Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 7990969, 17 pages
https://doi.org/10.1155/2022/7990969

Research Article

The Core Cluster-Based Subspace Weighted Clustering Ensemble

Xuan Huang,[1,2] Fang Qin,[3] and Lin Lin[4]

[1]*Chengdu College of University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*The School of Information Science and Technology Southwest Jiaotong University, Chengdu 610031, China*
[3]*School of Information Science and Technology, Dalian University of Science and Technology, Dalian 116052, Liaoning, China*
[4]*College of Information Engineering, Chengdu Aeronautic Polytechnic, Chengdu 610100, China*

Correspondence should be addressed to Fang Qin; qinfang@dlust.edu.cn

Received 24 January 2022; Accepted 18 April 2022; Published 24 May 2022

Academic Editor: Nima Jafari Navimipour

In recent years, the Internet of Things (IoT) technology has developed rapidly and is widely used in various fields. It is of great research significance to uncover underlying patterns and insights from the high-dimensional data of IoT, to excavate valuable information to guide people's production and life. Clustering can explore the natural cluster structure of the data, which is conducive to further understanding of the data, and is an essential preprocessing step for data analysis. However, clustering is highly dependent on the data. In order to reduce the complexity of the model, reduce the computational cost, and obtain a more robust clustering solution, we combine subspace clustering and ensemble learning to propose a novel subspace weighted clustering ensemble framework for high-dimensional data. The proposed framework first combines random feature selection and unsupervised feature selection to generate a set of base subspaces. Clustering is performed on each base subspace to achieve a set of subspace clustering solutions that generate a set of adaptive core clusters. The size of the core cluster is between the sample and the cluster. In the ensemble process, the core clusters are viewed as the basic unit, and the stability of the cluster is evaluated by measuring the distance between the core cluster pairs, and the similarity between the core clusters and the clusters in the base subspace, and then weighting the subspace clustering solution. Under this framework, we propose four subspace ensemble approaches based on core cluster to improve the accuracy of consensus clustering solutions. Extensive experiments are conducted on multiple real-world high-dimensional datasets, demonstrating that the proposed framework can process high-dimensional data for the IoT, and the proposed subspace clustering ensemble approaches are superior to the state-of-the-art clustering approaches.

1. Introduction

The Internet of Things is a network platform that combines various information collection devices and the Internet, which facilitates the sharing of information between various devices. Data acquisition devices are widely used in various application fields, and their collected data has the characteristics of high dimension. The impact of high-dimensional data on data mining is twofold. On the one hand, as the data dimension grows, there are more features that describe the sample from different perspectives, which brings richer and more comprehensive information to data analysis; on the other hand, high-dimensional data increases the complexity of the model and make data analysis difficult.

Clustering is an unsupervised learning technique that partitions unlabeled samples into clusters according to the specific criteria. Clustering can reveal the intrinsic properties of the study object and discover its underlying patterns. In the field of data mining and pattern recognition, clustering is often used as data preprocessing, which is the basis for subsequent data analysis. Clustering has always been an active research direction, and in different fields, there are existing multiple clustering algorithms and they achieve satisfactory results. But with the popularity of high-dimensional data, they pose a huge challenge to clustering.

Existing clustering approaches for high-dimensional data can initially be roughly divided into two categories: dimensionality reduction, and then clustering and subspace

clustering. The former typically uses feature selection or feature extraction techniques to extract features from a high-dimensional space that are relevant to subsequent clustering criteria, thereby reducing the dimensionality of the sample. Then, we performed the clustering in the lower-dimensional space. Subspace clustering assumes that the clusters of high-dimensional data are located in low-dimensional subspaces [1], and the goal is to find the clusters hidden in large dimensions.

Clustering is highly dependent on the data. As the dimensions increase, there are a large number of redundant and irrelevant features in the data. This makes the model more complex and the computational complexity grows exponentially. Different clustering algorithms are designed for different types of data, and they can discover the underlying structure of a particular data set, not valid for all types of data. Even with the same clustering algorithm, setting different parameters can group samples into different clusters. As a result, more researchers are focusing on clustering ensemble, trying to achieve better and more robust clustering results by merging multiple base clustering solutions. Recently, spectral clustering [2] has become one of the popular clustering approaches. It first calculates the similarity of the samples to construct an affinity matrix and then decomposes the Laplace vector of the affinity matrix to obtain the eigenvector associated with the feature values. The samples are then mapped to the lower-dimensional space for a final clustering solution.

Spectral clustering [3] can find clusters for complex samples, which is simple to implement and can achieve better results compared to traditional clustering approaches. Recently, some researchers have proposed clustering high-dimensional data through subspace clustering ensemble approaches [4–7]. Some of these employ spectral clustering approach [4, 5]. Cai et al. proposed the spectral clustering approach based on random subspace and graph fusion, termed SC-SRGF [5]. It combines the affinity graph of each subspace with the iterative similarity network fusion scheme and performs spectral clustering on the fusion matrix to obtain the final clustering solution. Huang et al. proposed multidiversified ensemble clustering (MDEC) [4], and under the proposed framework, consistent clustering solutions are obtained by performing spectral clustering, bipartite graph, and hierarchical clustering-based consensus function. These approaches [4–7] use random feature selection to generate a set of subspaces, that is, a certain proportion of features are randomly extracted from the original feature set to generate feature subset. The clusters of each subspace are ensemble to achieve the final clustering solution. However, these approaches [4–6] do not take into account the contributions of different feature subspaces during integration. Literature [7] considers the contribution of each subspace clustering solution in the ensemble process, but it selects members with larger contributions to participate in ensemble. This ensemble selection strategy ignores the feature subspace of small contributions, which is easy to bias the clustering results.

To the abovementioned problem, this paper proposes a novel adaptive core cluster-based subspace weighted clustering ensemble approach, termed C^2SWCE, which considers the similarity of clusters in subspaces in the ensemble process, and the weighted subspace clustering solution helps to improve the accuracy of consensus clustering result. The overall process of C^2SWCE is shown in Figure 1. The proposed framework first uses a combination of random feature selection and unsupervised feature selection to generate a set of base subspaces. Then, spectral clustering is performed in the base subspace to achieve the clustering solution for each base subspace, further generating a set of adaptive core clusters. The core cluster is viewed as the basic unit of clustering ensemble, and the clustering solution of the subspace is weighted by calculating the distance between the core clusters in the subspace, the similarity of the core cluster and the cluster. Finally, we propose four consensus functions under this framework, which combine the locally weighted subspace clustering solutions to achieve the consensus clustering solutions.

The main contributions of our approaches are summarized as follows:

(1) We combine random feature selection and unsupervised feature selection to generate a set of base subspaces. Random feature selection generates a set of random subspaces, and then, unsupervised feature selection is performed on random subspaces, and the features that retain the local structure of random subspaces are selected, which further reduces the dimensionality of random subspaces. This joint feature selection strategy ensures the diversity of feature subspaces and improves the computational efficiency.

(2) We introduce the concept of the core cluster, which is a collection of samples that are grouped into the same cluster in all subspaces. In this paper, the core cluster is viewed as the basic unit of the clustering ensemble, which improves the computational efficiency of the integration process

(3) We propose locally weighted subspace clustering ensemble framework, which evaluates the stability of clusters by calculating the distance of core clusters in each subspace, and the similarity of core clusters and clusters. We further propose four weighted ensemble approaches based on core clusters, which fuse clusters in base subspace and obtain the consensus clustering solution

(4) The experimental results on several real high-dimensional datasets show that the comprehensive performance of our approaches in clustering accuracy and time complexity is significantly better than that of the state of the art clustering approaches

The rest of this paper is organized as follows: We review the related work in Section 2, the locally weighted subspace clustering ensemble approaches are proposed in Section 3, experimental settings and related results and analysis in Section 4, and finally, in Section 5, we conclude this paper.

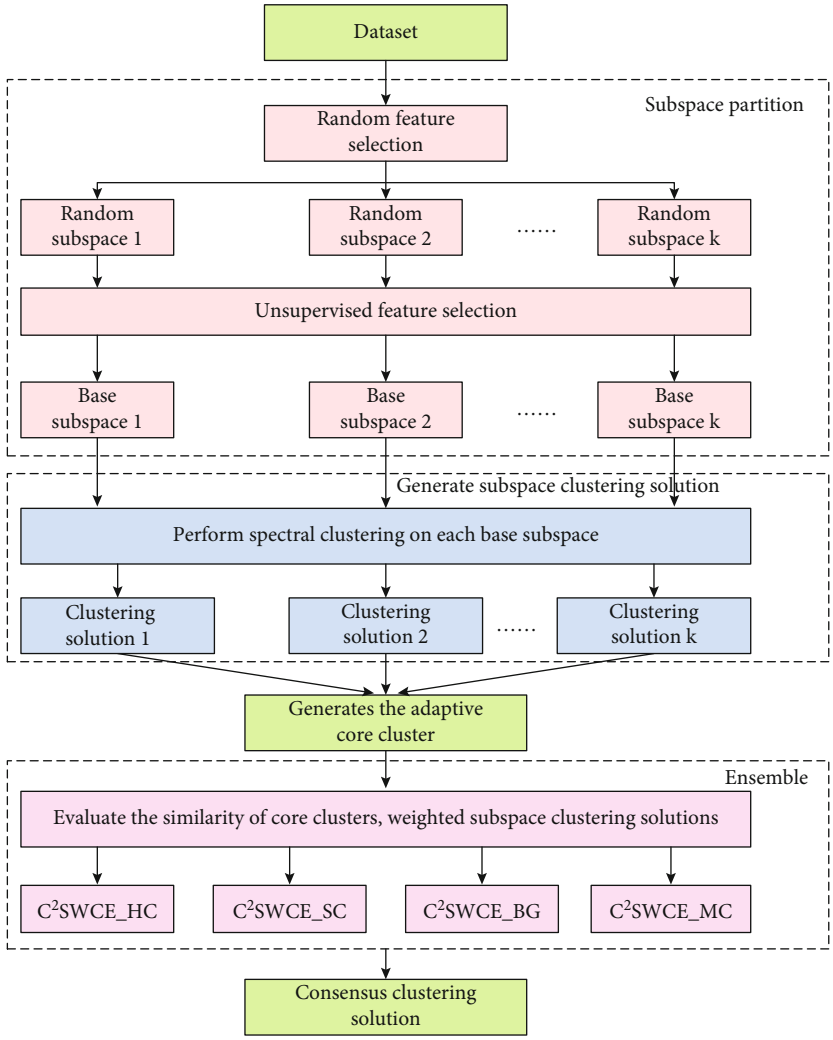


Figure 1: Flow diagram of the C^{2}SWCE.

2. Related Work

Clustering ensemble, also known as cluster integration, combines multiple base clustering solutions to obtain better and more robust consensus clustering result. The cluster integration process consists of two stages: the base clustering generation and the ensemble of clusters. In the first stage, different clustering algorithms are performed on the dataset, or different parameters are set by the same clustering algorithm to partition samples, and the goal of this stage is to generate diverse clustering solutions. In the second stage, the input base clustering solutions are fused by the consensus function to obtain the final clustering result.

The subspace clustering technique can explore the nature cluster structure of high-dimensional data in different low-dimensional spaces. Subspace clustering ensemble is a method that combines subspace clustering and ensemble learning, which fuses clusters in different feature subspaces to achieve consensus clustering solutions. Among the recently proposed subspace clustering ensemble approaches, Huang et al. proposed a multivariate subspace clustering ensemble framework, termed MDEC [4], in which random sampling was used to generate a set of feature subspaces. The novelty of this approach is that the randomizing a scaled exponential similarity kernel is used to get a large number of multivariate measures for each random subspace, forming a metric-subspace pairs. Perform spectral clustering on the similar matrix in each metric-subspace pair to obtain a clustering solution for each subspace, use the entropy criterion to weight the clusters of the subspace, and then use spectral clustering, bipartite graph and hierarchical clustering approaches to get a consensus clustering solution.

Cai et al. proposed a novel spectral clustering approach based on subspace, termed SC-SRGF [5], which first generates a set of random feature subspaces, uses the local structures information of each subspace to form the KNN affinity graph, and then use an iterative similarity network fusion scheme to fuse the affinity graphs of each subspace to obtain a unified affinity graph, and obtain the final clustering solution by performing spectral clustering on a unified affinity graph.

Shankar proposed subspace clustering integration framework AP^{2}CE [6], which uses the affinity propagation to produce a representative subset of features and employs

multiple distance function metric objects to produce diverse subspace clustering solutions and use the Ncut to partition the consensus matrix to get the final result.

Verma et al. proposed double weighting semisupervised ensemble clustering based on selected constraint projection, termed DCECP [7], which treats prior knowledge of experts as pairwise constraints and assigned different subsets of pairwise constraints to different integration members. In addition, an adaptive integration member weighting process is designed to associate different weight values with different integration members. Finally, the final clustering result is obtained using the weighted normalized cut algorithm.

Although many successful subspace ensemble clustering approaches have been developed, most existing approaches [4–6] treat each subspace clustering solution equally during the ensemble process. How to weigh the clustering solution according to the contribution of different feature spaces is worth considering in subspace ensemble.

3. Proposed Framework

In this section, we describe the overall process of the locally weighted subspace clustering ensemble approach. First, we give a brief overview of the proposed methods and then detail the proposed algorithm from three aspects: subspace generation, subspace clustering, and fusion of clusters in subspace.

3.1. Overview. In this paper, we introduce a locally subspace weighted clustering ensemble framework. First, we randomly select k feature subset from the original feature set to generate k random subspaces, and then selecting features representing the local structure from each random subspace to generate base subspace. Then, spectral clustering is performed on each base subspace to achieve subspace clustering solutions. The adaptive core clusters are generated from the base subspace clustering solution, which is a set of samples that are grouped into the same cluster in all subspaces. Then, by calculating the distance of the core cluster in different subspaces, the stability of the clusters is evaluated and weighted the base subspace clustering solution. Finally, in order to integrate the clustering solutions of the base subspaces and get the consensus results, in the C^{2}SWCE framework, we propose four adaptive core cluster-based consensus functions, to achieve the final clustering solutions.

Let $X \in R^{n\times m}$ be the input dataset that contains n samples, each with m features. Let x_i $(i = 1, \cdots n)$ denotes the i th sample, which corresponds to the ith row of X, so that the input data can be represented as $X = (x_1, x_2, \cdots, x_n)$. Let f_i $(j = 1, \cdots m)$ denotes the ith feature of the sample, which corresponds to the jth column of the X; therefore, matrix X can also be represented as $X = (f_1, f_2, \cdots, f_m)$.

3.2. Generate Subspaces. The natural cluster structure of high-dimensional data is hidden in low-dimensional subspaces [8]. Subspace clustering explores the possibility of grouping samples in different feature sets. Subspace clustering ensemble fuses clusters in different subspaces to obtain the final clustering solution, which is an effective method for clustering high-dimensional data. The recently proposed subspace clustering ensemble approaches [5, 9], which uses random feature selection to generate a set of random subspaces. Random subspace has diversity; it explored the potential cluster of high-dimensional dataset from the perspective of different features to achieve diverse subspace clustering solutions. Literature [10] uses stratified sampling method to generate feature groups and verifies that this method is superior to random sampling. This paper uses the random feature selection to obtain a variety of subspaces, and unsupervised feature selection is performed again on the random subspace to select the features that retain the local structure of random subspace.

The specific procedure of feature subspace segmentation is as follows. The original feature set $f_1, f_2 \cdots, f_m$ is randomly sampled according to the sampling ratio r, the original features are classified into k feature groups $\mathcal{G}_1, \mathcal{G}_2, \cdots, \mathcal{G}_k$, where $\mathcal{G}_1 (i = 1, \cdots k)$ represents the features contained in the ith random subspace. Let $\widetilde{X_i} \in R^{n\times|\mathcal{G}_i|}$ denotes the matrix corresponding to the ith random subspace, n is the number of samples, and $|\mathcal{G}_i| (i = 1, \cdots, k)$ is the number of features in the ith random subspace. Obviously, it holds that $\forall \mathcal{G}_i \neq \phi (i = 1, \cdots, k)$ and $|\mathcal{G}_i| = m \cdot r$.

A set of base subspaces is generated by calculating the Laplace score [11] of features in each random subspace and selecting the important features in random subspace. Then, we construct the KNN graph for each random subspace, which represents the local structure of the random subspace. The KNN graph of the ith subspace is defined as follows:

$$G^i_{\mathrm{KNN}} = \left\{ V^i, E^i \right\}, \tag{1}$$

where $V^i = \{v_1, v_2, \cdots, v_n\}$ is the set of nodes corresponding to the samples $x_1, x_2, \cdots, x_n$ in the ith subspace and E^i is the edge set of the ith subspace. We use the Gaussian kernel function to calculate the weights of the edges between the nodes in the subspace and their corresponding KNN nodes. E^i is defined as

$$E^i = \left\{ e^i_{\alpha\beta} \right\}_{n\times n}, \tag{2}$$

where

$$e^i_{\alpha\beta} = \begin{cases} \exp\left(-\dfrac{\left\| x^i_\alpha - x^i_\beta \right\|^2}{2\sigma^2} \right), & \text{if } x^i_\alpha \in \mathrm{KNN}\left(x^i_\beta\right) \text{ or } x^i_\beta \in \mathrm{KNN}\left(x^i_\alpha\right), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

In Equation (3), x^i_α and x^i_β are the two samples in the ith subspace, respectively. $\|x^i_\alpha - x^i_\beta\|^2$ is the Euclidean distance between x^i_α and x^i_β, and $\mathrm{KNN}(\cdot)$ is the K-nearest neighbor (KNN) operator, and σ is the mean of Euclidean distance between the sample and its KNN.

Let $D^i = \mathrm{diag}(\mathrm{E}^{\mathrm{i}}1)$ be the degree matrix of the ith subspace, where $1 = [1,\cdots,1]^{\mathrm{T}}$. The Laplacian of the graph is calculated as follows:

$$L^i = D^i - E^i. \tag{4}$$

In the ith subspace, the Laplace score of f_l^i $(l = 1, \cdots |\mathscr{G}_i|)$ is

$$L_{f_l}^i = \frac{\tilde{f}_l^{i\top} L^i \tilde{f}_l^i}{\tilde{f}_l^{i\top} D^i \tilde{f}_l^i}, \tag{5}$$

where

$$\tilde{f}_l^i = f_l^i - \frac{f_l^{i\top} D^i 1}{1^\top D^i 1} 1. \tag{6}$$

We calculate the Laplace score for each feature in the subspace and arrange them in descending order to select the top d features. We determine the number of second selected features based on the number of features in random subspace. Let $\tilde{r}$ represents unsupervised feature selection ratio in the random subspace, that is, from each random feature group $\mathscr{G}_i$ $(i = 1, \cdots k)$, select important features that represent the local structure of the random subspace to generate $\tilde{\mathscr{G}}_i$ $(i = 1, \cdots k)$, it holds that $\tilde{\mathscr{G}}_i \subset \mathscr{G}_i, |\tilde{\mathscr{G}}_i| = |\mathscr{G}_i| \cdot \tilde{r}(i = 1,\cdots k)(i = 1, \cdots k)$, and $\sum_{i=1}^k |\tilde{\mathscr{G}}_i| \ll m$. For convenience, let $S_i \in R^{n\times|\tilde{\mathscr{G}}_i|}$ $(i = 1, \cdots k)$ represent the matrix of the ith base subspace, where n is the number of samples and $\mathscr{G}_i$ is the number of features.

3.3. Generate Subspace Clustering Solution. Let $S_1, S_2, \cdots, S_k$ represent the base subspaces generated by the joint feature selection strategy, where$S_i \in R^{n\times d_i}$ $(i = 1, \cdots k)$ is the ith base subspace, and $d_i = |\tilde{\mathscr{G}}_i|$. Let $P^{(S_i)}$ $([i = 1, \cdots, k])$ be the clustering solution for S_i, formally, the clustering solution for the base subspace is $P^{(S_i)} = \{P^{(S_i)}(x_1), P^{(S_i)}(x_2), \cdots, P^{(S_i)}(x_n)\}$, where $P^{(S_i)}(x_j)$ $(j = 1, \cdots, n)$ is a cluster that contains x_j in S_i. Let $C_l^{(S_i)}(l = 1, \cdots, K_{S_i})$ denote the lth cluster in the S_i, then for $\forall x_j \in C_l^{(S_i)}$, it holds that $P^{(S_i)}(x_j) = C_l^{(S_i)}$. Thus, the clustering solution of S_i also be denoted as

$$P^{(S_i)} = \left\{C_1^{(S_i)}, C_2^{(S_i)}, \cdots, C_{K_{S_i}}^{(S_i)}\right\}, \tag{7}$$

where K_{S_i} is the number of clusters in S_i. Then, we can get the clustering solution set of k base subspace, which is represented as $\mathscr{P} = \{P^{(S_1)}, P^{(S_2)}, \cdots, P^{(S_k)}\}$. Subspace clustering ensemble is the fusion of clustering solutions from multiple subspaces to achieve consistent results $\mathscr{P}^*$.

In the clustering process, most approaches view input samples as base units and group them into different clusters. However, when there are more samples, the computational complexity also increases significantly. Huang et al. introduce the concept of superobject [12], which is defined as in the base clustering ensemble where two samples are partitioned into the same cluster, the two samples are in the same original superobject. The size of superobject is between clusters and samples, and it has been proven that viewing superobject as base units when integrated can significantly improve the scalability of data size and simplify the calculation.

Inspired by the concept of the superobject [12], this paper groups samples in base subspaces to achieve a set of base subspace clustering solutions $\mathscr{P}$ and then generates the adaptive core clusters. The core clusters in high-dimensional space are defined as follows.

(1) $\forall x_\alpha, x_\beta \in X, \quad \forall S_i (1 \le i \le k), \quad P^{(S_i)}(x_\alpha) = P^{(S_i)}(x_\beta)$, then$x_\alpha, x_\beta \in o$

(2) $\forall x_\alpha \in o, x_\beta \notin o, \exists S_i (1 \le i \le k), P^{(S_i)}(x_\alpha) \ne P^{(S_i)}(x_\beta)$

Definition 1. Let X be the input dataset and $\mathscr{P}$ be set of base subspace clustering solutions. Samples that simultaneously satisfying the following two conditions are core clusters, which are denoted as o.

Let $O = \{o_1, o_2, \cdots, o_\gamma\}$ be the set of core clusters in a high-dimensional space, where γ is the number of core clusters. It holds that $\forall o_i \ne \varnothing$, $o_i \cap o_j = \varnothing (\forall i \ne j, i, j = 1, 2, \cdots, \gamma)$, and $\bigcup_{i=1}^{\gamma} o_i = n$.

We provide examples of samples clustering in different subspaces. Given a dataset $X = \{x_1, \cdots, x_7\}$, where x_i is the ith sample. These samples are grouped into clusters in subspaces, and the relationships between the samples, core clusters, and clusters are described in Table 1. In subspace S_1, samples are grouped into two clusters, and in S_2 and S_3, they are grouped into three clusters, respectively.

According to Definition 1, there are four core clusters are generated in the above example. The relationship between the core cluster and the samples is shown in Table 1, namely, $x_1, x_2 \in o_1$, $x_3 \in o_2$, $x_4, x_5 \in o_3$, and $x_6, x_7 \in o_4$. The core cluster is viewed as the basic units of cluster in subspace, for example in base subspace S_1, $o_1, o_2 \subset C_1^{(S_1)}$, $o_3, o_4 \subset C_2^{(S_1)}$.

In the conventional subspace clustering approaches, samples are basic units. However, samples correspond to different features in different feature subspaces, so the implicit relationship between samples in different subspaces is also different. A core cluster is a set of samples that are grouped into the same clusters in all subspaces. In the process of subspace ensemble, we view the core cluster as the basic unit, that is, each cluster in subspace is composed of core clusters. We evaluate the stability of clusters in subspace by measuring the distance of core clusters pairs contained in the cluster.

Definition 2. $\forall S_i$ $(1 \le i \le k)$, $\forall o_\alpha, o_\beta$ $(\alpha, \beta = 1, \cdots \gamma)$, the distance between the core clusters pairs is defined as

$$\mathrm{CCS}^{(S_i)}(o_\alpha, o_\beta) = \frac{1}{|o_\alpha||o_\beta|} \sum_{x_a \in o_\alpha} \sum_{x_b \in o_\beta} d^{(S_i)}(x_a, x_b), \tag{8}$$

where $|o_\alpha|$ and $|o_\beta|$ are the number of samples contained in o_α and o_β, respectively. $d^{(S_i)}(\cdot,\cdot)$ is a distance metric

Table 1: Correspondence between samples, core clusters, and clusters in subspace.

X	x_1	x_2	x_3	x_4	x_5	x_6	x_7
S_1	$C_1^{(S_1)}$	$C_1^{(S_1)}$	$C_1^{(S_1)}$	$C_2^{(S_1)}$	$C_2^{(S_1)}$	$C_2^{(S_1)}$	$C_2^{(S_1)}$
S_2	$C_1^{(S_2)}$	$C_1^{(S_2)}$	$C_2^{(S_2)}$	$C_2^{(S_2)}$	$C_2^{(S_2)}$	$C_3^{(S_2)}$	$C_3^{(S_2)}$
S_3	$C_1^{(S_3)}$	$C_1^{(S_3)}$	$C_1^{(S_3)}$	$C_2^{(S_3)}$	$C_2^{(S_3)}$	$C_3^{(S_3)}$	$C_3^{(S_3)}$
O	o_1	o_1	o_2	o_3	o_3	o_4	o_4

function between samples in S_i, which can be selected as the Euclidean distance, the Manhattan distance, the Cosine distance, and so on. If the distance between the core clusters pairs is smaller, the more stable the clusters in the corresponding subspace, and vice versa. Clusters are evaluated by the average distance between core clusters.

For $\forall S_i$ $(1 \le i \le k)$, there are a total of K_{S_i} clusters, and each cluster $C_j^{(S_i)}(1 \le j \le K_{S_i})$ contains $\tilde{n}_j^{(S_i)}$ core clusters. Obviously, it holds that $\tilde{n}_j^{(S_i)} \ge 1$ and $\sum_{j=1}^{K_{S_i}} \tilde{n}_j^{(S_i)} = \gamma$.

Definition 3. For $\forall C_j^{(S_i)}$ $(1 \le i \le k, 1 \le j \le K_{S_i})$, the average distance of the core clusters is defined as

$$\mathrm{ACS}_j^{(S_i)} = \begin{cases} \dfrac{2}{\tilde{n}_j^{(S_i)}\left(\tilde{n}_j^{(S_i)} - 1\right)} \displaystyle\sum_{a=1}^{\tilde{n}_j^{(S_i)}} \sum_{b=a+1}^{\tilde{n}_j^{(S_i)}} \mathrm{CCS}^{(S_i)}(o_a, o_b), & \tilde{n}_j^{(S_i)} > 1, \\ 0, & \tilde{n}_j^{(S_i)} = 1. \end{cases} \tag{9}$$

In Equation (9), $\mathrm{CCS}^{(s_i)}(\cdot)$ is a distance metric function of core clusters pairs in S_i and $\tilde{n}_j^{(S_i)}$ is the number of core clusters in $C_j^{(S_i)}$.

The smaller the average distance between the core clusters in a cluster, the denser the cluster, that is, the greater the probability that the core clusters will be grouped into clusters in the base subspace, the higher the stability of the clusters. We introduced the cluster stability index (CSI) to describe this relationship.

Definition 4. For $\forall C_j^{(S_i)}$ $(1 \le i \le k, 1 \le j \le K_{S_i})$, its cluster stability index is defined as

$$\mathrm{CSI}_j^{(S_i)} = e^{-\mathrm{MCS}_j^{(S_i)}}, \tag{10}$$

where $\mathrm{MCS}_j^{(S_i)}$ is the average distance between the core clusters in $C_j^{(S_i)}$. For $\mathrm{MCS}_j^{(S_i)} \in [0, +\infty)$, the range of the cluster stability index is $\mathrm{CSI}_j^{(S_i)} \in (0, 1]$.

The CSI is an indicator that describes clusters in a subspace. If the cluster has only one core cluster, the distance between the core cluster is 0, and the stability index of the corresponding cluster is 1; the cluster can no longer be divided and is stable. If there are multiple core clusters in a cluster, the greater the average distance between the core clusters, the worse the stability of the clusters, and vice versa.

Thus, for $\forall S_i (1 \le i \le k)$, the average of the stability index for all clusters is

$$\mathrm{CSI}^{(S_i)} = \frac{1}{K_{S_i}} \sum_{j=1}^{K_{S_i}} CSI_j^{(S_i)} \quad \left(1 \le j \le K_{S_i}\right). \tag{11}$$

When fusing the clustering solutions of base subspace, we weight the subspace by the stability of the clusters in the corresponding base subspace. The weight of each base subspace is calculated by

$$\omega^{(S_i)} = \frac{\mathrm{CSI}^{(S_i)}}{\sum_{i=1}^{k} \mathrm{CSI}^{(S_i)}}, \tag{12}$$

where

$$\sum_{i=1}^{k} \omega^{(S_i)} = 1. \tag{13}$$

3.4. Subspace Clustering Ensemble. Clustering ensemble is an effective way to improve robustness and stability of clustering solution [13]. We propose four core cluster-based consensus functions that ensemble clustering solutions for each base subspace to achieve the final clustering results.

Define the core cluster similarity matrix in base subspace based on whether any two core clusters in the subspace are grouped into the same cluster.

Definition 5. For $\forall S_i$, the core cluster similarity matrix is defined as

$$A^{(S_i)} = \left\{ a_{\alpha\beta}^{(S_i)} \right\}_{\gamma \times \gamma} \quad (\alpha, \beta \in [1, \gamma], i \in [1, k]), \tag{14}$$

where

$$a_{\alpha\beta}^{(S_i)} = \begin{cases} \mathrm{CCS}^{(S_i)}\left(o_\alpha, o_\beta\right), & \text{if } P^{(S_i)}(o_\alpha) = P^{(S_i)}\left(o_\beta\right), \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

In Equation (15), $P^{(S_i)}(o_\alpha)$ and $P^{(S_i)}(o_\beta)$ represent clusters containing o_α and o_β in S_i, respectively. Unlike normal similarity matrices, we define the core cluster similarity matrices rather than similar matrices between samples. At the same time, we achieve the weighted core cluster similarity matrix based on the weights of base subspace, which is represented as

$$A - \frac{1}{k} \sum_{i=1}^{k} \omega^{(S_i)} \cdot A^{(S_i)}. \tag{16}$$

3.4.1. Hierarchical Clustering Based on the Core Cluster. In this section, we propose the consensus function for hierarchical clustering based on the core clusters, termed C^{2}SWCE_HC. The proposed method views the core cluster as the basic unit and constructs a dendrogram by the core cluster, where the root of the tree corresponds to the dataset and its leaves correspond to all the core clusters in $\mathscr{P}$. Each level of the dendrogram represents the clustering results of different numbers of core clusters, and the final clustering solution can be achieved by specifying a specific level of the dendrogram.

The specific steps of integration are as follows: First, the γ core clusters in $\mathscr{P}$ as the initial region, which is represented as

$$\Theta^{(0)} = \left\{\Theta_1^{(0)}, \cdots, \Theta_\gamma^{(0)}\right\}. \tag{17}$$

In Equation (17), $\Theta_i^{(0)} = \{o_i\}$ denotes the ith initial region, which corresponds to the ith core cluster. Let $M^{(0)}$ be the initial similar matrix. Merge the two most similar core clusters in $M^{(0)}$ into one larger cluster and update the similar matrix according to the average link for the next merge of core clusters. Repeat merge the two most similar regions of the similar matrix that were updated in the previous iteration. After each merge, the number of regions is reduced by 1, and after the ith iteration, the merged regions are represented as $\Theta^{(l)} = \{\Theta_1^{(l)}, \cdots, \Theta_{\gamma-l}^{(l)}\}$, and the corresponding similar matrix is updated as follows:

$$M^{(l)} = \left\{\frac{1}{\left|\Theta_i^{(l)}\right|\left|\Theta_j^{(l)}\right|}\sum_{o_\alpha\in\Theta_i^{(l)}, o_\beta\in\Theta_j^{(l)}} a_{\alpha\beta}^{(l)}\right\}_{(\gamma-l)\times(\gamma-l)}, \tag{18}$$

where $|\Theta_i^{(l)}|$ is the number of core clusters contained in the i th region after the ith iteration, and the maximum number of iterations of the dendrogram is $\gamma - 1$.

For clarity, C^{2}SWCE_HC is summarized in Algorithm 1.

3.4.2. Spectral Clustering Based on the Core Cluster. In this section, we introduce the core cluster-based spectral clustering consensus function to ensemble subspace clustering solutions. First, we build the affinity graph that treats the core clusters as graph nodes and the weighted core cluster similarity matrix as the adjacency matrix. The graph is defined as

$$G = (V, E), \tag{19}$$

where $V = O$ is the nodes set and E is the edge set. The weights of the edge between the nodes v_i and v_j is determined by matrix A, that is, $E_{ij} = A_{ij}$. Let $D \in R^{\gamma\times\gamma}$ be the degree matrix of E, the normalized graph Laplacian is computed as

$$L = D - E = I - D^{-1/2}ED^{-1/2} = I - D^{-1/2}AD^{-1/2}. \tag{20}$$

We perform eigendecomposition on L to achieve the first K eigenvalues, and matrix $\Xi \in R^{\gamma\times K}$ is constructed by corresponding K eigenvectors and perform K-means on the row vectors of the matrix.

For clarity, the C^{2}SWCE_SC algorithm is summarized in Algorithm 2.

3.4.3. Bipartite Graph Partition Based on the Core Clusters. Under the C^{2}SWCE framework, we propose the bipartite graph clustering ensemble based on the core clusters, termed C^{2}SWCE_BG. In ensemble process, we use the core clusters and clusters as graph nodes to construct the bipartite graph. The specific steps are described below.

In different base subspaces, the core clusters are grouped into a set of clusters. In all base subspaces, the set of clusters is

$$\mathscr{C} = \left\{C_1, C_2, \cdots, C_{n_c}\right\}, \tag{21}$$

where $n_c = \sum_{i=1}^{k} K_{S_i}$ is the number of clusters in $\mathscr{P}$. We view both the core cluster and cluster as graph nodes, and construct the bipartite graph. That is:

$$G_{\mathrm{BG}} = (U_{\mathrm{BG}}, V_{\mathrm{BG}}, E_{\mathrm{BG}}), \tag{22}$$

where $U_{\mathrm{BG}} = \mathrm{OUC}$ is the node set corresponding to the core cluster and the cluster; $V_{\mathrm{BG}} = \mathscr{C}$ is the node set corresponding to the cluster in $\mathscr{P}$; it holds that $|U_{\mathrm{BG}}| = \gamma + n_c$, $|V_{\mathrm{BG}}| = n_c$. E_{BG} is the edge set.

In $\mathscr{C}$, clusters in the same subspace contain different core clusters, while clusters in different subspaces may contain the same core cluster. Therefore, we use the Jaccard coefficient to measure the similarity of clusters. The core clusters are viewed as base units; the similarity matrix between clusters is defined as

$$Z = \left\{z_{ij}\right\}_{n_c\times n_c}, \tag{23}$$

where

$$z_{ij} = J\left(C_i, C_j\right) = \frac{\left|C_i \cap C_j\right|}{\left|C_i \cup C_j\right|} \quad (i, j \in [1, n_c]). \tag{24}$$

Clusters in the same base subspace do not intersect, so the Jaccard coefficient between clusters in the same subspace is 0. The similarity between clusters pairs is calculated by Equation (24) and the clusters are weighted based on similarity.

The similarity matrix between clusters and the core clusters is constructed as

$$\tilde{Z} = \left\{\tilde{z}_{ij}\right\}_{\gamma\times n_c}, \tag{25}$$

where

$$\tilde{z}_{ij} = \begin{cases} \omega^{(S_i)} \cdot \mathrm{CSI}_j, & \text{if } o_i \subseteq C_j, \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Input: $S_1, S_2, \cdots, S_k$, K
Output: $\mathscr{P}^*$
1. Performed clustering on each base subspace to generate a set of subspace clustering ensemble $\mathscr{P}$
2. Generate the adaptive core clusters according to Definition 1
3. Calculates the average distance between core clusters according to Equations (8) and (9)
4. Calculates the CSI of the cluster according to Equation (10), and $\omega^{(S_i)}$ is achieved according to Equation (11)–(13)
5. Construct the core cluster similarity matrix $A^{(S_i)}$ $(i = 1, \cdots, k)$ according to Equations (14) and (15), and the weighted core cluster similarity matrix A according to Equation (16).
6. Initialize $\Theta^{(0)} = \{\Theta_1^{(0)}, \cdots, \Theta_\gamma^{(0)}\}$, $M(0) = A$
7. Construct the dendrogram
    for $l = 1$ to $\gamma - 1$ do
      According to $M^{(l-1)}$, merge the two most similar regions to achieve $\Theta^{(l-1)}$
      Update $\Theta^{(l)}$, and achieve $M^{(l)}$
    end for
8. Select the level of the dendrogram according to K, and achieve K clusters
9. Map the labels of core cluster to the samples.

Algorithm 1

Input: $S_1, S_2, \cdots, S_k$, K
Output: $\mathscr{P}^*$
1. Performed clustering on each base subspace to generate a set of subspace clustering ensemble $\mathscr{P}$
2. Generate the adaptive core clusters according to Definition 1
3. Calculates the average distance between core clusters according to Equations (8) and (9)
4. Calculates the CSI of the cluster according to Equation (10), and $\omega^{(S_i)}$ is achieved according to Equations (11)–(13)
5. Construct the core cluster similarity matrix $A^{(S_i)}$ $(i = 1, \cdots, k)$ according to Equations (14) and (15), and the weighted core cluster similarity matrix A according to Equation (16).
6. Build the graph $G = (V, E)$ with $V = O$,$E = A$
7. Constructed the normalized graph Laplacian L according to Equation (20).
8. Perform eigendecomposition on L to achieve the first K eigenvalues and corresponding eigenvectors to build Ξ
9. After normalizing Ξ, perform K-means to categorized the core clusters
10. Map the labels of core cluster to the samples

Algorithm 2

In Equation (26), CSI_j is the cluster stability index of C_j in S_l, $\omega^{(S_l)}$ is the weight of C_j in S_l ($l[l \in [1, k]], i \in [i, \gamma], j \in [1, n_c]$).

Connect the matrix $\tilde{Z}$ and matrix Z to generate matrix $E_{BG} \in R^{(\gamma+n_c)\times n_c}$, that is,

$$E_{\mathrm{BG}} = \left\{e_{ij}\right\}_{(\gamma+n_c)\times n_c} \tag{27}$$

The entry of E_{BG} corresponds to the weight of the edge between the two nodes in G_{BG}, denoted as

$$e_{ij} = \begin{cases} \omega^{(S_l)} \cdot \mathrm{CSI}_j, & \text{if} \quad u_i \in \mathscr{C}, v_j \in O, v_j \subseteq u_i, \\ J\left(u_i, v_j\right), & \text{if} \quad \tilde{u}_i \in \mathscr{C}, \tilde{v}_j \in \mathscr{C}, \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

In Equation (28), CSI_j is the cluster stability index of the jth cluster. $J(\cdot, \cdot)$ is the Jaccard coefficient operator, which is calculated according to Equation (24). In G_{BG}, there are no edges between cluster nodes or between the core cluster nodes, only edges between cluster nodes and core cluster nodes.

According to the corresponding features in different base subspaces, we use the weight of the subspace corresponding to the cluster to which the core cluster belongs as the weight of the edges of the core cluster and the cluster. The higher the stability of the cluster, the greater its impact during integration and the greater the weights assigned.

Finally, we use Tcut [14] to segment G_{BG}. All nodes are partitioned into K disjoint groups. The samples contained in the core clusters and cluster of the same group are partitioned in the same cluster.

The specific steps of C^{2}SWCE_BG are summarized in Algorithm 3.

3.4.4. Metacluster-Based Ensemble Clustering. Under the proposed C^{2}SWCE framework, we propose the metacluster-based consensus clustering algorithm, termed C^{2}SWCE_MC. In the proposed approach, clusters are regarded as the basic units that use the similarities between them to divide clusters into

Input:$S_1, S_2, \cdots, S_k, K$
Output: $\mathscr{P}^*$
1. Performed clustering on each base subspace to generate a set of subspace clustering ensemble $\mathscr{P}$
2. Generate the adaptive core clusters according to Definition 1
3. Calculated the similarity of the cluster according to Equations (23) and (24)
4. Calculate the similarity of clusters and core clusters according to Equations (25) and (26)
5. Construct the bipartite graph according to Equation (22) with $U = \mathrm{OUC}$, $V = \mathscr{C}$
6. Combine matrix $\tilde{Z}$ and matrix Z to generate E_{BG}
7. Segment the G_{BG} by Tcut
8. Map the labels of core cluster to the samples

Algorithm 3

different groups, and the samples of the same group are divided into the same cluster.

C^{2}SWCE_MC first treats all clusters as nodes and constructs the similar graph, which is defined as:

$$G_M = (V_M, E_M), \tag{29}$$

where $V_M = \mathscr{C}$ is the node set for all clusters in $\mathscr{P}$ and E_M is the edge set. The weights of the edges between nodes are defined by the similarity matrix Z of the clusters by Equations (23) and (24).

$$E_M = \left\{e_{ij}\right\}_{n_c \times n_c}, \tag{30}$$

where $e_{ij} = z_{ij}$.

Finally, we adopt Ncut [15] to partition nodes into K disjoint sets of nodes, which denotes as

$$\mathrm{MC} = \left\{\mathrm{MC}_1, \cdots, \mathrm{MC}_K\right\}, \tag{31}$$

where MC_i ($i \in [1, k]$) represents the ith metacluster, which is the set of clusters. The core clusters are treated as base units in the ensemble. In each subspace, $o_i (i \in [1, \gamma])$ is partitioned into a cluster. o_i may appear multiple times in MC. Define the discriminant function δ to represent the relationship:

$$\delta = \begin{cases} 1, & \text{if } o_i \subseteq C_\beta, \\ 0, & \text{otherwise.} \end{cases} \tag{32}$$

The probability that o_i ($i \in [1, \gamma]$) is grouped into metacluster is

$$\phi\left(o_i, \mathrm{MC}_j\right) = \frac{1}{\left|\mathrm{MC}_j\right|} \sum_{C_\beta \subseteq \mathrm{MC}_j} \delta \quad (j \in [1, n_c]), \tag{33}$$

where $\left|\mathrm{MC}_j\right|$ is the number of clusters in the jth metacluster.

Finally, o_i is assigned to the metacluster with the highest probability, namely,

$$\arg\max_{MC_j \in MC} \phi\left(o_i, \mathrm{MC}_j\right). \tag{34}$$

We summarize C^{2}SWCE_MC in Algorithm 4.

4. Experiments

In this section, we conduct experiments on eight high-dimensional datasets to compare the proposed four subspace clustering ensemble algorithms against several clustering approaches. All the experiments in this paper are conducted in MATLAB R2016a on a PC with 8 Intel 3.40 GHz processors and 8 GB of RAM.

4.1. Datasets. In our experiments, we use eight high-dimensional datasets, which including four cancer gene expression datasets and four image datasets. The 4 gene expression datasets, namely, Yeoh02v1[1], Yeoh02v2[1], Bhattacharjee2001[1], and Golub1999v1[1]. The Yeoh02v1 dataset and Yeoh02v2 dataset are the pediatric acute lymphoblastic leukemia dataset, in which Yeoh02v1 dataset contains 2 categories of genes expressed in leukemic blasts, and Yeoh02v2 dataset contains 6 categories of genes expressed in leukemic blasts. Bhattacharjee2001 is the lung tumor samples dataset, which contains 186 lung tumor samples and 17 normal lung tissues. Golub1999v1 is the leukemia dataset, which contains acute myeloid leukemia samples and acute lymphoblastic leukemia samples. The other 4 image datasets, including COIL_20[3], USPS[2], Semeion[4] and Multi Featureples[4]. The COIL_20 dataset is an image dataset containing 20 item objects, and the other 3 image datasets are all handwritten digit datasets. The USPS dataset contains a total of 10 categories and 11,000 samples. For facilitating comparison, we randomly selected 10% of the samples from each category of the USPS dataset to form a dataset containing 1,100 samples, which is represented as USPS_10P.

To simplify the description, the 8 datasets will be abbreviated as D1 to D8, respectively. The details of the datasets are given in Table 2.

(1) https://schlieplab.org/Supplements/CompCancer/

(2) http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html

(3) https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

(4) http://archive.ics.uci.edu/ml/index.php

4.2. Evaluation Criterion. We use the normalized mutual information (NMI) and the clustering accuracy (ACC) to

Input: $S_1, S_2, \cdots, S_k, K$
Output: $\mathscr{P}^*$
1. Performed clustering on each base subspace to generate a set of subspace clustering ensemble $\mathscr{P}$
2. Generate the adaptive core clusters according to Definition 1
3. Calculate the similarity matrix of clusters in $\mathscr{C}$ according to Equation (23) and (24)
4. Build the graph G_M according to Equations (29)–(31)
5. Partition the graph nodes into K disjoint groups by Tcut, generating MC
6. The core clusters are treated as the base unit, the metaclusters corresponding to the core clusters are determined according to Equations (32), (34), and (35)
7. Map the labels of core cluster to the samples

Algorithm 4

Table 2: Description of the datasets.

Dataset	Abbreviation	Sample	Dimension	Class
Yeoh02v1	D1	248	2526	2
Yeoh02v2	D2	248	2526	6
Bhattacharje2001	D3	203	1543	5
Golub99v1	D4	72	1868	2
USPS_10P	D5	1100	256	10
COIL_20	D6	1440	1024	20
Semeion	D7	1593	256	10
Multiple features	D8	2000	649	10

evaluate the quality of clustering results. The NMI metrics measure the accuracy of the clustering results according to the shared information of the ground-truth clustering solution and the test clustering solution. Let P denote the clustering solution of the proposed method, and P^G denote the groundtruth clustering solution. The NMI score of P with respect to P^G is calculated as

$$\mathrm{NMI}\left(P, P^G\right) = \frac{\sum_{i=1}^{n^P}\sum_{j=1}^{n^G} n_{ij} \log\left(n_{ij} n / n_i^P n_j^G\right)}{\sqrt{\sum_{i=1}^{n^P} n_i^P \log\left(n_i^P/n\right) \sum_{j=1}^{n^G} n_j^G \log\left(n_j^G/n\right)}}, \tag{35}$$

where n^P and n^G are the number of clusters in P and P^G, respectively. n_i^P is the number of samples for the ith cluster in P, and n_j^G is the number of samples for the jth cluster in P^G; n is the number of input samples, and n_{ij} is the number of samples that the ith cluster in P and the jth cluster in P^G jointly contain.

The ACC measures the ratio of the number of samples that are correctly classified to the number of all samples. The ACC is an indicator that evaluates the accuracy of the clustering result, which is defined as

$$\mathrm{ACC}\left(P, P^G\right) = \frac{\sum_{i=1}^{n} \delta\left(P^G(x_i), \mathrm{map}\left(P(x_i)\right)\right)}{n}, \tag{36}$$

where $P^G(x_i)$ is the groundtruth label corresponding to x_i, and $P(x_i)$ is the test clustering labels corresponding to x_i in proposed approach. $\mathrm{map}(\cdot)$ is the relabel mapping function that aligns the test clustering label with the groundtruth label. $\delta(\cdot)$ is an indicator function, which is defined as

$$\delta(x, y) = \begin{cases} 1, & \text{if } x = y, \\ 0, & \text{otherwise.} \end{cases} \tag{37}$$

The value interval for ACC and NMI score is $[0, 1]$, and a higher ACC score or NMI score indicates a better clustering effect.

4.3. Discussion of the Parameter Selection. There are several parameters in proposed C^{2}SWCE, where k is the number of subspaces, r is the random feature selection ratio, $\tilde{r}$ is the unsupervised feature selection ratio, and K is the number of clusters. To increase the diversity of ensemble members, each subspace randomly generates a different number of clusters within range $[2, \sqrt{n}]$. We set the random feature selection ratio to $r = 0.5$, and the number of generated random subspaces is in the range of $[5, 25]$, with the increment of 5. For each parameter setting, we run the spectral clustering methods 10 times in each random subspace, generating core clusters according to Definition 1 at each time. The average number of generated core clusters is shown in Figure 2.

(1) The relationship between k and γ

In the proposed C^{2}SWCE framework, the core cluster is the basic unit of integration.

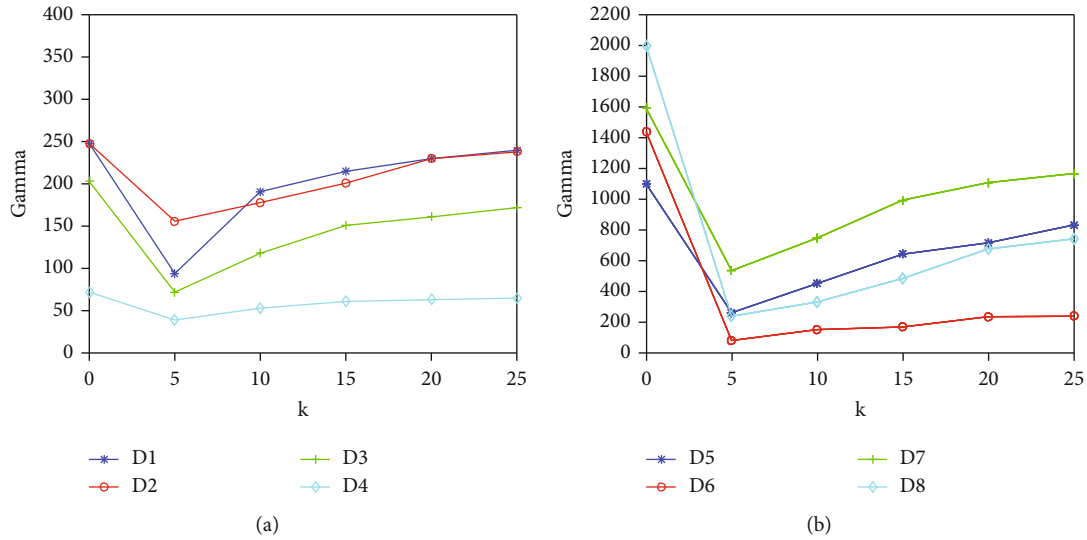


Figure 2: The correspondence between k and γ ($r = 0.5$).

In Figure 2, k is the number of generated random subspaces, $k = 0$ corresponds to the number of input samples, and s_{ij} is the number of core clusters. It is observed from Figure 2 that in the corresponding dataset, as the number of subspaces increases, so does the number of generated core clusters. For the image datasets (Figure 2(b)), even if the number of subspaces is set to $k = 25$, the number of core clusters is still much smaller than the number of input samples, and this trend is most pronounced on the D6 and D8 datasets. For the gene express datasets (Figure 2(a)), due to the small number of input samples, when the number of subspaces is set to $k = 25$, the number of generated core clusters is close to the number of input samples.

(2) The influence of parameters k and $\tilde{r}$ on clustering accuracy

In this section, we will compare the effect of the number of a random subspace k and the unsupervised feature selection ratio $\tilde{r}$ on the accuracy of clustering ensemble. We set the value interval of parameter k to $[5, 25]$ and the increment to 5. For each random subspace, the unsupervised feature selection ratio $\tilde{r}$ is set in range of $[0.3, 0.8]$, with increments set to 0.1. For each parameter setting of k and $\tilde{r}$, we run the proposed methods 10 times, and the average of the NMI scores is shown in Figure 3.

As shown in Figure 3, for the image datasets (D5-D7), parameters k and $\tilde{r}$ are insensitive to clustering results. When $\tilde{r}$ is fixed, the number of subspaces has little effect on the consensus clustering solutions. When k is fixed, the NMI score is relatively high in the range of $[0.6, 0.8]$. For the D1-D3 datasets, as shown in Figures 3(a)–3(c), the higher NMI scores are distributed in parameter k in an interval of $[10, 20]$ and parameter $\tilde{r}$ in an interval of $[0.5, 0.8]$ region. For the D4 dataset, it can be clearly observed that when $\tilde{r} = 0.8$, the corresponding NMI score is higher.

For each dataset, the more base subspaces is divided, the more adaptive core clusters are generated, which adds computational complexity and algorithm runtime. As observed from Figure 2 that the number of core clusters generated by parameter k in range $[10, 20]$ is moderate. At the same time, when $r = 0.5$ and $\tilde{r}$ is set in the range of 0.5, 0.8, the ensemble result has a higher NMI score. Therefore, in subsequent experiments, we set $k = 10$, $r = 0.8$, and $\tilde{r} = 0.8$ in experiments on all datasets.

4.4. Compare the Effects of Feature Selection and Weighted Strategies on Ensemble Results. In this section, we first compare the impact of the hybrid feature selection strategies on consensus clustering solutions. For each dataset, the proposed four ensemble methods are run on random subspaces and base subspaces generated by mixed feature selection strategies and compare their clustering results. For fair comparison, each ensemble approach runs 10 times in the unweighted manner, recording the average of its NMI scores. The comparison results are recorded in Table 3, with the notation "N" corresponding to the clustering result on random subspace, and the notation "Y" corresponding to the result of clustering result on the base space.

As shown in Table 3, the clustering effect of the proposed clustering ensemble approaches on the base subspace is differently better than that in the random subspace. Therefore, in subsequent experiments, we use the hybrid feature selection strategy, perform unsupervised feature selection on random subspace, generate a set of base subspaces, and then generate ensemble members by performing the clustering on the base subspace.

The proposed four ensemble approaches treat the core cluster as the base unit. The base subspace clustering solutions are weighted by calculating the distance between the core clusters pairs contained in the cluster, or the similarity of core clusters to each cluster. To verify the effect of the weighted clustering integration method, we compared the results of the weighted integration method and the unweighted integration method on the base subspace, respectively. Each ensemble method is run 10 times, and the average NMI of the proposed methods is recorded in Table 4. In Table 4, the notation "N" corresponds to the unweighted clustering results of the

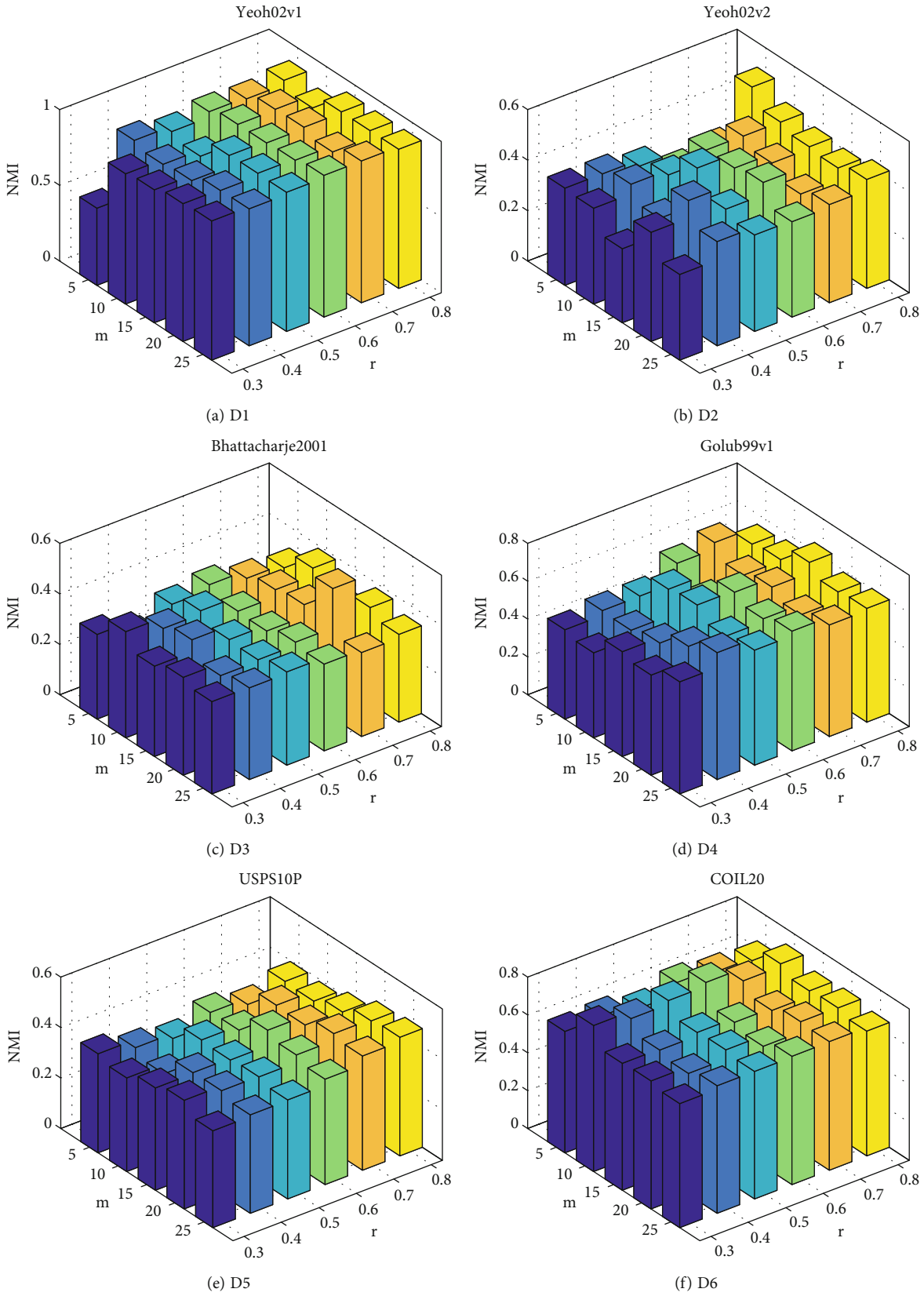
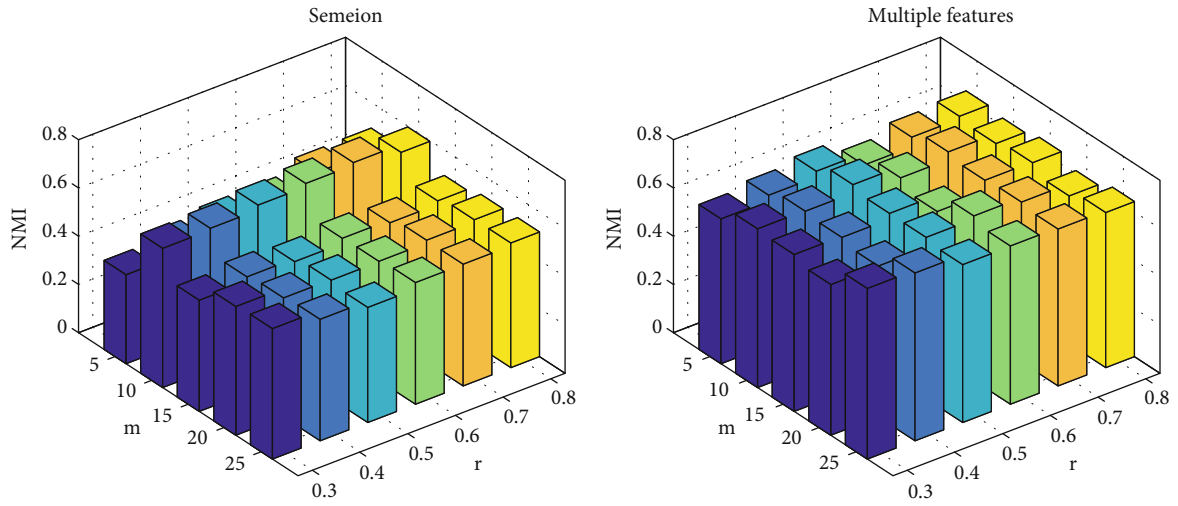


Figure 3: Continued.



(g) D7 (h) D8

Figure 3: The parameters k and $\tilde{r}$ correspond to clustering results (NMI).

Table 3: Comparison of clustering results for random subspaces with clustering results for base subspaces (NMI).

Datasets	C^2SWCE_HC		C^2SWCE_SC		C^2SWCE_BG		C^2SWCE_MC	
	N	Y	N	Y	N	Y	N	Y
D1	0.6625	0.8270	0.6381	0.8449	0.7823	0.8436	0.8041	0.8808
D2	0.3322	0.4081	0.3524	0.3845	0.4312	0.4899	0.5202	0.5302
D3	0.4179	0.5209	0.2324	0.3533	0.3502	0.4598	0.4190	0.4948
D4	0.7593	0.8257	0.8203	0.8955	0.8203	0.8955	0.9477	0.8203
D5	0.4377	0.6152	0.5133	0.5815	0.4358	0.6118	0.4374	0.6123
D6	0.8177	0.9113	0.7621	0.7818	0.8082	0.9006	0.8083	0.9104
D7	0.4742	0.6596	0.5331	0.6412	0.5318	0.6704	0.4913	0.6402
D8	0.7417	0.8881	0.7515	0.9324	0.7663	0.8967	0.7865	0.9015

Table 4: Comparison of weighted ensemble manner and unweighted ensemble manner (NMI).

Datasets	C^2SWCE_HC		C^2SWCE_SC		C^2SWCE_BG		C^2SWCE_MC	
	Y	N	Y	N	Y	N	Y	N
D1	0.9449	0.8270	0.8943	0.8449	0.8821	0.8436	0.8809	0.8808
D2	0.5386	0.4081	0.5666	0.3845	0.5638	0.4899	0.5211	0.5302
D3	0.5855	0.5209	0.4919	0.3533	0.5514	0.4598	0.5078	0.4948
D4	0.9101	0.8257	0.8955	0.8955	0.8955	0.8955	0.8203	0.8203
D5	0.6675	0.6152	0.5859	0.5815	0.5669	0.6118	0.5460	0.6123
D6	0.9224	0.9113	0.8780	0.7818	0.9277	0.9006	0.8956	0.9104
D7	0.7588	0.6596	0.6597	0.6412	0.6800	0.6704	0.6458	0.6402
D8	0.9099	0.8881	0.9446	0.9324	0.9363	0.8967	0.9021	0.9015

proposed methods, and the notation "Y" corresponds to the weighted clustering results of the proposed methods.

As observed from Table 4, on the other datasets except D4, the fusion results of the proposed four ensemble methods under the weighted ensemble manner are better than the ensemble results of the unweighted ensemble manner. For the D4 dataset, the NMI score of the weighted ensemble manner of the proposed C^2SWCE_HC is 0.9101, which is higher than the NMI score of 0.8257 corresponding to the unweighted ensemble manner. However, on the D4 dataset, there are no significant advantages in the clustering results obtained by the other three ensemble methods using a weighted ensemble manner. For other datasets, the consensus clustering results achieved by the weighted ensemble

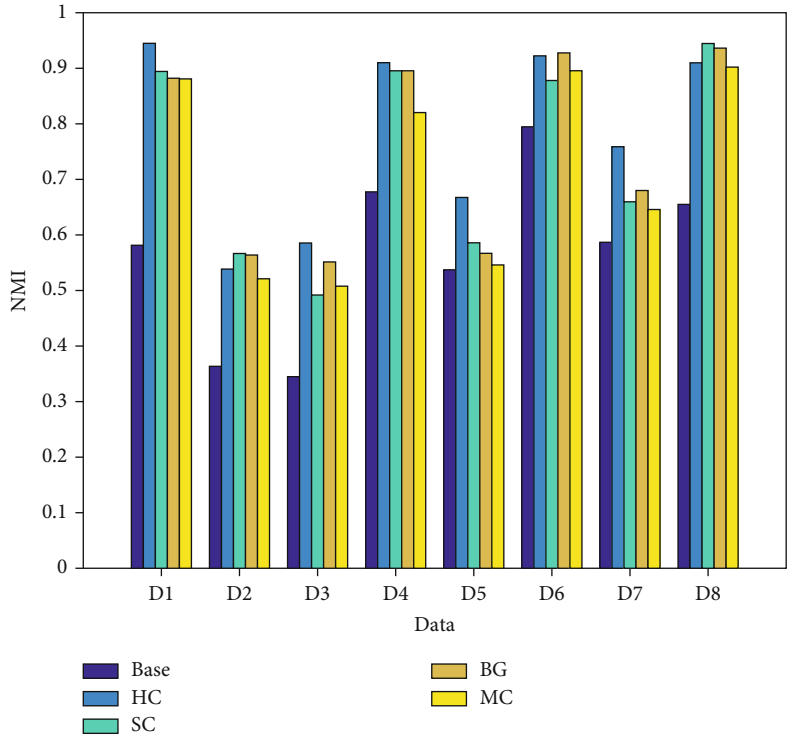


Figure 4: Performance comparison of the proposed approaches with the base clustering.

manner are better than those obtained by the unweighted ensemble manner. This illustrates the effectiveness of the proposed weighted ensemble strategies.

4.5. Comparison with Base Clustering. Subspace clustering ensemble can fuse multiple subspace clustering solutions to obtain a better accuracy, more robust consensus solution. In this section, we compare the clustering effects of the proposed locally weighted subspace clustering ensemble approaches, namely, C^{2}SWCE_HC, C^{2}SWCE_SC, C^{2}SWCE_BG, and C^{2}SWCE_MC against spectral clustering approach. Each method runs 20 times, and their average NMI scores are described in Figure 4.

In Figure 4, "base" corresponds to base clustering results, and "HC," "SC," "BG," and "MC" correspond to the clustering results of C^{2}SWCE_HC, C^{2}SWCE_SC, C^{2}SWCE_BG, and C^{2}SWCE_MC, respectively. As can be seen from Figure 4, on each experimental dataset, the proposed approaches can achieve better and more robust consensus clustering results than the spectral clustering approach. In particular, on the D1, D2, D3, D6, D7, and D8 datasets, the performance of the proposed 4 clustering integration approaches are significantly better than that of the spectral clustering approach.

4.6. Comparison with Other Clustering Methods. In this section, we compare the proposed consensus clustering approaches with the state-of-the-art clustering approaches to evaluate their effectiveness. Among comparative clustering approaches, K-means and GNMF [16] are classic clustering methods and SSC [17] and LSC [2] are the clustering methods based on spectral clustering. In the contrasting ensemble methods, MDEC_SC [4], MDEC_HC [4], and MDEC_BG [4] all use spectral clustering to generate the clustering ensemble members, and then, they use spectral clustering, hierarchical clustering, and bipartite graph partition to fuse base clustering solutions, respectively. SC_SRGF [5] adopts spectral clustering to obtain clustering solutions for each subspace; however, ECPCS-HC [18], ECPCS-MC [18], WEAC-AL [19], GP-MGLA [19], LWGP [20], and LWEA [20] all use K-means to generate base clustering solutions. All parameters of the comparison approaches are set as suggested by the corresponding papers.

For a fair comparison, we run each clustering approach 20 times on each dataset and the average performance in terms of NMI and ACC are summarized in Tables 5 and 6, respectively. In Tables 5 and 6, the highest score is highlighted in bold; the symbol "_" indicating that the algorithm cannot be performed.

The landmark-based spectral clustering (LSC) [2] approach is not suitable for datasets with number of features greater than the number of samples, so there are no corresponding clustering results on D1-D4 datasets. As shown in Table 5, the SSC achieves the best average performance in terms of NMI on D3 dataset. For the remaining datasets, the clustering ensemble approaches yield better average performance than the traditional clustering approaches. This also confirms the previous conclusion that the clustering ensemble approaches are more suitable than the traditional clustering approaches for high-dimensional data clustering scenarios.

As can be seen from Tables 5 and 6, the proposed C^{2}SWCE_HC achieves the highest average NMI scores on

Table 5: Average performance in terms of NMI over 20 runs by different clustering methods.

	D1	D2	D3	D4	D5	D6	D7	D8
K-means	0.2481	0.2238	0.3487	0.5222	0.4575	0.7470	0.5219	0.7152
GNMF	0.5962	0.3369	0.3485	0.7223	0.5572	0.8803	0.5885	0.9216
SSC	0.2938	0.3461	**0.5916**	0.1276	0.4283	0.8640	0.6042	0.8640
LSC	_	_	_	_	0.6646	0.8852	0.6189	0.9266
MDEC_SC	0.8831	0.4938	0.4843	0.7352	**0.5922**	**0.8896**	0.6562	0.9385
SC_SRGF	0.2371	0.1931	0.3621	0.5853	0.5524	0.8870	0.5947	0.8460
C^2SWCE_SC	**0.8943**	**0.5666**	**0.4919**	**0.8955**	0.5859	0.8780	**0.6597**	**0.9446**
ECPCS-HC	0.0044	0.2483	0.5141	0.5264	0.5192	0.7751	0.6329	0.7976
MDEC_HC	0.8855	0.4661	0.5284	0.7853	0.6501	0.9177	0.7072	0.8822
WEAC_AL	0.2994	0.3764	0.2886	0.2240	0.4962	0.7921	0.5959	0.7009
LWEA	0.9447	0.2511	0.3533	0.5974	0.5081	0.7636	0.5999	0.7953
C^2SWCE_HC	**0.9449**	**0.5386**	**0.5855**	**0.9101**	**0.6675**	**0.9224**	**0.7588**	**0.9099**
MDEC_BG	**0.8863**	0.5584	0.5079	0.6923	**0.5882**	0.9162	0.6775	0.9331
GP_MGLA	0.2794	0.3542	0.2847	0.2194	0.4808	0.7974	0.5315	0.6814
LWGP	0.8620	0.3152	0.3630	0.5974	0.5246	0.7924	0.6193	0.8736
C^2SWCE_BG	0.8821	**0.5638**	**0.5514**	**0.8955**	0.5669	**0.9277**	**0.6800**	**0.9363**
ECPCS-MC	0.0816	0.3178	0.4256	0.5660	0.5451	0.8048	**0.6600**	0.8580
C^2SWCE_MC	**0.8809**	**0.5211**	**0.5078**	**0.8203**	**0.5460**	**0.8956**	0.6458	**0.9021**

Table 6: Average performance in terms of ACC over 20 runs by different clustering methods.

	D1	D2	D3	D4	D5	D6	D7	D8
K-means	0.8182	0.4286	0.5655	0.8597	0.4225	0.5944	0.0411	0.7334
GNMF	0.9395	0.5524	0.7783	0.9444	0.5931	0.8118	0.6390	0.9625
SSC	0.8508	0.5565	0.8177	0.6528	0.4418	0.7821	0.6109	0.8200
LSC	_	_	_	_	0.6505	0.7972	0.6309	0.9675
MDEC_SC	**0.9899**	0.5395	0.7320	0.9583	0.5031	0.7698	0.0210	**0.9713**
SC_SRGF	0.7766	0.4097	0.5448	0.9458	0.4911	0.7990	**0.0657**	0.8465
C^2SWCE_SC	**0.9899**	**0.6109**	**0.7433**	**0.9861**	**0.5423**	**0.8583**	0.0135	0.9793
ECPCS-HC	0.8161	0.4762	0.7695	0.8944	0.4692	0.6307	**0.0485**	0.8244
MDEC_HC	0.9899	0.5548	**0.7744**	0.9611	**0.6229**	0.8543	0.0157	0.8539
WEAC_AL	0.6774	0.5000	0.3693	0.2639	0.4509	0.7063	0.0232	0.6295
LWEA	**0.9960**	0.4194	0.5419	0.9167	0.5218	0.6076	0.0308	0.7820
C^2SWCE_HC	0.9859	**0.6109**	0.6133	**0.9861**	0.5423	**0.8583**	0.0135	**0.9693**
MDEC_BG	**0.9899**	0.6129	**0.7488**	0.9417	0.4951	**0.8460**	0.0239	0.9682
GP_MGLA	0.7097	0.4960	0.3596	0.2639	0.4364	0.7354	0.0113	0.5565
LWGP	0.9879	0.4879	0.4975	0.9167	0.5255	0.6764	**0.0797**	0.9360
C^2SWCE_BC	**0.9899**	**0.6532**	0.6552	**0.9861**	**0.5341**	0.8358	0.0135	**0.9788**
ECPCS-MC	0.8363	0.5258	0.6384	0.9069	**0.5117**	0.6729	0.0207	**0.9260**
C^2SWCE_MC	**0.9899**	**0.6210**	**0.7414**	**0.9722**	0.4696	**0.8333**	**0.0546**	0.9115

all datasets among the comparative hierarchical clustering ensemble approaches. Especially on the D1-D4 datasets, the performance of the C^2SWCE_HC is significantly better than that of WEAC_AL and ECPCS-HC, and it also achieves the highest average ACC score on the D2, D4, D6, and D8 datasets. Comparing the ensemble approaches based on spectral clustering, the average performance in terms of NMI of C^2SWCE_SC is significantly higher than that of other methods on the D1-D4 dataset. For example, on the D1 and D2 datasets, the average NMI scores of SC_SRGF are 0.2371 and 0.1931, respectively, while the average NMI scores of C^2SWCE_SC are 0.8943 and 0.5666, respectively, which increased by about 4 times. In addition, compared with the other bipartite graph partitioning ensemble methods, the proposed C^2SWCE_BG achieves the highest average NMI score on 6 datasets and the highest average

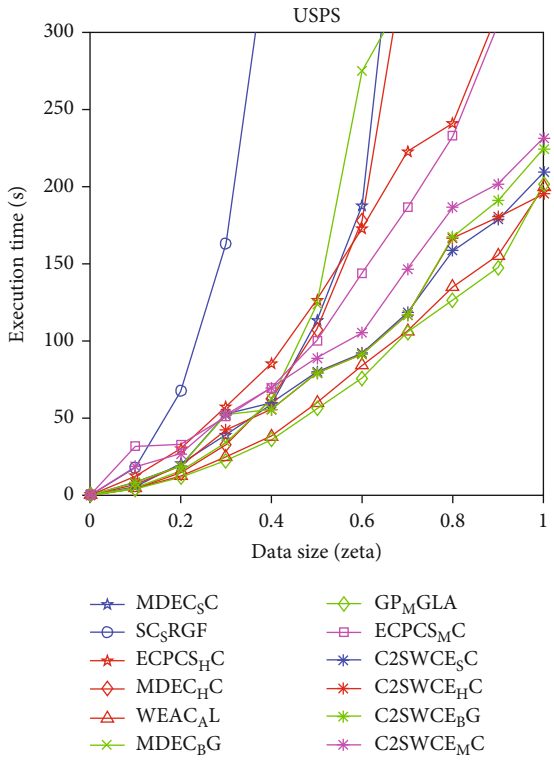


Figure 5: Execution time of clustering ensemble approaches with different data sizes.

ACC score on 5 datasets. Its performance is significantly better than MDEC_BG, GP_MGLA, and LWGP. In particular, on the D1, D3, and D4 datasets, the average NMI scores of the C^{2}SWCE_BG are 0.8821, 0.5514, and 0.8955, respectively, significantly exceeding the corresponding NMI scores of 0.2794, 0.2847, and 0.2194 for the GP_MGLA. Finally, the metacluster-based ensemble clustering proposed in this paper is significantly better than that of the ECPCS-MC method on all datasets except D7.

4.7. Execution Time. In this section, we compare the execution times of different clustering ensemble approaches at the integration phase on the USPS dataset. The USPS dataset contains 10 categories with a total of 11,000 samples. To compare the clustering effects of different ensemble approaches on datasets of different sizes, first, according to the method described in Section 4.1, samples of different proportions are randomly selected for each category of the input dataset, to generate the USPS dataset of different sizes. In the experiment, the sample sampling ratio ς is set in interval of $[0, 1]$, and the increment set to 0.1. That is, the generated dataset contains $n\varsigma$ samples, where n is the number of samples in the USPS dataset. Running various ensemble approaches on the USPS datasets of different sizes, Figure 5 compares their execution time during the integration phase.

As can be seen from Figure 5, the execution time of all ensemble approaches increases significantly as the size of the USPS dataset increases, with SC_SRGF and MDEC_SC approaches increasing the fastest. When clustering all samples of the USPS dataset, the proposed C^{2}SWCE_HC, C^{2}SWCE_SC, C^{2}SWCE_BG, and C^{2}SWCE_MC approaches are 195.23 s, 209.15 s, 220.72 s, and 231.65 s, respectively. Compared with other ensemble methods, the proposed methods have obvious advantages, especially the computational efficiency of C^{2}SWCE_SC is higher than that of the contrasting spectral clustering ensemble methods MDEC_SC and SC_SRGF. The results show that the spectral clustering method based on the core clusters can reduce the complexity of the spectral clustering method to a certain extent. Further, compared with other hierarchical clustering ensemble methods, the execution time of C^{2}SWCE_HC is comparable to that of the WEAC_AL, but much faster than that of ECPCS-HC and MDEC_HC. At the same time, we also observed that the execution time of the C^{2}SWCE_BG is slightly higher than that of the GP_MGLA, but it has a significant advantage over MDEC_BG.

In summary, the proposed approaches have a modest computational cost for ensemble tasks on USPS datasets of different sizes. Compared with the same type of ensemble approaches, the execution time of the proposed 4 approaches have advantages.

5. Conclusion

In this paper, we propose a novel locally weighted subspace clustering ensemble framework, termed C^{2}SWCE. It first uses the hybrid feature selection strategy that combines random feature selection and unsupervised feature selection to generate a set of base subspaces. The strategy combines the diversity of random feature selection to select representative features from each random subspace in an unsupervised manner, continuously reducing the dimensionality of the random subspace. To increase the diversity of subspace ensemble members, in addition to using random feature selection to generate feature subspaces, we also randomly group the samples into different numbers of clusters. Furthermore, we introduce concept of the core cluster. In the ensemble process, the core cluster is viewed as the base unit, which improves the ensemble efficiency to a certain extent. The subspace clustering solution is weighted by evaluating the stability of the cluster. Last but not least, under the proposed framework, four weighted ensemble approaches are proposed to integrate the clustering solutions of the base subspace to achieve the final clustering result. Extensive experiments are conducted on 8 real-world datasets to verify the effectiveness of the proposed ensemble approaches. Experimental results show that compared with the state-of-the-art ensemble methods, our methods have stronger robustness, and the comprehensive performance of clustering accuracy and efficiency has advantages.

Data Availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest

The author states that this article has no conflict of interest.

References

[1] E. Elhamifar and R. Vidal, "Sparse subspace clustering: algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.

[2] X. Chen and D. Cai, "Large scale spectral clustering with landmark-based representation," *In Proceedings of AAAI Conference on Artificial Intelligence*, vol. 25, no. 1, pp. 313–318, 2011.

[3] W. Y. Chen, Y. Song, H. Bai, C. J. Lin, and E. Y. Chang, "Parallel spectral clustering in distributed systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 568–586, 2011.

[4] D. Huang, C.-D. Wang, J.-H. Lai, and C.-K. Kwoh, "Toward multidiversified ensemble clustering of high-dimensional data: from subspaces to metrics and beyond," *IEEE Transactions on Cybernetics*, pp. 1–14, 2021.

[5] X. Cai, D. Huang, C.-D. Wang, and C.-K. Kwoh, "Spectral clustering by subspace randomization and graph fusion for high-dimensional data," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 330–342, Springer, Cham, 2020.

[6] K. Shankar, "Fuzzy clustering and classification based iris recognition: a medical application," *American Journal of Business and Operations Research*, vol. 1, no. 1, pp. 19–27, 2020.

[7] S. Verma and S. Gain, "Mitigating hot spot problem in wireless sensor networks using political optimizer based unequal clustering technique," *Journal of Cybersecurity and Information Management*, vol. 8, no. 2, pp. 42–50, 2021.

[8] H.-P. Kriegel, P. Kröger, and A. Zimek, "Clustering high-dimensional data," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 1, pp. 1–58, 2009.

[9] 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.

[10] L. Jing, K. Tian, and J. Z. Huang, "Stratified feature sampling method for ensemble clustering of high dimensional data," *Pattern Recognition*, vol. 48, no. 11, pp. 3688–3702, 2015.

[11] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," *In Proceedings of International Conference on Neural Information Processing Systems*, pp. 507–514, 2005.

[12] D. Huang, J. Lai, and C.-D. Wang, "Ensemble clustering using factor graph," *Pattern Recognition*, vol. 50, pp. 131–142, 2016.

[13] C. Domeniconi and M. S. Al-Razgan, "Weighted cluster ensembles," *ACM Transactions on Knowledge Discovery from Data*, vol. 2, no. 4, pp. 1–40, 2009.

[14] Z. Li, X.-M. Wu, and S.-F. Chang, "Segmentation using superpixels: a bipartite graph partitioning approach," in *Proceedings of IEEE Conference Computer Vision Pattern Recognition*, pp. 789–796, Providence, RI, USA, June 2012.

[15] J. Jianbo 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.

[16] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[17] S. Matsushima and M. Brbic, "Selective sampling-based scalable sparse subspace clustering," *In Proceedings of Advances in Neural Information Processing Systems*, pp. 12416–12425, 2019.

[18] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 1, pp. 508–520, 2021.

[19] D. Huang, J. H. Lai, and C. D. Wang, "Combining multiple clusterings via crowd agreement estimation and multi- granularity link analysis," *Neurocomputing*, vol. 170, pp. 240–250, 2015.

[20] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE Transactions On Cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2018.