independent trials considering the randomness and compute the means of validity measures. We can see that the values of CA and NMI vary slightly along with the growth of sample rate, verifying the feasibility of random sampling.

5.3. Performance of the Spectral Ensemble Clustering with Random Projection. Since cluster ensemble consists of two parts: basic partition clustering and ensemble clustering, we below combine different basic partition clustering algorithms and different ensemble clustering algorithms to get different cluster ensemble algorithms. Thus, the performance of new ensemble clustering algorithm (Algorithm 2) and new cluster ensemble algorithm (combination of Algorithms 1 and 2) can both be manifested. Following is the list of information of different cluster ensemble algorithms in comparison:

(i) *CK-SE*: the basic partition clustering algorithm "CK" is the constrained k-means clustering algorithm [9],

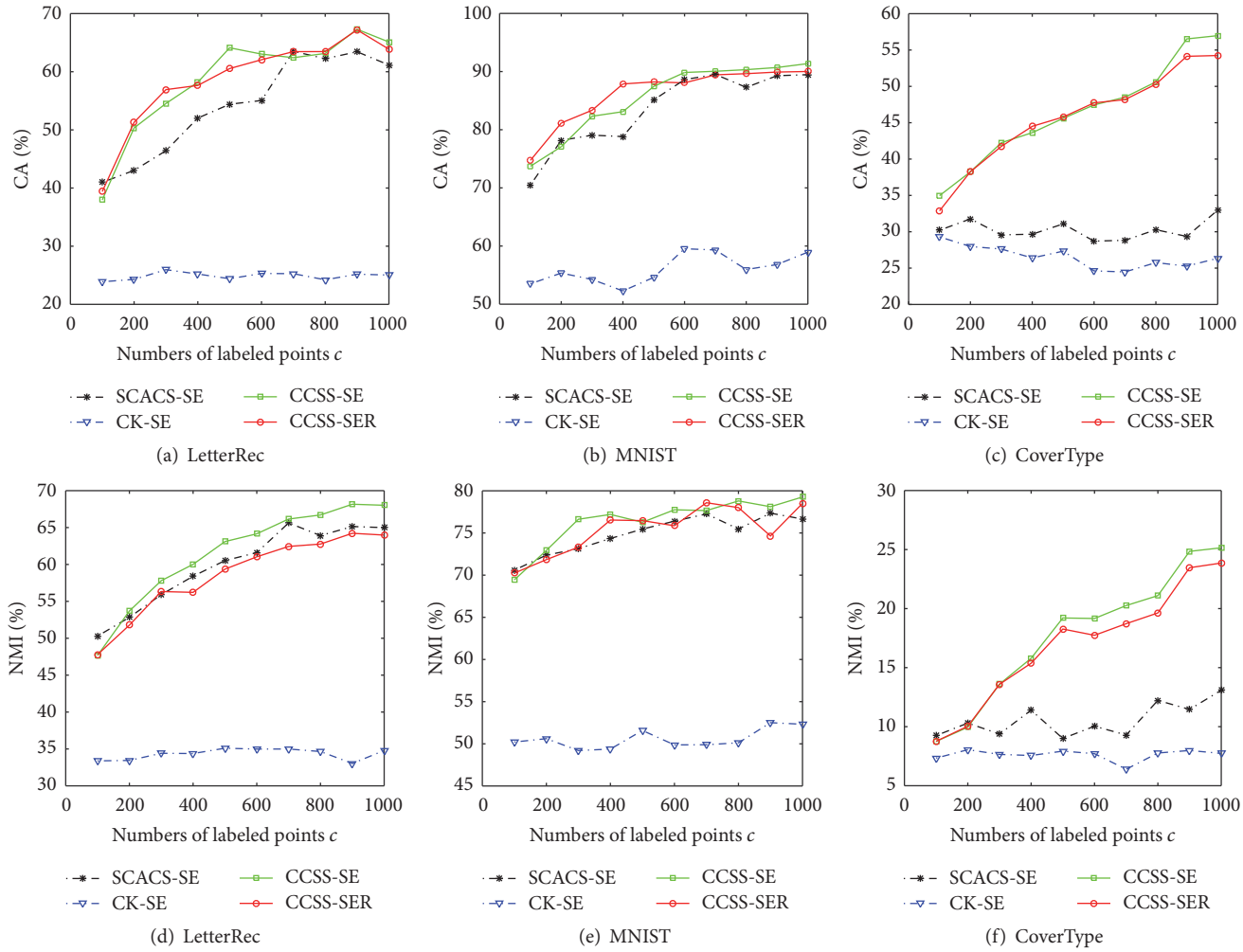

Figure 3: Performance of ensemble clustering algorithms with different constraint information.

and the ensemble clustering algorithm "SE" is the spectral ensemble clustering (SEC) algorithm [22].

(ii) *SCACS-SE*: the basic partition clustering algorithm is SCACS [16] in Section 5.2, and the ensemble clustering algorithm is also SE [22].

(iii) *CCSS-SE*: the basic partition clustering algorithm "CCSS" is our fast CSC algorithm (Algorithm 1), and the ensemble clustering algorithm is also SE [22].

(iv) *CCSS-SER*: the basic partition clustering algorithm is CCSS, and the ensemble clustering algorithm "SER" is our spectral ensemble clustering with random projection (Algorithm 2).

In the phase of basic partition clustering, we fix the number of basic partitions as 50 and the parameters of basic clustering algorithms are the same as those in the last subsection. In addition, similar to the operation of SE [22], the basic partitions are obtained by varying the cluster number from $k - 5$ to $k + 4$. We repeat each cluster ensemble algorithm 10 times and present the average values of results.

First, we show the comparison of different cluster ensemble algorithms in terms of different constraint information in Figure 3. Here the dimensionality rd of CCSS-SER reduced by random projection is 40 and we change the number of labeled points c from 100 to 1000 with step size 100. In the figure, the validity measures of Figures 3(a)–3(c) and Figures 3(d)–3(f) are related to CA and NMI, respectively. Just like the results of last subsection, CCSS-SE has similar performance to that of SCACS-SE on LetterRec and MNIST data sets and has much better performance on CoverType data set. From the comparison between Figure 1 and 3, we can see that the two validity measures are both higher than those of the basic partition dramatically, verifying ensemble clustering's improvement in clustering quality. Compared with CK-SE, CCSS-SE and CCSS-SER both have better performance significantly, which indicates that the basic partitions have an obvious impact on the final result and also verify the high accuracy of our new constrained spectral cluster ensemble method. In addition, the little difference of performance between CCSS-SE and CCSS-SER implies that the random projection can preserve the results of spectral ensemble clustering approximately on different constraint information.

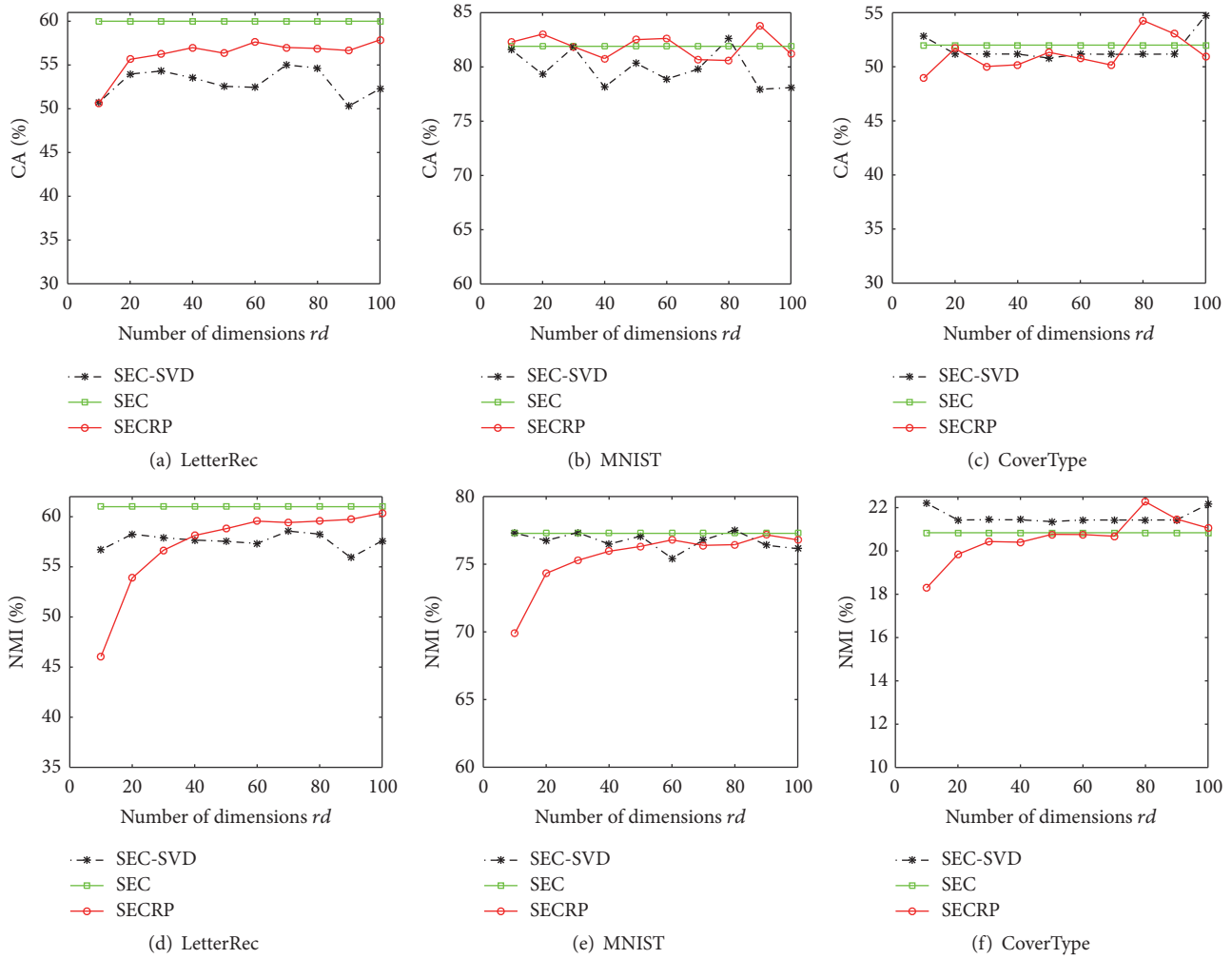

Figure 4: Performance of ensemble clustering algorithms with different dimension.

Table 2: Decrease of running time of SECRP from SEC with different dimensions rd.

rd	10	20	30	40	50	60	70	80	90	100
LetterRec	2.44	2.39	2.11	2.07	2.04	2.03	1.92	1.74	1.67	1.56
MNIST	2.76	2.68	2.66	2.58	2.51	2.34	2.31	2.11	2.16	2.03
CoverType	18.85	18.64	15.34	15.26	14.04	11.43	9.73	8.31	7.72	7.44

Second, we inspect the influence of dimensions of random projection on the performance of our algorithm in Figure 4 and Table 2. In Figure 4, the "SEC-SVD" denotes the SEC algorithm with dimensionality reduction of SVD. When rd is above certain bound, the validity measures of "SECRP" (denote our algorithm SECRP) are almost stable and similar to those of SEC over all three data sets. This indicates that the accuracy of clustering algorithm can be kept when the dimensions surpass a certain bound, which verifies Theorem 7. The small bound of dimensions ($rd = 40$) also reveals the effectiveness of dimensionality reduction of random projection. With respect to SEC-SVD, although it can also preserve the accuracy of clustering algorithm, its running time is not encouraging. Even letting $rd = 20$, the running time comparisons of original algorithm and SVD method over three data sets are 3.47 s/10.85 s, 4.91 s/14.54 s, and 22.06 s/326.61 s. These phenomena may be caused by the tardiness of SVD on large matrix and the breaking of sparseness of binary matrix $\mathbf{B}$. In Table 2, the decrease of running time verifies the efficiency of our new spectral ensemble clustering. Combining this and subfigures (g,h,i) in Figure 1, the efficiency of new constrained cluster ensemble method is also verified. In addition, we can see the decrease of running time caused by random projection is declining with the growth of dimensions, indicating the relative small dimensionality with random projection is preferable.

6. Conclusion

To handle large scale data sets, we propose a fast CSC algorithm. The new algorithm can decrease the space and time complexity of a recently introduced CSC model through landmark-based graph construction and improve its efficiency further by random sampling. The new algorithm not only has the similar property of original model asymptotically, but also is the most efficient and suitable to a wide range of data sets empirically. Taking the new CSC algorithm as basic partition algorithm, we design an efficient semisupervised cluster ensemble algorithm. In the stage of consensus clustering, we reduce the dimensionality of input of spectral ensemble clustering by sparse random projection and prove that the sparse random projection can keep the clustering quality approximately. The experimental results over several data sets also verify the efficiency and effectiveness of new cluster ensemble algorithm. Moreover, in the process of spectral ensemble clustering, the influence analysis of dimensionality reduction with random projection can also give the theoretical guarantee for the weighted k-means clustering with random projection. In the future, we will use techniques such as applying several different basic partition methods, selecting the results of basic partitions, and giving different weights for basic partitions to improve the performance of our cluster ensemble algorithm further.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Nature Science Foundation of China under Grants 61502527 and 61379150 and in part by the Open Foundation of State Key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) (no. SKLNST-2013-1-06).

References

[1] J. Shen, D. Liu, J. Shen, Q. Liu, and X. Sun, "A secure cloud-assisted urban data sharing framework for ubiquitous-cities," *Pervasive and Mobile Computing*, 2017.

[2] Q. Liu, W. Cai, J. Shen, Z. Fu, X. Liu, and N. Linge, "A speculative approach to spatial-temporal efficiency with multi-objective optimization in a heterogeneous cloud environment," *Security and Communication Networks*, vol. 9, no. 17, pp. 4002–4012, 2016.

[3] , *Data Classification: Algorithms and Applications*, C. C. Aggarwal and C. K. Reddy, Eds., CRC Press, New York, NY, USA, 2014.

[4] Y. Zheng, B. Jeon, D. Xu, Q. M. J. Wu, and H. Zhang, "Image segmentation by generalized hierarchical fuzzy C-means algorithm," *Journal of Intelligent and Fuzzy Systems*, vol. 28, no. 2, pp. 961–973, 2015.

[5] Z. Zhou, Q. J. Wu, F. Huang, and X. Sun, "Fast and accurate near-duplicate image elimination for visual sensor networks," *International Journal of Distributed Sensor Networks*, vol. 13, no. 2, 2017.

[6] H. Rong, T. Ma, M. Tang, and J. Cao, "A novel subgraph K+-isomorphism method in social network based on graph similarity detection," *Soft Computing*, pp. 1–19, 2017.

[7] T. Ma, Y. Wang, M. Tang et al., "LED: a fast overlapping communities detection algorithm based on structural clustering," *Neurocomputing*, vol. 207, pp. 488–500, 2016.

[8] Z. Yu, H. Chen, J. You et al., "Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 4, pp. 887–901, 2015.

[9] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proceedings of the 18th International Conference on Machine Learning (ICML '01)*, pp. 577–584, Williams College, Williamstown, Mass, USA, June 2001.

[10] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning with application to clustering with side-information," in *Proceedings of the 15th International Conference on Neural Information Processing Systems (NIPS '02)*, pp. 505–512, Vancouver, Canada, December 2002.

[11] S. D. Kamvar, D. Klein, and C. D. Manning, "Spectral learning," in *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI '03)*, pp. 561–566, Acapulco, Mexico, August 2003.

[12] Z. Lu and M. Á. Carreira-Perpiñán, "Constrained spectral clustering through affinity propagation," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '08)*, Anchorage, Ala, USA, June 2008.

[13] Z. Li, J. Liu, and X. Tang, "Constrained clustering via spectral regularization," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 421–428, Miami, Fla, USA, June 2009.

[14] Z. Lu and H. H. Ip, "Constrained spectral clustering via exhaustive and efficient constraint propagation," in *Proceedings of the 11th European Conference on Computer Vision on Computer Vision (ECCV '10)*, vol. 6316, pp. 1–14, Crete, Greece, September 2010.

[15] X. Wang, B. Qian, and I. Davidson, "On constrained spectral clustering and its applications," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 1–30, 2014.

[16] J. Li, Y. Xia, Z. Shan, and Y. Liu, "Scalable constrained spectral clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 589–593, 2015.

[17] M. Cucuringu, I. Koutis, S. Chawla, G. L. Miller, and R. Peng, "Simple and scalable constrained clustering: a generalized spectral method," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS '16)*, pp. 445–454, Cadiz, Spain, May 2016.

[18] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm," in *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic (NIPS '01)*, pp. 849–856, Vancouver, Canada, December 2001.

[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[20] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation," in *Proceedings of the 25th AAAI Conference on Artificial Intelligence and the 23rd Innovative Applications of Artificial Intelligence Conference*, pp. 313–318, August 2011.

[21] D. Cai and X. Chen, "Large scale spectral clustering via landmark-based sparse representation," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1669–1680, 2015.

[22] H. Liu, T. Liu, J. Wu, D. Tao, and Y. Fu, "Spectral ensemble clustering," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, pp. 715–724, Sydney, Australia, August 2015.

[23] X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: a cluster ensemble approach," in *Proceedings of the 20th International Conference on Machine Learning (ICML '03)*, vol. 3, pp. 186–193, August 2003.

[24] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 835–850, 2005.

[25] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 583–617, 2003.

[26] J. Wu, H. Liu, H. Xiong, and J. Cao, "A theoretic framework of K-means-based Consensus Clustering," in *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI '13)*, pp. 1799–1805, Beijing, China, August 2013.

[27] Z. Yu, L. Li, J. Liu, J. Zhang, and G. Han, "Adaptive noise immune cluster ensemble using affinity propagation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 12, pp. 3176–3189, 2015.

[28] Z. Yu, P. Luo, J. You et al., "Incremental semi-supervised clustering ensemble for high dimensional data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 701–714, 2016.

[29] M. Ye, W. Liu, J. Wei, and X. Hu, "Fuzzy *c*-means and cluster ensemble with random projection for big data clustering," *Mathematical Problems in Engineering*, vol. 2016, Article ID 6529794, 13 pages, 2016.

[30] W. B. Johnson and J. Lindenstrauss, "Extensions of lipschitz mappings into a hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[31] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the 13th Annual ACM Symposium on Theory of Computing*, pp. 604–613, ACM, 1998.

[32] D. Achlioptas, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of Computer and System Sciences*, vol. 66, no. 4, pp. 671–687, 2003.

[33] J. A. Tropp, "Improved analysis of the subsampled randomized Hadamard transform," *Advances in Adaptive Data Analysis. Theory and Applications*, vol. 3, no. 1-2, pp. 115–126, 2011.

[34] D. M. Kane and J. Nelson, "Sparser Johnson-Lindenstrauss transforms," *Journal of the ACM*, vol. 61, no. 1, article 4, 2014.

[35] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas, "Random projections for linear support vector machines," *ACM Transactions on Knowledge Discovery from Data*, vol. 8, no. 4, article 22, 2014.

[36] C. Boutsidis, A. Zouzias, and P. Drineas, "Random projections for κ-means clustering," in *Proceedings of the 24th Annual Conference on Neural Information Processing Systems (NIPS '10)*, pp. 298–306, December 2010.

[37] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for c-means clustering," *IEEE Transactions on Information Theory*, vol. 61, no. 2, pp. 1045–1062, 2015.

[38] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu, "Dimensionality reduction for k-means clustering and low rank approximation," in *Proceedings of the 47th Annual ACM Symposium on Theory of Computing (STOC '15)*, pp. 163–172, June 2015.

[39] Q. Ding and E. D. Kolaczyk, "A compressed PCA subspace method for anomaly detection in high-dimensional data," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7419–7433, 2013.

[40] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06)*, pp. 801–808, Vancouver, Canada, 2006.

[41] M. Popescu, J. Keller, J. Bezdek, and A. Zare, "Random projections fuzzy c-means (RPFCM) for big data clustering," in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE '15)*, pp. 1–6, Istanbul, Turkey, August 2015.

[42] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques: Concepts and Techniques*, Elsevier, 2011.

[43] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time (1+)-approximation algorithm for k-means clustering in any dimensions," in *Proceedings of the 45th Symposium on Foundations of Computer Science (FOCS '04)*, pp. 454–462, Rome, Italy, October 2004.

[44] D. Arthur and S. Vassilvitskii, "k-means++: the advantages of careful seeding," in *Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[45] A. Fahad, N. Alshatri, Z. Tari et al., "A survey of clustering algorithms for big data: taxonomy and empirical analysis," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 3, pp. 267–279, 2014.