Incremental Semi-Supervised Clustering Ensemble for High Dimensional Data Clustering

Zhiwen Yu, *Senior Member, IEEE*, Peinan Luo, Jane You, Hau-San Wong, Hareton Leung, Si Wu, Jun Zhang, *Senior Member, IEEE*, and Guoqiang Han

Abstract—Traditional cluster ensemble approaches have three limitations: (1) They do not make use of prior knowledge of the datasets given by experts. (2) Most of the conventional cluster ensemble methods cannot obtain satisfactory results when handling high dimensional data. (3) All the ensemble members are considered, even the ones without positive contributions. In order to address the limitations of conventional cluster ensemble approaches, we first propose an incremental semi-supervised clustering ensemble framework (ISSCE) which makes use of the advantage of the random subspace technique, the constraint propagation approach, the proposed incremental ensemble member selection process, and the normalized cut algorithm to perform high dimensional data clustering. The random subspace technique is effective for handling high dimensional data, while the constraint propagation approach is useful for incorporating prior knowledge. The incremental ensemble member selection process is newly designed to judiciously remove redundant ensemble members based on a newly proposed local cost function and a global cost function, and the normalized cut algorithm is adopted to serve as the consensus function for providing more stable, robust, and accurate results. Then, a measure is proposed to quantify the similarity between two sets of attributes, and is used for computing the local cost function in ISSCE. Next, we analyze the time complexity of ISSCE theoretically. Finally, a set of nonparametric tests are adopted to compare multiple semi-supervised clustering ensemble approaches over different datasets. The experiments on 18 real-world datasets, which include six UCI datasets and 12 cancer gene expression profiles, confirm that ISSCE works well on datasets with very high dimensionality, and outperforms the state-of-the-art semi-supervised clustering ensemble approaches.

Index Terms—Cluster ensemble, semi-supervised clustering, random subspace, cancer gene expression profile, clustering analysis

1 Introduction

Recently, cluster ensemble approaches are gaining more and more attention [1], [2], [3], [4], due to its useful applications in the areas of pattern recognition [2], [3], [4], [5], data mining [6], [7], bioinformatics [8], [9], [10], and so on. When compared with traditional single clustering algorithms, cluster ensemble approaches are able to integrate multiple clustering solutions obtained from different data sources into a unified solution, and provide a more robust, stable and accurate final result.

However, conventional cluster ensemble approaches have several limitations: (1) They do not consider how to make use of prior knowledge given by experts, which is represented by pairwise constraints. Pairwise constraints are often defined as the must-link constraints and the cannot-link constraints. The must-link constraint means that two feature vectors should be assigned to the same cluster, while the cannot-link constraints means that two feature vectors cannot be assigned to the same cluster. (2) Most of the cluster ensemble methods cannot achieve satisfactory results on high dimensional datasets. (3) Not all the ensemble members contribute to the final result. In order to address the first and second limitations, we first propose the random subspace based semi-supervised clustering ensemble framework (RSSCE), which integrates the random subspace technique [11], the constraint propagation approach [12], and the normalized cut algorithm [13] into the cluster ensemble framework to perform high dimensional data clustering. Then, the incremental semi-supervised clustering ensemble framework (ISSCE) is designed to remove the redundant ensemble members. When compared with traditional semi-supervised clustering algorithm, ISSCE is characterized by the incremental ensemble member selection (IEMS) process based on a newly proposed global objective function and a local objective function, which selects ensemble members progressively. The local objective function is calculated based on a newly designed similarity function which determines how similar two sets of attributes are in the subspaces. Next, the computational cost and the space consumption of ISSCE are analyzed theoretically. Finally, we adopt a number of nonparametric tests to compare multiple semi-supervised clustering ensemble approaches over different datasets. The experiment results show the improvement of ISSCE over traditional semi-supervised clustering ensemble approaches or conventional cluster ensemble methods on six real-world datasets from UCI machine learning repository [14] and 12 real-world datasets of cancer gene expression profiles.

- *Z. Yu, P. Luo, S. Wu, and G. Han are with the School of Computer Science and Engineering, South China University of Technology, China. E-mail: {zhwyu, csgqhan}@scut.edu.cn, hareton.leung@comp.polyu.edu.hk, 411139073@qq.com, ez.wusi@gmail.com.*
- *J. You and H. Leung are with the Department of Computing, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR. E-mail: csyjia@comp.polyu.edu.hk.*
- *H.-S. Wong is with the Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR. E-mail: cshswong@cityu.edu.hk.*
- *J. Zhang is with the School of Advanced Computing, Sun Yat-Sen University, Guangzhou, 510275, P. R. China. E-mail: junzhang@ieee.org.*

Manuscript received 24 May 2015; revised 26 Oct. 2015; accepted 30 Oct. 2015. Date of publication 10 Nov. 2015; date of current version 2 Feb. 2016. Recommended for acceptance by J. Gama.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2015.2499200

The contributions of the paper is fourfold. First, we propose an incremental ensemble framework for semi-supervised clustering in high dimensional feature spaces. Second, a local cost function and a global cost function are proposed to incrementally select the ensemble members. Third, the newly designed similarity function is adopted to measure the extent to which two sets of attributes are similar in the subspaces. Fourth, we use non-parametric tests to compare multiple semi-supervised clustering ensemble approaches over different datasets.

The remainder of the paper is organized as follows. Section 2 describes previous work related to semi-supervised clustering and cluster ensemble. Section 3 presents the incremental semi-supervised clustering ensemble framework. Section 4 analyzes the proposed algorithm theoretically. Section 5 experimentally evaluates the performance of our proposed approach. Section 6 describes our conclusion and future work.

2 Related Work

Cluster ensemble, also referred to as consensus clustering, is one of the important research directions in the area of ensemble learning, which can be divided into two stages: the first stage aims at generating a set of diverse ensemble members, while the objective of the second stage is to select a suitable consensus function to summarize the ensemble members and search for an optimal unified clustering solution. To attain these objectives, Strehl and Ghosh [1] first proposed a knowledge reuse framework which integrates multiple clustering solutions into a unified one. After that, a number of researchers followed up Strehl's work, and proposed different kinds of cluster ensemble approaches [15], [16], [17], [18], [19], [20], [21]. While there are different kinds of cluster ensemble techniques, few of them consider how to handle high dimensional data clustering, and how to make use of prior knowledge. High dimensional datasets have too many attributes relative to the number of samples, which will lead to the overfitting problem. Most of the conventional cluster ensemble methods do not take into account how to handle the overfitting problem, and cannot obtain satisfactory results when handling high dimensional data. Our method adopts the random subspace technique to generate the new datasets in a low dimensional space, which will alleviate this problem.

There are also other research works which study the properties of the cluster ensemble theoretically, such as the stability of k-means based cluster ensemble [2], the efficiency of the cluster ensemble [22], the convergence property of consensus clustering [23], the scalability property of the cluster ensemble [24], the effectiveness of cluster ensemble methods [25], and so on. Cluster ensemble approaches have been applied to different areas, such as bioinformatics [26], [27], image segmentation [28], language processing [29], Internet security [30], and so on.

Recently, some researchers realized that not all the ensemble members contribute to the final result, and investigate how to select a suitable subset of members to obtain better results [31], [32], [33], [34], [35]. For example, Yu et al. [33], [34] treated the ensemble members as features, and explored how to use suitable feature selection techniques to choose the ensemble members. In summary, most of the cluster ensemble approaches only consider using a similarity score or feature selection technique to remove the redundant ensemble members, and few of them study how to apply an optimization method to search for a suitable subset of ensemble members. In the current work, the proposed ISSCE framework uses a newly designed incremental ensemble member selection process to generate an optimal set of members.

In addition, conventional cluster ensemble methods do not take into account how to make use of prior knowledge, which is usually represented in the form of pairwise constraints or a very small set of labeled data. Single semi-supervised clustering algorithms have the ability to handle prior knowledge, and use them to guide the search in the process of clustering. A number of semi-supervised clustering algorithms have been proposed [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], such as semi-supervised maximum margin clustering [36], semi-supervised kernel mean shift clustering [37], semi-supervised linear discriminant clustering [38], semi-supervised hierarchical clustering [39], active learning based semi-supervised clustering [40], semi-supervised affinity propagation [41], semi-supervised non-negative matrix factorization [42], and so on. It is natural to adopt a suitable single semi-supervised clustering method as the basic clustering algorithm in the cluster ensemble. In this paper, we consider the constraint propagation approach (E^2CP) proposed in [12], which propagate constraints in a more exhaustive and efficient way, as the basic clustering algorithm in ISSCE. This approach has two advantages: (1) The time complexity of E^2CP is proportional to the total number of all possible pairwise constraints, which is $O(Kn^2)$ (where K is the number of neighbors in the K-NN graph, and n is the number of feature vectors in the dataset). It is much smaller than that of conventional constraint clustering approaches, which is $O(n^4)$. (2) E^2CP achieves good results on different real-world datasets, such as image datasets, UCI datasets, cross-modal multimedia retrieval, and so on. Greene and Cunningham [55] studied constraint selection by identifying the constraints which are useful for improving the accuracy of the clustering solution. When compared with their work, our proposed incremental semi-supervised clustering ensemble framework adopts the more effective constraint propagation approach to convey supervised information from the labeled data samples to the unlabeled samples, and solve the label propagation problem in parallel.

3 Incremental Semi-supervised Clustering Ensemble Framework

We focus on semi-supervised clustering ensemble approaches, which have been successfully applied to different areas, such as data mining [46], bioinformatics [47], [48], and so on. For example, Wang et al. [46] proposed the constraint neighborhood projection based semi-supervised clustering ensemble approach, and achieved good performance on UCI machine learning datasets. Yu et al. represented prior knowledge provided by experts as pairwise constraints, and proposed the knowledge based cluster ensemble method [47] and the double selection based

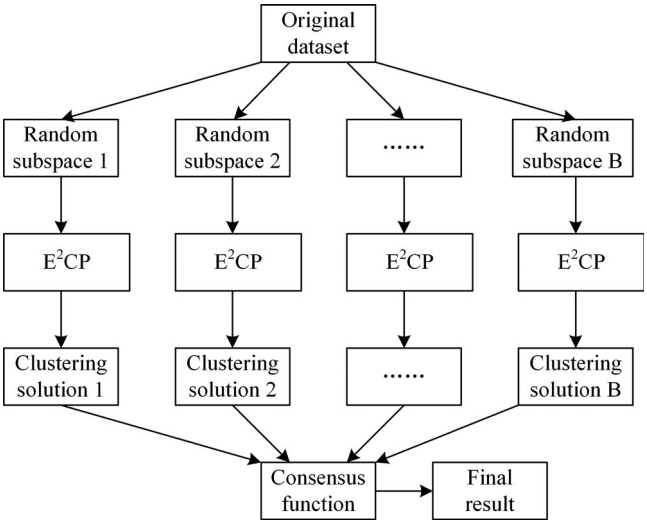


Fig. 1. An overview of the random subspace based semi-supervised clustering ensemble approach.

semi-supervised clustering ensemble approach [48]. Both of them are successfully used for clustering gene expression data. However, few of them consider how to handle high dimensional datasets. In order to address this limitation, we first propose the random subspace based semi-supervised clustering ensemble approach, as shown in Fig. 1. RSSCE first adopts the random subspace technique to generate B random subspaces A_1, A_2, ..., A_B. Then, the constraint propagation approach (E^2CP, [12]) is applied to perform clustering in the subspaces, and generate a set of clustering solutions I^1, I^2, ..., I^B. Next, a consensus matrix is constructed based on the set of clustering solutions. Finally, the normalized cut algorithm (Ncut, [13]) is used to serve as the consensus function, partition the consensus matrix, and obtain the final result.

Specifically, given a very high dimensional dataset $P = \{p_1, p_2, \ldots, p_n\}$, with each feature vector p_i ($i \in \{1, \ldots, n\}$) containing m attributes, RSSCE uses the random subspace technique to generate a set of random subspaces $A = \{A_1, \ldots A_B\}$ in the first step. Specifically, a sampling rate $\tau \in [\tau_{min}, \tau_{max}]$ of the number of attributes in the subspace over the total number of attributes in the original space is first generated as follows:

$$\tau = \tau_{min} + \lfloor \varsigma_1(\tau_{max} - \tau_{min}) \rfloor, \tag{1}$$

where ς_1($\varsigma_1 \in [0, 1]$) is a uniform random variable. Then, the attribute is selected by RSSCE one by one, whose index is determined as follows:

$$j = \lfloor 1 + \varsigma_2 m \rfloor, \tag{2}$$

where j is the index of the selected attribute, and ς_2 is a uniform random variable. The above operation will continue until τm attributes are selected. The new subspace is constructed by these selected attributes. Finally, RSSCE will generate a set of random subspaces A_1, A_2, ..., A_B by repeating the process B times. The advantage of the random subspace technique is to provide multiple ways to explore the underlying structure of the data in a low dimensional space.

In the second step, the constraint propagation approach [12] is adopted to serve as the semi-supervised clustering model χ to generate a set of clustering solutions. E^2CP considers a limited number of must-link and cannot-link constraints between pairs of feature vectors given by experts, and decomposes a constraint propagation problem into a set of semi-supervised classification problems. The dataset $P = \{p_1, p_2, \ldots, p_n\}$ can be modeled by an undirected weighted graph $G(P, X)$, in which P is a set of feature vectors corresponding to the vertices, and X denotes the similarity matrix with a weight value x_{ij} associated with the edge between p_i and p_j:

$$x_{ij} = \begin{cases} \varrho_{ij} & \text{if } p_i \in \mathcal{N}_q(p_j) \text{ or } p_j \in \mathcal{N}_q(p_i) \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

$$\varrho_{ij} = \exp\left(-\frac{\left\|p_i - p_j\right\|^2}{\overline{d}^2}\right), \tag{4}$$

where $\overline{d}$ is set to the average Euclidean distance between all pairs of q-nearest neighbors, and $\mathcal{N}_q(p_i)$ denotes the q-nearest neighbor set for p_i.

Given a set of initial must-link constraints $M = \{(p_i, p_j) : y_i = y_j, 1 \le i, j \le n\}$ and a set of initial cannot-link constraints $N = \{(p_i, p_j) : y_i \ne y_j, 1 \le i, j \le n\}$ (where y_i is the label of the feature vector p_i), the constraint matrix $R = \{r_{ij}\}_{n \times n}$ can be constructed as follows:

$$R = \begin{cases} +1, & (p_i, p_j) \in M \\ -1, & (p_i, p_j) \in N \\ 0, & otherwise. \end{cases} \tag{5}$$

The goal of E^2CP is to propagate supervised information from labeled data samples to unlabeled ones, which can be solved by label propagation based on q-nearest neighbor graphs. The propagation procedure in the j-th column(or row) of F (F is an $n \times n$ propagation matrix, whose initial values are set to R) can be viewed as a semi-supervised binary classification subproblem with respect to p_j. The semi-supervised classification problem with respect to p_j in the vertical direction can be formulated as minimizing a Laplacian regularized objective function according to [54]:

$$\min_{F_{\cdot j}} \frac{1}{2}\left\|F_{\cdot j} - R_{\cdot j}\right\|_2^2 + \frac{\mu}{2} F_{\cdot j}^T L F_{\cdot j}, \tag{6}$$

where $F_{\cdot j}$ denotes the j-th column of F, and L is the Laplacian matrix of the dataset P. The first term in the above equation denotes the fitting error, which penalizes large changes between the propagated pairwise constraints and the initial ones, and the second term measures how far the points in the dataset differ from each other with respect to the similarity. By integrating all these subproblems with respect to each column, the constraints propagation problem in the vertical direction can be formulated as follows:

$$\min_{F} \frac{1}{2}\|F - R\|_2^2 + \frac{\mu}{2} F^T L F. \tag{7}$$

Only choosing the vertical direction to propagate might not be enough to fully utilize the prior knowledge, since some columns of R do not have any pairwise constraints.

Fortunately, the problem can be fixed by propagating both in the horizontal direction and vertical direction iteratively using the method in [12]. The pairwise constraint propagation with respect to p_j in the horizontal direction can be formulated as follows:

$$\min_{F_{j\cdot}} \frac{1}{2}\left\|F_{j\cdot} - R_{j\cdot}\right\|_2^2 + \frac{\mu}{2} F_{j\cdot} L F_{j\cdot}^T, \tag{8}$$

where $F_{j\cdot}$ denotes the j-th row of F. If all the horizontal propagation subproblems are merged together, the following objective function can be defined:

$$\min_{F} \frac{1}{2}\|F - R\|_2^2 + \frac{\mu}{2} F L F^T, \tag{9}$$

E^2CP takes into account the propagation process in the horizontal and vertical directions together, and optimizes the following objective function [12]:

$$\min_{F} \|F - R\|_F^2 + \frac{\beta_1}{2} tr(F^T L F + F L F^T), \tag{10}$$

where β_1 is a pre-defined parameter. The advantage of the above formulation is that we can perform the constraint propagation in the two directions at the same time. A closed-form solution of F^* can be derived, which is as follows:

$$F^* = (1-\beta_2)^2 (I - \beta_2 \bar{L})^{-1} R (I - \beta_2 \bar{L})^{-1}, \tag{11}$$

where $\beta_2 = \beta_1/(\beta_1 + 1)$ and $\bar{L} = I - L$. F^* is used to adjust the similarity matrix X of the dataset P as follows:

$$\widetilde{x}_{ij} = \begin{cases} 1 - (1 - f_{ij}^*)(1 - x_{ij}), & f_{ij}^* \geq 0 \\ (1 + f_{ij}^*) x_{ij}, & f_{ij}^* < 0, \end{cases} \tag{12}$$

E^2CP applies a spectral clustering algorithm to the adjusted similarity matrix $\widetilde{X}$ of the dataset P to obtain a clustering solution.

Finally, E^2CP will obtain a set of corresponding clustering solutions $I = \{I^1, I^2, \ldots, I^B\}$ in terms of different ensemble members $\widehat{\Gamma} = \{(A_1, \chi_1), (A_2, \chi_2), \ldots, (A_B, \chi_B)\}$.

In the third step, RSSCE generates a consensus matrix O by summarizing the clustering solutions $\{I^1, I^2, \ldots, I^B\}$ generated by the semi-supervised clustering model E^2CP. Each clustering solution I^b ($b \in \{1, \ldots, B\}$) can first be transformed to an adjacency matrix O^b with entries o_{ij}^b as follows:

$$o_{ij}^b = \begin{cases} 1 & \text{if ybi} = \text{ybj}, \\ 0 & \text{if ybi} \neq \text{ybj}, \end{cases} \tag{13}$$

where y_i^b and y_j^b denote the predicted labels of the data samples p_i and p_j in the b-th clustering solution, respectively. An $n \times n$ consensus matrix O is then constructed by combining all the adjacency matrices $(O^1, O^2, \ldots, O^B)$ as follows:

$$O = \frac{1}{B} \sum_{b=1}^{B} O^b. \tag{14}$$

In the fourth step, RSSCE adopts the normalized cut algorithm (Ncut, [13]) as the consensus function to partition the feature vector set P based on the consensus matrix O. We construct a graph ($G = (P, O)$) whose vertices denote the feature vectors, and whose edges correspond to the values of o_{ij} in O, which represent the probability that the feature vectors belong to the same cluster. Ncut partitions the graph G into two subgraphs recursively until k subgraphs are obtained. The cost function $\Omega(P_1, P_2)$ of Ncut, which is used to maximize the association within the cluster and minimize the disassociation between the clusters, is defined as follows:

$$\Omega(P_1, P_2) = \frac{\Phi(P_1, P_2)}{\Psi(P_1, P)} + \frac{\Phi(P_1, P_2)}{\Psi(P_2, P)}, \tag{15}$$

$$\Phi(P_1, P_2) = \sum_{p_i \in P_1, p_j \in P_2} o_{ij}, \tag{16}$$

$$\Psi(P_1, P) = \sum_{p_i \in P_1, p_h \in P} o_{ih}, \tag{17}$$

where $\Omega(P_1, P_2)$ is a dissimilarity measure between P_1 and P_2, and o_{ij} is the ij-th entry of O. The above cost function can be converted to an alternative form as follows:

$$\Omega(P_1, P_2) = \frac{\sum_{(v_i > 0, v_j < 0)} -o_{ij} v_i v_j}{\sum_{v_i > 0} \vartheta_i} + \frac{\sum_{(v_i < 0, v_j > 0)} -o_{ij} v_i v_j}{\sum_{v_i < 0} \vartheta_i}, \tag{18}$$

where $v = [v_1, \ldots, v_n]^T$ is an n-dimensional indicator vector (n is the number of feature vectors in P), v_i ($i \in \{1, \ldots, n\}$) takes on values in $\{-1, 1\}$, $v_i = 1$ if the i-th vertex belongs to P_1. Otherwise, $v_i = -1$, and $\vartheta_i = \sum_j o_{ij}$.

The corresponding optimization problem is defined as follows [13]:

$$\min_{v} \Omega(v) = \min_{\alpha} \frac{\alpha^T (U - O) \alpha}{\alpha^T U \alpha}, \tag{19}$$

$$\alpha = (1 + v) - \iota (1 - v), \tag{20}$$

$$\iota = \frac{\sum_{v_i > 0} \vartheta_i}{\sum_{v_i < 0} \vartheta_i}, \tag{21}$$

with the constraints:

$$\alpha_i \in \{-\iota, 1\}, \alpha^T U I = 0, \tag{22}$$

where U is an $n \times n$ diagonal matrix with ϑ_i ($i \in \{1, \ldots, n\}$) on its diagonal, I denotes the identity matrix, and η_i is the i-th component of α.

However, the above optimization problem in its exact form is NP-complete. One possible way to solve the problem is to search for an approximate solution in the real value domain. Since the above equation is in the form of a Rayleigh quotient, the above optimization problem can be solved through the following generalized eigenvalue system when α is relaxed to take on real values, which is as follows:

$$(U - O)\alpha = \Lambda U \alpha, \tag{23}$$

where Λ denotes the eigenvalues. The second smallest eigenvector of the generalized eigenvalue system is the solution to the normalized cut problem [13].

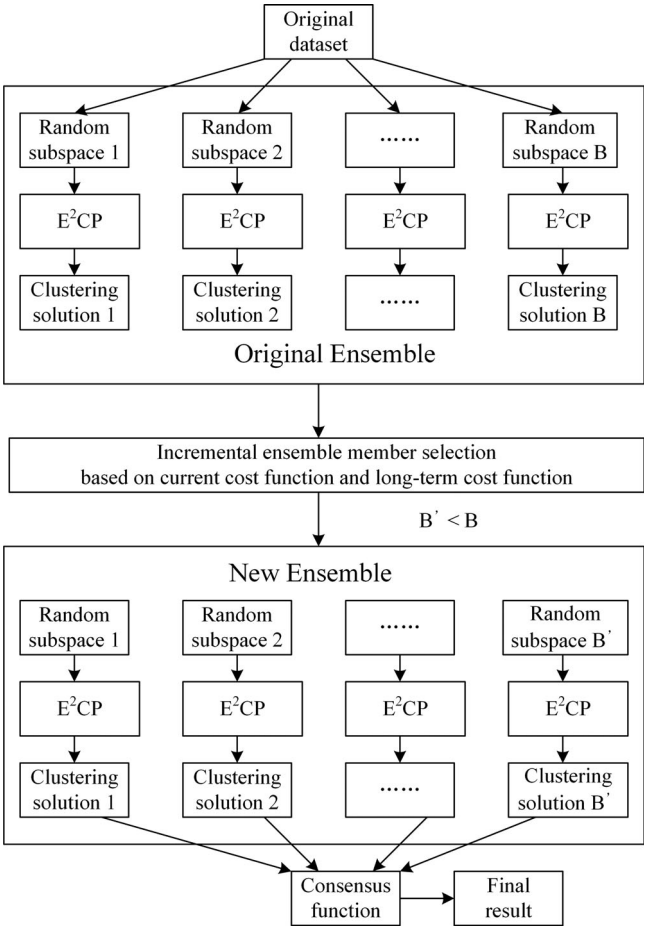


Fig. 2. An overview of the incremental ensemble member selection approach.

Fig. 2 shows an overview of the incremental semi-supervised clustering ensemble framework, and Algorithm 1 provides a flowchart of the approach. When compared with RSSCE, ISSCE adopts the incremental ensemble member selection process based on a local cost function and global cost function to generate the new ensemble set $\Gamma = \{(A_1, \chi_1), (A_2, \chi_2), \ldots, (A_{B'}, \chi_{B'})\}$ (where $B' < B$) from the original ensemble $\widehat{\Gamma}$. There are two reasons for our proposed incremental process: (1) The global objective function is useful for performing global search, while the local objective function is suitable for local search. Our proposed incremental process will improve the efficiency by using both search modes. (2) The local objective function takes into account the similarity of two subspaces and the cost function of the clustering solutions, which will increase the adaptivity of the process.

4 Incremental Ensemble Member Selection

Algorithm 2 provides an overview of the incremental ensemble member selection process. The input is the original ensemble, while the output is a newly generated ensemble with smaller size. Specifically, IEMS considers the ensemble members one by one, and calculates the objective function $\Delta(I^b)$ for each clustering solution I^b generated by E^2CP with respect to the subspace A_b in the first step. In the second step, it sorts all the ensemble members in $\widehat{\Gamma}$ in ascending order according to the corresponding Δ values. The first ensemble member (A_t, χ_t) (where $t = 1$) is picked up and inserted into the new ensemble $\Gamma = \{(A_t, \chi_t)\}$. At the same time, (A_t, χ_t) is removed from the original ensemble $\widehat{\Gamma}$ as follows:

$$\widehat{\Gamma} = \widehat{\Gamma} \backslash \{(A_t, \chi_t)\}. \tag{24}$$

In the third step, each ensemble member (A_b, χ_b) in $\widehat{\Gamma}$ is considered in turn, and the local objective function ζ_b with respect to the ensemble member (A_t, χ_t) in Γ is calculated. All the ensemble members in $\widehat{\Gamma}$ are sorted in ascending order according to the corresponding local objective function ζ_b in the fourth step. IEMS will consider the ensemble member in $\widehat{\Gamma}$ one by one in the fifth step. It generates the new ensemble $\Gamma' = \Gamma + \{(A_b, \chi_b)\}$ (where $(A_b, \chi_b) \in \widehat{\Gamma}$), and calculates the global objective function $\Delta(I')$ and $\Delta(I)$for the clustering solutions I' and I generated by Γ' and Γ, respectively. If $\Delta(I') \leq \Delta(I)$, the ensemble member (A_b, χ_b) is inserted into the new ensemble: $\Gamma = \Gamma + \{(A_b, \chi_b)\}$, and removed from the original ensemble $\widehat{\Gamma}$: $\widehat{\Gamma} = \widehat{\Gamma} - (A_b, \chi_b)$. IEMS sets the ensemble member (A_t, χ_t) to (A_b, χ_b), and proceeds to the third step. It will perform the above last three steps repeatedly until B' $(B' < B)$ ensemble members are selected.

Algorithm 1. Incremental Semi-Supervised Clustering Ensemble Approach

Require:
Input: a high dimensional dataset P;
Ensure:
1: Original ensemble generation;
2: Generate B random subspaces $\{A_1, A_2, \ldots, A_B\}$;
3: Generate the semi-supervised clustering models $\chi_1, \chi_2, \ldots, \chi_B$ using E^{2}CP;
4: Call incremental ensemble member selection process in Algorithm 2;
5: New ensemble generation;
6: Generate B' random subspaces $\{A_1, A_2, \ldots, A_{B'}\}$ $(B' < B)$;
7: Generate semi-supervised clustering models $\chi_1, \chi_2, \ldots, \chi_{B'}$ using E^{2}CP;
8: Obtain consensus matrix O by summarizing the clustering solutions $\{I^1, I^2, \ldots, I^{B'}\}$ generated by the semi-supervised clustering models;
9: Consensus function for the final results using the normalized cut approach;
Output: the labels of the samples in P.

The incremental ensemble member selection process makes use of the global objective function Δ and the local objective function ζ. Assume that (1) M and N are the sets of must-link and cannot-link pairs, respectively. (2) $W^M = \{w_{ij}^M\}$ and $W^N = \{w_{ij}^N\}$ denote the weight sets which contain weights corresponding to the must-link and cannot-link constraints, respectively. (3) $Y = \{y_1, y_2, \ldots, y_n\}$ are the set of labels for the feature vectors in P. (4) The cost of violating constraints is specified by the generalized Potts metric. Given a clustering solution I, the global objective function $\Delta(I)$ of IEMS, which is motivated from the cost function of PC-Kmeans [49], is defined as follows:

$$\Delta(I) = \frac{1}{2}\sum_{p_i \in P}\sum_{h=1}^{k}\theta(y_i = h)d(p_i, \mu_h) + \sum_{p_i, p_j \in M} w_{ij}^{M}\theta(y_i \neq y_j) + \sum_{p_i, p_j \in N} w_{ij}^{N}\theta(y_i = y_j), \tag{25}$$

$$\mu_h = \frac{\sum_{i=1}^{h}\theta(y_i = h)p_i}{\sum_{i=1}^{h}\theta(y_i = h)}, \tag{26}$$

where $d(p_i, \mu_h)$ denotes the Euclidean distance between the feature vectors p_i and μ_h, θ denotes an indicator function, $\theta(true) = 1$ and $\theta(false) = 0$. The objective of the cost function is to optimize the squared distances of the feature vectors from the centers, such that as many constraints are satisfied as possible.

Algorithm 2. Incremental Ensemble Member Selection

Require:
Input: the sample set $P = (p_1, p_2, \ldots, p_l)$;
the must-link set M;
the cannot-link set N;
a set of random subspaces $A = \{A_1, \ldots A_B\}$;
a set of semi-supervised clustering models $\chi = \{\chi_1, \ldots, \chi_B\}$;
a set of ensemble members $\widehat{\Gamma} = \{(A_1, \chi_1), (A_2, \chi_2), \ldots, (A_B, \chi_B)\}$;
the empty ensemble Γ;
Ensure:
1: **For** b in $1, \ldots, B$
2: Calculate the objective function $\Delta(I^b)$ using Eq. 25 for each clustering solution I^b generated by E^2CP;
3: Set $t = 1$;
4: Sort ensemble members in ascending order according to the corresponding Δ, and pick up the first ensemble member (A_t, χ_t);
5: Add to new ensemble: $\Gamma = \{(A_t, \chi_t)\}$, $\widehat{\Gamma} = \widehat{\Gamma} - \{(A_t, \chi_t)\}$;
6: **Repeat**
7: t=t+1;
8: **For** each (A_b, χ_b) in $\widehat{\Gamma}$
9: Calculate the local objective function ζ_b;
10: Sort ensemble members in $\widehat{\Gamma}$ in ascending order according to the corresponding local objective function ζ_b;
11: Set $b = 0$;
12: **Repeat**
13: Set $b = b + 1$;
14: Generate new ensemble $\Gamma' = \Gamma + \{(A_b, \chi_b)\}$ (where $(A_b, \chi_b) \in \widehat{\Gamma}$);
15: Calculate the global objective function $\Delta(I')$ and $\Delta(I)$for the clustering solutions I' and I generated by Γ' and Γ respectively;
16: **Until** $\Delta(I') \leq \Delta(I)$;
17: Add to new ensemble: $\Gamma = \Gamma + \{(A_b, \chi_b)\}$, $\widehat{\Gamma} = \widehat{\Gamma} - \{(A_b, \chi_b)\}$;
18: **Until** $t \geq B'$ or $\Gamma = \emptyset$;
Output: the new ensemble Γ.

Given the original ensemble $\widehat{\Gamma}$ and the new ensemble Γ, the local objective function ζ_b for the local b-th ensemble member $(A_b, \chi_b) \in \widehat{\Gamma}$ with respect to the ensemble member $(A_t, \chi_t) \in \Gamma$ is defined as follows:

$$\zeta_b = \sum_{\forall A_t \in \Gamma} \frac{S(A_b, A_t)}{\Delta(I^b)}, \tag{27}$$

where $\Delta(I^b)$ denotes the global objective function for the clustering solution I^b, and $S(A_b, A_t)$ denotes the similarity function between two subspaces A_b and A_t.

Given the subspaces A_b and A_t, the set of attributes in these subspaces can be represented by Gaussian mixture models (GMMs) $\Omega^b = \{\Phi_1^b = (w_1^b, \mu_1^b, \Sigma_1^b), \Phi_2^b = (w_2^b, \mu_2^b, \Sigma_2^b), \ldots, \Phi_{k_1}^b = (w_{k_1}^b, \mu_{k_1}^b, \Sigma_{k_1}^b)\}$ and $\Omega^t = \{\Phi_1^t = (w_1^t, \mu_1^t, \Sigma_1^t), \Phi_2^t = (w_2^t, \mu_2^t, \Sigma_2^t), \ldots, \Phi_{k_2}^t = (w_{k_2}^t, \mu_{k_2}^t, \Sigma_{k_2}^t)\}$ respectively (where $w_{h_1}^b$, $\mu_{h_1}^b$ and $\Sigma_{h_1}^b$, ($h_1 \in \{1, \ldots, k_1\}$) denote the weight value, the mean vector and the covariance matrix for the h_1-th component $\Phi_{h_1}^b$ of Ω^b, respectively. $w_{h_2}^t$, $\mu_{h_2}^t$ and $\Sigma_{h_2}^t$, ($h_2 \in \{1, \ldots, k_2\}$) denote the weight value, the mean vector and the covariance matrix for the h_2-th component $\Phi_{h_2}^t$ of Ω^t, respectively). The expectation-maximization approach (EM) is adopted to perform clustering on the set of attributes in the subspace, and determine the optimal parameter values of GMMs.

Algorithm 3 provides a flowchart of the similarity function (SF) for $S(A_b, A_t)$. The input of SF is two Gaussian mixture models Ω^b and Ω^t, while the output is the similarity value $S(A_b, A_t)$ between two subspaces A_b and A_t. Specifically, the similarity function first considers the similarity of all the pairs of components in Ω^b and Ω^t. The Bhattacharyya distance $\varphi(\Phi_{h_1}^b, \Phi_{h_2}^t)$ is used to calculate the similarity between two components $\Phi_{h_1}^b$ in Ω^b and $\Phi_{h_2}^t$ in Ω^t, which is as follows:

$$\varphi(\Phi_{h_1}^b, \Phi_{h_2}^t) = \frac{1}{8}(\mu_{h_1}^b - \mu_{h_2}^t)^T \left(\frac{\Sigma_{h_1}^b + \Sigma_{h_2}^t}{2}\right)^{-1}(\mu_{h_1}^b - \mu_{h_2}^t) + \frac{1}{2}\ln\frac{|\frac{\Sigma_{h_1}^b + \Sigma_{h_2}^t}{2}|}{\sqrt{|\Sigma_{h_1}^b||\Sigma_{h_2}^t|}}. \tag{28}$$

Then, SF sorts all the component pairs $(\Phi_{h_1}^b, \Phi_{h_2}^t)$ in ascending order according to the corresponding Bhattacharyya distance values, and inserts them into a queue. Next, it sets $S(A_b, A_t) = 0$, performs a de-queue operation, and considers the component pair $(\Phi_{h_1}^b, \Phi_{h_2}^t)$ one by one. If $w_{h_1}^b > 0$ and $w_{h_2}^t > 0$, the minimum weight w between the two is selected in the first step, which is as follows:

$$w = \min(w_{h_1}^b, w_{h_2}^t). \tag{29}$$

In the second step, the similarity value $S(A_b, A_t)$ is assigned a new value as follows:

$$S(A_b, A_t) = S(A_b, A_t) + w\varphi(\Phi_{h_1}^b, \Phi_{h_2}^t). \tag{30}$$

The weights $w_{h_1}^b$ and $w_{h_2}^t$ are updated in the third step as follows:

$$w_{h_1}^b = w_{h_1}^b - w, w_{h_2}^t = w_{h_2}^t - w \tag{31}$$

Finally, the similarity value $S(A_b, A_t)$ will be obtained by considering all the component pairs $(\Phi_{h_1}^b, \Phi_{h_2}^t)$ in the queue.

Algorithm 3. Similarity Function

Require:

Input: the Gaussian mixture model $\Omega^b = \{\Phi_1^b = (w_1^b, \mu_1^b, \Sigma_1^b), \ldots, \Phi_{k_1}^b = (w_{k_1}^b, \mu_{k_1}^b, \Sigma_{k_1}^b)\}$;
the Gaussian mixture model $\Omega^t = \{\Phi_1^t = (w_1^t, \mu_1^t, \Sigma_1^t), \ldots, \Phi_{k_2}^t = (w_{k_2}^t, \mu_{k_2}^t, \Sigma_{k_2}^t)\}$;

Ensure:

For h_1 in $1, \ldots, k_1$
  For h_2 in $1, \ldots, k_2$
    Calculate the Bhattacharyya distance $\varphi(\Phi_{h_1}^b, \Phi_{h_2}^t)$ using Eq. 28 ;
Sort all the component pairs $(\Phi_{h_1}^b, \Phi_{h_2}^t)$ in ascending order according to the corresponding Bhattacharyya distance values, and insert them into a queue;
$S(A_b, A_t) = 0$;
For j in $1, \ldots, k_1 k_2$ (where $k_1 k_2$ is the number of component pairs)
  Perform a de-queue operation and obtain the component pair $(\Phi_{h_1}^b, \Phi_{h_2}^t)$;
    If $w_{h_1}^b > 0$ and $w_{h_2}^t > 0$
      $w = \min(w_{h_1}^b, w_{h_2}^t)$;
      $S(A_b, A_t) = S(A_b, A_t) + w\varphi(\Phi_{h_1}^b, \Phi_{h_2}^t)$;
      $w_{h_1}^b = w_{h_1}^b - w$;
      $w_{h_2}^t = w_{h_2}^t - w$;

Output: the similarity value $S(A_b, A_t)$.

5 Theoretical Analysis

We also perform a theoretical analysis of ISSCE in terms of its computational cost. The corresponding time complexity T_{ISSCE} is estimated as follows:

$$T_{ISSCE} = T_{OE} + T_{IEMS} + T_{FR}, \tag{32}$$

where T_{OE}, T_{IEMS}, and T_{FR} denote the computational costs for the original ensemble generation step, the incremental ensemble member selection step, and the final result generation step using the new ensemble, respectively. T_{OE} is related to the number of feature vectors n, the number of attributes m, the number of random subspaces B, and the number of neighbors K in the K-NN graph in E^2CP as follows:

$$T_{OE} = O(BKn^2). \tag{33}$$

T_{IEMS} will be affected by the number of feature vectors n, the number of attributes m, the number of random subspaces B, the number of selected subspace B', and the number of clusters k in the clustering solution as follows:

$$T_{IEMS} = O(Bkn^2 + B\log B + B'nm^2). \tag{34}$$

T_{FR} is related to the number of feature vectors n and the number of clusters k in the clustering solution as follows:

$$T_{FR} = O(kn^3). \tag{35}$$

Since B, B', K and k are constants, which are significantly smaller than n^3 or nm^2, the computational cost of ISSCE is approximately $O(e^3)$, where $e = \max(n, m)$.

The space consumption of ISSCE is composed of the consumption of the original ensemble $O(Bnm)$, the new ensemble $O(B'nm)$, and the consensus matrix $O(n^2)$. As a result, the overall space consumption of ISSCE is $O(e^2)$.

TABLE 1
A Summary of Real-World Datasets (Where n Denotes the Number of Data Samples, m Denotes the Number of Attributes, and k Denotes the Number of Classes)

Dataset	Source	n	m	k
Iris	[14]	150	4	3
Movement_libras	[14]	360	90	15
RobotExecution1	[14]	88	90	4
RobotExecution2	[14]	47	90	5
RobotExecution4	[14]	117	90	3
Syntheticcontro	[14]	600	60	6
Alizadeh-2000-v3	[50]	62	2,093	4
Armstrong-2002-v2	[50]	72	2,194	3
Bredel-2005	[50]	50	1,739	3
Dyrskjot-2003	[50]	40	1,203	3
Liang-2005	[50]	35	1,411	3
Alizadeh-2000-v3(o)	[50]	62	4,026	4
Nutt-2003-v1	[50]	50	1,377	4
Pomeroy-2002-v2	[50]	42	1,379	5
Ramaswamy-2001	[50]	190	1,363	14
Risinger-2003	[50]	42	1,771	4
Su-2001	[50]	174	1,571	10
Tomlins-2006-v1	[50]	104	2,315	5

6 Experiment

The performances of ISSCE and other semi-supervised clustering approaches are evaluated using 18 real-world datasets as shown in Table 1 (where n denotes the number of data samples, m denotes the number of attributes, and k denotes the number of classes), which includes six datasets from UCI machine learning repository and 12 high dimensional datasets of cancer gene expression profiles. The preprocessing process for the cancer datasets is the same as that in [49].

Normalized mutual information (NMI, [51]) and adjusted Rand index (ARI, [51]) are used to evaluate the performances of ISSCE and its competitors. The performance of the proposed approach is evaluated by the average value and the corresponding standard deviation of NMI and ARI, respectively, after 10 runs.

Given the ground truth result I with k clusters $I = \{C_1, C_2, \ldots, C_k\}$, and the result I' obtained by ISSCE with k' clusters $I' = \{C'_1, C'_2, \ldots, C'_{k'}\}$, we use NMI [51] to evaluate the quality of the final clustering result, which is defined as follows:

$$NMI(I, I') = \frac{2H_1(I; I')}{H_2(I) + H_2(I')}, \tag{36}$$

$$H_1(I; I') = \sum_h \sum_l \frac{|C_h \cap C'_l|}{n} \log \frac{n|C_h \cap C'_l|}{|C_h||C'_l|}, \tag{37}$$

$$H_2(I) = -\sum_h \frac{|C_h|}{n} \log \frac{|C_h|}{n}, \tag{38}$$

$$H_2(I') = -\sum_l \frac{|C'_l|}{n} \log \frac{|C'_l|}{n}, \tag{39}$$

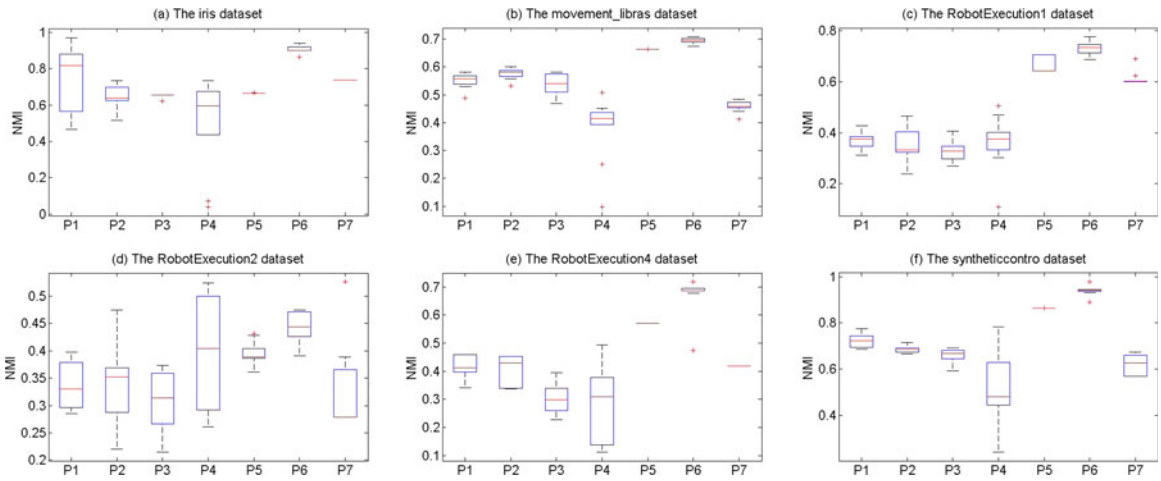


Fig. 3. The comparison of semi-supervised clustering ensemble approaches on six real-world datasets from the UCI machine learning repository in Table 1 (where P1, ..., P7 denote the semi-supervised clustering ensemble method based on the random subspace technique and PC-Kmeans (P1), the cluster ensemble approach based on the random subspace technique and K-means (P2), the cluster ensemble approach based on the bagging technique and K-means (P3), the cluster ensemble approach based on hierarchical clustering (P4), the semi-supervised clustering ensemble method based on the constraint propagation approach (P5), the incremental semi-supervised clustering ensemble approach (P6), and the semi-supervised kernel based K-means clustering ensemble algorithm (P7), respectively.)

where $h \in \{1, \ldots, k\}$, $l \in \{1, \ldots, k'\}$, n is the total number of data samples, and $|\cdot|$ denotes the cardinality of the cluster. The larger the NMI value is, the better the quality of clusters becomes.

The adjusted Rand index $ARI(I, I')$ [51] is defined as follows:

$$ARI(I, I') = \frac{\sum_{h=1}^{k}\sum_{l=1}^{k'}\binom{|C_h \cap C'_l|}{2} - \varrho_3}{\frac{1}{2}(\varrho_1 + \varrho_2) - \varrho_3}, \tag{40}$$

$$\varrho_1 = \sum_{h=1}^{k}\binom{|C_h|}{2}, \varrho_2 = \sum_{l=1}^{k'}\binom{|C'_l|}{2}, \varrho_3 = \frac{2\varrho_1\varrho_2}{n(n-1)}. \tag{41}$$

A higher ARI value corresponds to a higher clustering solution quality.

In the following experiments, we first study the effect of the parameters. Then, the effect of the incremental ensemble member selection process is explored. Next, the proposed approach ISSCE is compared with single semi-supervised clustering methods and semi-supervised clustering ensemble approaches on the real-world datasets. Finally, a set of nonparametric tests are adopted to compare multiple semi-supervised clustering ensemble approach over different datasets.

6.1 The Effect of the Parameters

We conduct experiments on six cancer gene expression data with respect to the average NMI values to investigate the effect of the parameters, which include the sampling rate in the random subspace technique, the number of pairwise constraints, and the number of nearest neighbors used.

In order to explore the effect of the sampling rate τ, we vary its value from 0.1 to 0.5 with an increment 0.05. Fig. 1 in the supplementary file shows the effectiveness of the sampling rate in the random subspace technique. It is observed that when $\tau = 0.3$, ISSCE achieves good performance on most of the datasets, such as the Armstrong-2002-v2 dataset, the Dyrskjot-2003 dataset, the Liang-2005 dataset and the Ramaswamy-2001 dataset. A possible reason could be that when $\tau = 0.3$, the underlying structure of the datasets can be adequately captured by the subspace generated by the random subspace technique. As a result, τ is set to 0.3 in the following experiments.

We also vary the number of pairwise constraints from 0.5 to $2\,n$ with an increment $0.5\,n$ to study its effect. As shown in Fig. 2 in the supplementary file, when the number of pairwise constraints increases, the average NMI values increase gradually on most of the datasets, such as the Armstrong-2002-v2 dataset, the Liang-2005 dataset the Su-2001 dataset, and so on. The possible reason is that more number of pairwise constraints will provide additional useful information, which will improve the final results. However, a large number of pairwise constraints require more efforts from experts. In order to keep a balance between the workload of the experts and the performance of the algorithm, we set n as the default value for the number of pairwise constraints.

In order to explore the effect of the number of nearest neighbors in Eq. (3), we vary q from 3 to 11 with an increment of two. The performance of ISSCE is not sensitive to q as shown in Fig. 3 in the supplementary file. As a result, q is set to 3 to avoid additional computational cost.

6.2 The Effect of the Incremental Ensemble Member Selection Process

In order to investigate the effect of the incremental ensemble member selection process, we compare ISSCE and PCKE-IEMS with E^{2}CPE and PCKE with respect to NMI on all the datasets in Table 1. PCKE-IEMS denotes ISSCE with the selection process incorporated and using PC-Kmeans [50] as the basic semi-supervised clustering model. E^{2}CPE denotes ISSCE without using the selection process and applying E^{2}CP [12] as the basic semi-supervised clustering model. PCKE denotes ISSCE without the selection process and using PC-Kmeans as the basic semi-supervised clustering model.

TABLE 2
The Comparison of ISSCE with Selection and E^{2}CPE without Selection (Where the Best Values Are Highlighted in Bold)

Datasets	E^{2}CPE	ISSCE	Difference
Iris	0.6646 ± 0.0100	**0.9050** ± 0.0200	0.2404
Movemen_libras	0.6626 ± 0	**0.6946** ± 0.0100	0.0320
RobotExetcution1	0.6675 ± 0.0316	**0.7330** ± 0.0265	0.0655
RobotExecution2	0.3946 ± 0.0224	**0.4433** ± 0.0283	0.0487
RobotExecution4	0.5715 ±0	**0.6720** ± 0.0707	0.1005
Syntheticcontro	0.8635 ±0	**0.9396**± 0.0224	0.0761
Alizadeh-2000-v3	0.6985 ± 0	**0.7210** ± 0.0141	0.0225
Armstrong-2002-v2	0.8389±0.0100	**0.8536** ± 0.0141	0.0147
Bredel-2005	0.4425 ± 0	**0.5666** ± 0.0300	0.1241
Dyrskjot-2003	0.6620 ± 0	**0.7345** ± 0.0836	0.0725
Liang-2005	0.5950 ± 0.0173	**0.6270** ± 0.0458	0.0320
Alizadeh-2000-v3(o)	0.7884 ± 0	**0.8221** ± 0.0265	0.0337
Nutt-2003-v1	0.3983 ± 0.0173	**0.4749** ± 0.0400	0.0766
Pomeroy-2002-v2	0.5688 ± 0.0224	**0.6408** ± 0.0671	0.0720
Ramaswamy-2001	0.6527 ± 0	**0.6719** ± 0.0141	0.0192
Risinger-2003	0.5018 ± 0.0600	**0.5240** ± 0.0141	0.0222
Su-2001	0.7273 ± 0	**0.7408** ± 0.0200	0.0135
Tomlins-2006-v1	0.5770 ± 0.0100	**0.6557** ± 0.0245	0.0787

TABLE 3
The Comparison of PCKE-IEMS with Selection and PCKE without Selection

Datasets	PCKE	PCKE-IEMS
Iris	**0.7398** ± 0.1797	0.6558 ± 0.2241
movemen_libras	0.5499 ± 0.0265	**0.5516** ± 0.0224
RobotExetcution1	**0.3725** ± 0.0374	0.3560 ± 0.0529
RobotExecution2	0.3344 ± 0.0424	**0.3680** ± 0.0775
RobotExecution4	0.4137 ±0.0400	**0.4556** ±0.0374
syntheticcontro	**0.7239** ±0.0316	0.7134 ± 0.0265
Alizadeh-2000-v3	0.4774 ± 0.1649	**0.4987** ± 0.1575
Armstrong-2002-v2	0.6400 ± 0.1523	**0.6867** ± 0.1404
Bredel-2005	0.5280 ± 0.0922	**0.5401** ±0.2278
Dyrskjot-2003	0.5835 ±0.1523	**0.6192** ± 0.2198
Liang-2005	0.3799 ± 0.1200	**0.4458** ± 0.1476
Alizadeh-2000-v3(o)	**0.6274** ± 0.1140	0.6157 ± 0.0970
Nutt-2003-v1	0.4463 ± 0.1049	**0.4862** ± 0.0889
Pomeroy-2002-v2	0.4904 ± 0.0995	**0.6181** ± 0.0877
Ramaswamy-2001	**0.4823** ±0.0316	0.4390 ± 0.0436
Risinger-2003	0.4111 ± 0.0616	**0.4438** ± 0.0954
Su-2001	0.6060 ± 0.0954	**0.6118** ± 0.1049
Tomlins-2006-v1	0.4162 ± 0.0995	**0.4305** ± 0.0480

Table 2 shows the results obtained by ISSCE and E^{2}CPE, while Table 3 shows the results obtained by PCKE-IEMS and PCKE. It can be seen that ISSCE outperforms E^{2}CPE on all of the datasets, and PCKE-IEMS obtains better results on 13 out of 18 datasets when compared with PCKE. For example, the NMI values 0.9050, 0.7330, 0.6720, 0.9396, 0.5666, 0.7345, 0.4749, 0.6408 and 0.6557 obtained by ISSCE on the Iris, RobotExetcution1, RobotExecution4, Syntheticcontro, Bredel-2005, Dyrskjot-2003, Nutt-2003-v1, Pomeroy-2002-v2 and Tomlins-2006-v1 datasets, respectively, are 0.2404, 0.0655, 0.1005, 0.0761, 0.1241, 0.0725, 0.0766, 0.0720, 0.0787 larger than those obtained by E^{2}CPE. This indicates that ISSCE is significantly better than E^{2}CPE on the above nine datasets with a significance level of 0.05. ISSCE is suitable for the datasets with an underlying subspace structure, such as the Iris, RobotExecution4 and Bredel-2005 . The possible reasons could be that IEMS make use of the local and global objective functions to select the useful ensemble members and remove the redundant members. As a whole, the incremental ensemble member selection process is a general technique which can be incorporated into different ensemble methods to improve their performance.

Another interesting observation from Tables 2 and 3 is that ISSCE outperforms PCKE-IEMS on 17 out of 18 datasets. This indicates that the basic semi-supervised clustering model E^{2}CP plays an important role in ISSCE. Once E^{2}CP is replaced by PC-Kmeans, the performances of ISSCE on most of the datasets become less satisfactory.

6.3 The Comparison of Single Semi-Supervised Clustering Approaches

We have also compared ISSCE with the pairwise constraints based K-means algorithm (PC-Kmeans, [50]), the constraint propagation approach (E^{2}CP, [12]), and the semi-supervised kernel based K-means algorithm (SSKK, [56]) based on NMI on all the datasets in Table 1. The possible reason for the selection of PC-Kmeans [50] is that it is one of the most popular semi-supervised clustering algorithms. It will serve as the baseline for the comparison of other semi-supervised clustering algorithms.

Table 4 shows the results obtained by ISSCE and single semi-supervised clustering approaches with respect to the average value and standard deviations of NMI on all the real-world datasets. As shown in Table 4, ISSCE outperforms other semi-supervised clustering algorithms, such as E^{2}CP, on 17 out of 18 datasets. Since most of them are high dimensional datasets, this indicates that ISSCE is able to alleviate the effect of high dimensionality. The possible reasons are as follows: (1) The random subspace technique is able to reduce the dimension of the original space, which is useful for discovering the underlying structure of the dataset in the low dimensional space. (2) ISSCE integrates multiple clustering solutions generated from different subspaces into a unified clustering solution, which provides more accurate, robust and stable final results. While the computation time is longer for ISSCE when compared with single semi-supervised clustering algorithms, it generates final results of higher quality, which represents a worthwhile tradeoff.

6.4 The Comparison of Semi-Supervised Clustering Ensemble Approaches

In the following experiments, ISSCE (P6) is compared with a number of state-of-the-art cluster ensemble algorithms and semi-supervised clustering ensemble algorithms, which include the semi-supervised clustering ensemble method based on the random subspace technique and PC-Kmeans (RSPCKE, P1, [50]), the cluster ensemble approach based on the random subspace technique and K-means (RSKE, P2, [8]), the cluster ensemble approach based on the bagging technique and K-means (BAGKE, P3, [2]), the cluster ensemble approach based on hierarchical clustering (HCCE, P4, [15]), the semi-supervised clustering ensemble method based on the constraint propagation approach (E^{2}CPE, P5, [12]), and the semi-supervised kernel based K-means clustering ensemble algorithm (SSKKE, P7, [56]).

Figs. 3 and 4 show the performance of seven semi-supervised clustering methods or clustering ensemble approaches in terms of the average value and standard

TABLE 4
The Comparison with Single Semi-Supervised Clustering Approaches
(Where the Best Values Are Highlighted in Bold)

Datasets	ISSCE	PC-kmeans	E^2CP	SSKK
Iris	**0.9050** $\pm$ 0.0200	0.6644 $\pm$ 0.2005	0.6646 $\pm$ 0.0100	0.4454 $\pm$ 0.0624
Movement_libras	**0.6946** $\pm$ 0.0100	0.5455 $\pm$ 0.0316	0.604 $\pm$ 0	0.4328 $\pm$ 0.0173
RobotExecution1	**0.7330** $\pm$ 0.0265	0.4034 $\pm$ 0.0860	0.7177 $\pm$ 0	0.6668 $\pm$ 0.0728
RobotExecution2	**0.4433** $\pm$ 0.0283	0.3532 $\pm$ 0.0480	0.4293 $\pm$ 0.0173	0.4008 $\pm$ 0.0671
RobotExecution4	**0.6720** $\pm$ 0.0707	0.4863 $\pm$ 0.0100	0.4829 $\pm$ 0.0265	0.5002 $\pm$ 0
Syntheticcontro	**0.9396** $\pm$ 0.0224	0.7078 $\pm$ 0.0283	0.9234 $\pm$ 0	0.5649$\pm$ 0.0387
Alizadeh-2000-v3	**0.7210** $\pm$ 0.0141	0.5210 $\pm$ 0.1565	0.7084 $\pm$ 0	0.3487 $\pm$ 0.0728
Armstrong-2002-v2	**0.8536** $\pm$ 0.0141	0.6244 $\pm$ 0.1217	0.8424$\pm$ 0	0.5752 $\pm$ 0.2848
Bredel-2005	**0.5666** $\pm$ 0.0300	0.5426 $\pm$ 0.0693	0.3763 $\pm$ 0	0.2126 $\pm$ 0.0100
Dyrskjot-2003	**0.7345** $\pm$ 0.0837	0.6402 $\pm$ 0.1212	0.6891 $\pm$ 0	0.2585 $\pm$ 0.0332
Liang-2005	**0.6270** $\pm$ 0.0458	0.4519 $\pm$ 0.1442	0.6011 $\pm$ 0.0245	0.2869 $\pm$ 0
Alizadeh-2000-v3(o)	**0.8221** $\pm$ 0.0265	0.6213 $\pm$ 0.1020	0.8138 $\pm$ 0.0141	0.3531 $\pm$ 0.0361
Nutt-2003-v1	0.4749 $\pm$ 0.0400	**0.5490** $\pm$ 0.1105	0.4426 $\pm$ 0.0245	0.2387 $\pm$ 0.0906
Pomeroy-2002-v2	**0.6408** $\pm$ 0.0671	0.4340 $\pm$ 0.1364	0.6198 $\pm$ 0.0469	0.2989 $\pm$ 0.0837
Ramaswamy-2001	**0.6719** $\pm$ 0.0141	0.5229 $\pm$ 0.0224	0.6637 $\pm$ 0.0100	0.4428 $\pm$ 0.0510
Risinger-2003	**0.5240** $\pm$ 0.0141	0.4255 $\pm$ 0.1389	0.4962 $\pm$ 0	0.3332 $\pm$ 0.0721
Su-2001	**0.7428** $\pm$ 0.0200	0.6203 $\pm$ 0.0860	0.7050 $\pm$ 0.0100	0.3043 $\pm$ 0.1158
Tomlins-2006-v1	**0.6557** $\pm$ 0.0245	0.4269 $\pm$ 0.0854	0.6203 $\pm$ 0.0100	0.5678 $\pm$ 0.0283

deviation of NMI on six real-world datasets from the UCI machine learning repository, and 12 high dimensional datasets of cancer gene expression profiles in Table 1. Three interesting observations can be obtained from these figures. First, ISSCE (P6) achieves the best performance on 5 out of 6 datasets in Fig. 3 and nine out of 12 datasets in Fig. 4. The

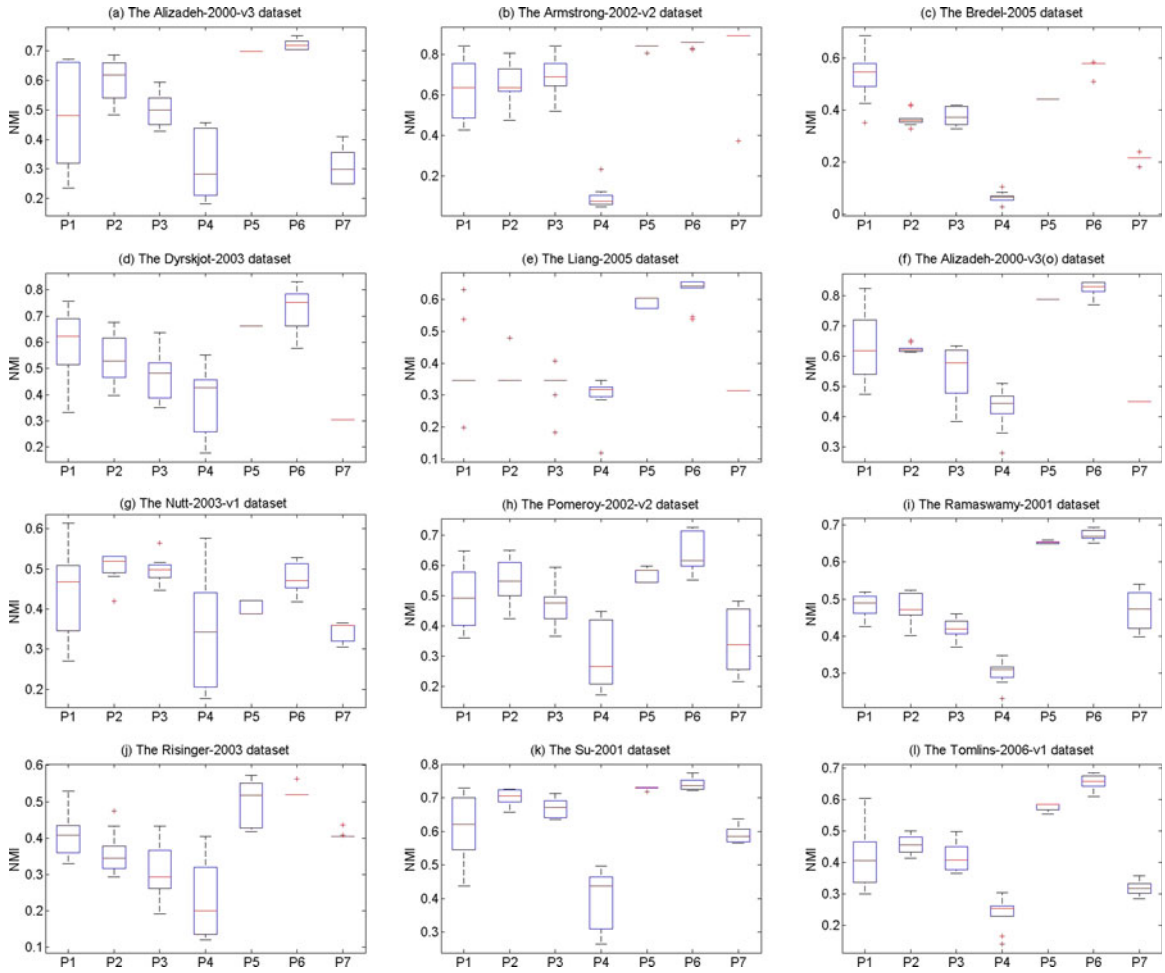


Fig. 4. The comparison of semi-supervised clustering ensemble approaches on 12 high dimensional datasets from cancer gene expression profiles in Table 1.

TABLE 5
The Comparison with Semi-Supervised Clustering Ensemble Approaches on the Datasets in Table 1 with Respect to ARI (Where the Best Values Are Highlighted in Bold)

Datasets	RSPCKE	RSKE	BAGKE	HCCE	E^{2}CPE	ISSCE	SSKKE
Iris	0.8519 ± 0.1651	0.8661 ± 0.1303	0.9216 ± 0.0076	0.955 ± 0.0543	0.9549 ± 0.0495	$\mathbf{1} \pm 0$	0.8324 ± 0.0884
Movement_libras	0.3157 ± 0.0084	0.3053 ± 0.0195	0.2865 ± 0.0343	0.1149 ± 0.0379	$\mathbf{0.3757} \pm 0$	0.3572 ± 0.0196	0.1227 ± 0.013
RobotExecution1	0.1612 ± 0.0326	0.1112 ± 0.0917	0.2545 ± 0.0797	0.0833 ± 0.1521	0.4827 ± 0.0015	$\mathbf{0.6968} \pm 0.0499$	0.4464 ± 0.0629
RobotExecution2	0.2959 ± 0.0986	0.2866 ± 0.1089	0.1545 ± 0.0939	0.3248 ± 0.1612	0.2172 ± 0.0255	$\mathbf{0.2976} \pm 0.087$	0.2721 ± 0.101
RobotExecution4	0.356 ± 0.0171	$\mathbf{0.3725} \pm 0.0344$	0.2299 ± 0.0276	0.2391 ± 0.1142	0.3378 ± 0.0001	0.2604 ± 0.103	0.4251 ± 0
Syntheticcontro	0.5653 ± 0.0248	0.5503 ± 0.026	0.5086 ± 0.02	0.3141 ± 0.2346	0.7464 ± 0.0014	$\mathbf{0.9357} \pm 0.0821$	0.2832 ± 0.0297
Alizadeh-2000-v3	0.266 ± 0.1093	0.373 ± 0.069	0.4155 ± 0.093	0.1487 ± 0.0723	0.4634 ± 0	$\mathbf{0.4928} \pm 0.0705$	0.1477 ± 0.0273
Armstrong-2002-v2	0.6506 ± 0.1352	0.5796 ± 0.1573	0.6427 ± 0.1253	0.0271 ± 0.067	0.8431 ± 0	$\mathbf{0.8442} \pm 0.0434$	0.6922 ± 0.1708
Bredel-2005	0.4309 ± 0.1225	0.3206 ± 0.0333	0.2706 ± 0.0645	0.0109 ± 0.0333	0.4873 ± 0.0479	$\mathbf{0.4975} \pm 0.0244$	0.0513 ± 0.0222
Dyrskjot-2003	0.5209 ± 0.0933	0.543 ± 0.0542	0.4251 ± 0.1142	0.3865 ± 0.0842	0.4924 ± 0.0722	$\mathbf{0.6674} \pm 0.0569$	0.1837 ± 0.0487
Liang-2005	0.1436 ± 0.0616	0.1573 ± 0.0032	0.1104 ± 0.1936	0.1274 ± 0.056	0.1431 ± 0	$\mathbf{0.3767} \pm 0.0213$	0.0109 ± 0.0028
Alizadeh-2000-v3(o)	0.4086 ± 0.0863	0.4205 ± 0.0354	0.3253 ± 0.0583	0.1626 ± 0.0906	$\mathbf{0.7545} \pm 0$	0.5981 ± 0.0119	0.2536 ± 0.0318
Nutt-2003-v1	0.3391 ± 0.0605	$\mathbf{0.3805} \pm 0.0295$	0.3171 ± 0.0346	0.1325 ± 0.0724	0.2597 ± 0.0143	0.2719 ± 0.0747	0.2646 ± 0.0106
Pomeroy-2002-v2	0.3541 ± 0.159	0.3422 ± 0.0865	0.2409 ± 0.0803	0.0606 ± 0.0418	0.3471 ± 0	$\mathbf{0.3627} \pm 0.0375$	0.0482 ± 0.093
Ramaswamy-2001	0.1577 ± 0.0532	0.1205 ± 0.028	0.2275 ± 0.0382	0.0652 ± 0.0332	0.4431 ± 0.0047	$\mathbf{0.465} \pm 0.0887$	0.1421 ± 0.073
Risinger-2003	$0.3216 \pm 0.0.0812$	0.3017 ± 0.0638	0.221 ± 0.0907	0.0815 ± 0.1062	0.3309 ± 0	$\mathbf{0.3475} \pm 0.0426$	0.2278 ± 0.1428
Su-2001	0.4607 ± 0.1043	0.5283 ± 0.0702	0.5344 ± 0.0552	0.1763 ± 0.0932	0.587 ± 0.0135	$\mathbf{0.6643} \pm 0.0416$	0.2599 ± 0.0593
Tomlins-2006-v1	$0.2354 \pm 0.0.064$	0.2842 ± 0.0662	0.2436 ± 0.0537	0.1045 ± 0.0558	0.5523 ± 0.0348	$\mathbf{0.4815} \pm 0.0317$	0.1527 ± 0.046

main reasons are as follows: (1) The random subspace technique is useful for ISSCE to alleviate the effect of high dimensionality. (2) The incremental ensemble member selection process removes the redundant members in ISSCE, and improves its performance. (3) The normalized cut algorithm is effective for partitioning the consensus matrix and obtaining the final result. Second, E^{2}CPE (P5) obtains satisfactory results on most of the datasets. The possible reason could be that the constraint propagation approach is good at conveying information encapsulated in the pairwise constraints to the unlabeled samples such that the overall quality of the final result could be improved. Third, the results obtained by the cluster ensemble approaches, such as RSKE (P2), BAGKE (P3) and HCCE (P4) are less satisfactory than those by semi-supervised clustering approaches. This indicates that prior knowledge, which is represented by the pairwise constraints, is useful for improving the performance of the ensemble approaches. In summary, ISSCE is the best choice when dealing with different kinds of real-world datasets, especially the high dimensional datasets.

6.5 Nonparametric Tests

We adopt a set of nonparametric tests [52], [53] to compare multiple semi-supervised clustering ensemble approaches, which include ISSCE, PCKE, RSPCKE, PCKE-IEMS, RSKE, BAGKE, HCCE, E^{2}CP, SSKKE, and E^{2}CPE, over 18 real-world datasets in Table 1, in order to identify the significant difference among the results obtained by the different approaches in Figs. 3 and 4. Tables 1, 2, and 3 in the supplementary file show the results of multiple comparison of the semi-supervised clustering ensemble approaches using different nonparametric statistical procedures, which include the Bonferroni-Dunn test, the Holm test, the Hochberg test and the Hommel test. The average ranking of the classifier ensemble approaches is shown in Table 1 in the supplementary file. It is observed that ISSCE attains the highest average ranking, when compared with other semi-supervised clustering ensemble approaches. In general, the results show the extent of improvement of ISSCE over other semi-supervised clustering ensemble approaches.

7 Applications on More Real-World Datasets

To further evaluate the performances of the proposed approach, we apply ISSCE on more large real-world datasets, which are shown in Table 6. These include the cane-9 document dataset (Cane-9), the Mfeat handwritten digit dataset (Mfeat1), the Semeion handwritten digit dataset (Semeion), the multiple feature handwritten digit dataset (Mfeat2), and the USPS handwritten digit dataset (USPS). Most of them are image datasets. Tables 7 and 8 show the results obtained by different semi-supervised clustering ensemble approaches on the new datasets with respect to the average value and the corresponding standard deviation of NMI and ARI respectively. It can be seen that ISSCE outperforms other approaches on four out of five datasets with respect to NMI, and all of the datasets based on ARI, which indicates that ISSCE is effective for handling high dimensional datasets.

We also adopt the adjusted Rand index to evaluate the performance of the approaches. Tables 5 and 8 show the results obtained by different semi-supervised clustering ensemble approaches with respect to the average values and the corresponding standard deviations of ARI on the datasets in Tables 1 and 6 respectively.

8 Conclusion And Future Work

In this paper, we propose a new semi-supervised clustering ensemble approach, which is referred to as the incremental

TABLE 6
A Summary of the Real-World Datasets (Where n Denotes the Number of Data Samples, m Denotes the Number of Attributes, and k Denotes the Number of Classes)

Dataset	Source	n	m	k
Cane-9	[14]	1,080	856	9
Mfeat1	[14]	2,000	649	10
Semeion	[14]	1,593	256	10
Mfeat2	[14]	2,000	520	10
USPS	[14]	400	256	10

TABLE 7
The Comparison with Semi-Supervised Clustering Ensemble Approaches on the Dataset in Table 6 with Respect to NMI (Where the Best Values Are Highlighted in Bold)

Datasets	RSPCKE	RSKE	BAGKE	HCCE	E^{2}CPE	ISSCE	SSKKE
Cane-9	0.4114 ± 0.0682	0.3095 ± 0.0455	0.369 ± 0.0179	0.0433 ± 0.0148	**0.6497** ± 0	0.5152 ± 0.0633	0.2619 ± 0.0317
Mfeat1	0.5859 ± 0.0531	0.5908 ± 0.0468	0.5152 ± 0.0452	0.3457 ± 0.2093	0.7371 ± 0	**0.8565** ± 0.0498	0.3349 ± 0.0172
Semeion	0.0.5392 ± 0.159	0.4996 ± 0.0119	0.5275 ± 0.0389	0.1295 ± 0.1238	0.6452 ± 0	**0.6561** ± 0.028	0.5039 ± 0.0306
Mfeat2	0.6263 ± 0.0234	0.6179 ± 0.0318	0.6048 ± 0.0425	0.3427 ± 0.2531	0.7745 ± 0.0004	**0.9** ± 0.0017	0.5022 ± 0.0217
USPS	0.5112 ± 0.0577	0.5632 ± 0.0418	0.5822 ± 0.0191	0.2573 ± 0.1646	0.5813 ± 0.001	**0.608** ± 0.0174	0.4965 ± 0.0441

TABLE 8
The Comparison with Semi-Supervised Clustering Ensemble Approaches on the Dataset in Table 6 with Respect to ARI

Datasets	RSPCKE	RSKE	BAGKE	HCCE	E^{2}CPE	ISSCE	SSKKE
Cane-9	0.236 ± 0.0593	0.2303 ± 0.036	0.2193 ± 0.0273	0.0178 ± 0.0143	0.4088 ± 0.0007	**0.4207** ± 0.0322	0.0819 ± 0.0148
Mfeat1	0.4477 ± 0.0665	0.5089 ± 0.0505	0.3815 ± 0.047	0.2487 ± 0.1734	0.6076 ± 0.0435	**0.6727** ± 0.1738	0.2591 ± 0.0654
Semeion	0.4161 ± 0.0266	0.3352 ± 0.0586	0.395 ± 0.0644	0.0527 ± 0.0634	0.5811 ± 0.0017	**0.6067** ± 0.0296	0.2724 ± 0.0404
Mfeat2	0.5762 ± 0.0082	0.5027 ± 0.0569	0.4976 ± 0.0382	0.1981 ± 0.1284	0.6849 ± 0.0003	**0.7953** ± 0.0047	0.2173 ± 0.0283
USPS	0.4241 ± 0.0558	0.4766 ± 0.0234	0.4129 ± 0.0886	0.193 ± 0.1618	0.3716 ± 0.0	**0.4994** ± 0.0262	0.2609 ± 0.0341

semi-supervised clustering ensemble approach. Our major contribution is the development of an incremental ensemble member selection process based on a global objective function and a local objective function. In order to design a good local objective function, we also propose a new similarity function to quantify the extent to which two sets of attributes in the subspaces are similar to each other. We conduct experiments on six real-world datasets from the UCI machine learning repository and 12 real-world datasets of cancer gene expression profiles, and obtain the following observations: (1) The incremental ensemble member selection process is a general technique which can be used in different semi-supervised clustering ensemble approaches. (2) The prior knowledge represented by the pairwise constraints is useful for improving the performance of ISSCE. (3) ISSCE outperforms most conventional semi-supervised clustering ensemble approaches on a large number of datasets, especially on high dimensional datasets. In the future, we shall perform theoretical analysis to further study the effectiveness of ISSCE, and consider how to combine the incremental ensemble member selection process with other semi-supervised clustering ensemble approaches. We shall also investigate how to select parameter values depending on the structure/complexity of the datasets.

Acknowledgments

The authors are grateful for the constructive advice received from the anonymous reviewers of this paper. The work described in this paper was partially funded by the grant from the National High-Technology Research and Development Program (863 Program) of China No. 2013AA01A212, the grant from the NSFC for Distinguished Young Scholars 61125205, and the grants from the NSFC No. 61332002, No. 61300044, No. 61472145, No. 61572199, and No. 61502173, the grant from the Guangdong Natural Science Funds for Distinguished Young Scholars (No. S2013050014677), the Fundamental Research Funds for the Central Universities (No. 2014G0007 and 2015PT016), the grant from the key lab of cloud computing and big data in Guangzhou (No. SITGZ [2013]268-6), the grant from Science and Technology Planning Project of Guangdong Province, China (No. 2015A050502011), the grant from Key Enterprises and Innovation Organizations in Nanshan District in Shenzhen (Project No. KC2013ZDZJ0007A), the grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CityU 11300715), the grant from City University of Hong Kong (No. 7004220), the grant from Hong Kong General Research Grant (No. B-Q44D), and the grants from the Hong Kong Polytechnic University (No. G-YM05 and G-YN39).

References

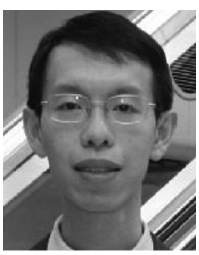
[1] A. Strehl and J. Ghosh, "Cluster ensembles — A knowledge reuse framework for combining multiple partitions," *J. Machine Learning Res.*, vol. 3, pp. 583–617, 2002.
[2] L. I. Kuncheva and D. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1798–1808, Nov. 2006.
[3] A. P. Topchy, A. K. Jain, and W. F. Punch, "Cluster ensembles: Models of consensus and weak partitions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1866–1881, Dec. 2005.
[4] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 16–173, Jan. 2008.
[5] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2396–2409, Dec. 2011.
[6] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based cluster ensemble approach for categorical data clustering," *IEEE Trans. Know. Data Eng.*, vol. 24, no. 3, pp. 413–425, Mar. 2012.
[7] Y. Yang and K. Chen, "Temporal data clustering via weighted clustering ensemble with different representations," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 2, pp. 307–320, Feb. 2011.
[8] Z. Yu, H. Chen, J. You, G. Han, and L. Li, "Hybrid fuzzy cluster ensemble framework for tumor clustering from bio-molecular data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 10, no. 3, pp. 657–670, May/Jun. 2013.
[9] Z. Yu, L. Li, J. You, and G. Han, "SC3: Triple spectral clustering based consensus clustering framework for class discovery from cancer gene expression profiles," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 9, no. 6, pp. 1751–1765, 2012.

[10] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, G. Han, and L. Li, "Adaptive fuzzy consensus clustering framework for clustering analysis of cancer data," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 12, no. 4, pp. 887–901, Jul./Aug. 2015.
[11] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.
[12] Z. Lu and Y. Peng, "Exhaustive and efficient constraint propagation: A graph-based learning approach and its applications," *Int. J. Comput. Vis.*, vol. 103, no. 3, pp. 306–325, 2013.
[13] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[14] M. Lichman, "UCI machine learning repository [http://archive.ics.uci.edu/ml]," Univ. California, School of Information and Computer Science, Irvine, CA, USA, 2013.
[15] S. Monti, P. Tamayo, J. Mesirov, and T. Glub, "Consensus clustering: A resampling based method for class discovery and visualization of gene expression microarray data," *J Mach. Learn. Res.*, vol. 52, nos. 1/2, pp. 91–118, Jul./Aug. 2003.
[16] A. L. N. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 835–850, Jun. 2005.
[17] C. Domeniconi and M. Alrazgan, "Weighted cluster ensembles: Methods and analysis," *ACM Trans. Knowl. Discovery Data*, vol. 2, no. 4, pp. 1–40, 2009.
[18] H.G. Ayad and M.S. Kamel, "On voting-based consensus of cluster ensembles," *Pattern Recog.*, vol. 43, no. 5, pp. 1943–1953, 2010.
[19] Z. Yu, H.-S. Wong, J. You, G. Yu, and G. Han, "Hybrid cluster ensemble framework based on the random combination of data transformation operators," *Pattern Recog.*, vol. 45, no. 5, pp. 1826–1837, 2012
[20] Z. Yu, J. You, H.-S. Wong, and G. Han, "From cluster ensemble to structure ensemble," *Inf. Sci.*, vol. 168, pp. 81–99, 2012.
[21] Z. Yu, L. Li, H.-S. Wong, J. You, G. Han, Y. Gao, and G. Yu, "Probabilistic cluster structure ensemble," *Inf. Sci.*, vol. 267, 16–34, 2014.
[22] T. Wang, "CA-tree: A hierarchical cluster for efficient and scalable coassociation-based cluster ensembles," *IEEE Trans. Syst., Man, Cybern.*, vol. 41, no. 3, pp. 686–698, Jun. 2011.
[23] A. P. Topchy, M. H. C. Law, A. K. Jain, and A. L. N. Fred, "Analysis of consensus partition in cluster ensemble,"in *Proc. Int. Conf. Data Mining*, 2004, pp. 225–232.
[24] P. Hore, L. O. Hall, and D. B. Goldgof, "A scalable framework for cluster ensembles," *Pattern Recog.*, vol. 42, no. 5, pp. 676–688, 2009.
[25] S. Zhang, H.-S. Wong, "ARImp: A generalized adjusted rand index for cluster ensembles,"in *Proc. 20th Int. Conf. Pattern Recog.*, 2010, pp. 778–781.
[26] Z. Yu, H.-S. Wong, H. Wang, "Graph-based consensus clustering for class discovery from gene expression data," *Bioinformatics*, vol. 23, pp. 2888–2896, 2007.
[27] Z. Yu and H.-S. Wong, "Class discovery from gene expression data based on perturbation and cluster ensemble," *IEEE Trans. NanoBioSci.*, vol. 8, no. 2, pp. 147–160, Jun. 2009.
[28] X. Zhang, L. Jiao, F. Liu, L. Bo, and M. Gong, "Spectral clustering ensemble applied to SAR image segmentation," *IEEE Trans. Geosci. Remote Sensing*, vol. 46, no. 7, pp. 2126–2136, Jul. 2008.
[29] N. Bassiou, V. Moschou, and C. Kotropoulos, "Speaker diarization exploiting the eigengap criterion and cluster ensembles," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2134–2144, Nov. 2010.
[30] W. Zhuang, Y. Ye, Y. Chen, and T. Li, "Ensemble clustering for internet security applications," *IEEE Trans. Syst., Man, Cybern., Part C: Appl. Rev.*, vol. 42, no. 6, pp. 1784–1796, Nov. 2012.
[31] X. Z. Fern and W. Lin, "Cluster ensemble selection," *Statist. Anal. Data Mining* , vol. 1, no. 3, pp. 787–797, 2008.
[32] J. Azimi and X. Fern, "Adaptive cluster ensemble selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 992–997.
[33] Z. Yu, L. Li, Y. Gao, J. You, J. Liu, H.-S. Wong, and G. Han, "Hybrid clustering solution selection strategy," *Pattern Recog.*, vol. 47, no. 10, pp. 3362–3375, 2014.
[34] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, and G. Han, "Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 4, pp. 1–14, Jul./Aug. 2014.
[35] E. Akbari, H.M. Dahlan, R. Ibrahim, and H. Alizadeh, "Hierarchical cluster ensemble selection," *Eng. Appl. Artif. Intell.*, vol. 39, pp. 146–156, 2015.
[36] H. Zeng and Y.-M. Cheung, "Semi-supervised maximum margin clustering with pairwise constraints," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 926–939, May 2012.
[37] S. Anand, S. Mittal, O. Tuzel, and P. Meer, "Semi-supervised kernel mean shift clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1201–1215, Jun. 2014.
[38] C.-L. Liu, W.-H. Hsaio, C.-H. Lee, and F.-S. Gou, "Semi-supervised linear discriminant clustering," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 989–1000, Jul. 2014.
[39] L. Zheng and T. Li, "Semi-supervised hierarchical clustering," in *Proc. IEEE 11th Int. Conf. Data Mining*, 2011, pp. 982–991.
[40] S. Xiong, J. Azimi, X. Z. Fern, "Active learning of constraints for semi-supervised clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 43–54, Jan. 2014.
[41] N. M. Arzeno and H. Vikalo, "Semi-supervised affinity propagation with soft instance-level constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1041–1052, May 2015.
[42] D. Wang, X. Gao, and X. Wang, "Semi-supervised nonnegative matrix factorization via constraint propagation," *IEEE Trans. Cybern.*, 2015, Doi: 10.1109/TCYB.2015.2399533.
[43] H. Liu, G. Yang, Z. Wu, and D. Cai, "Constrained concept factorization for image representation," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1214–1224, Jul. 2014.
[44] L.C. Jiao, F. Shang, F. Wang, and Y. Liu, "Fast semi-supervised clustering with enhanced spectral embedding," *Pattern Recog.*, vol. 45, no. 12, pp. 4358–4369, 2012.
[45] X. Yin, S. Chen, E. Hu, and D. Zhang, "Semi-supervised clustering with metric learning: An adaptive kernel method," *Pattern Recog.*, vol. 43, no. 4, pp. 1320–1333, 2010.
[46] H. Wang, T. Li, T. Li, and Y. Yang, "Constraint neighborhood projections for semi-supervised clustering," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 636–643, May 2014.
[47] Z. Yu, H.-S. Wong, J. You, Q. Yang, and H. Liao, "Knowledge based Cluster Ensemble for Cancer Discovery from Biomolecular Data", *IEEE Trans. NanoBioSci.*, vol. 10, no. 2, pp. 76–85, Jun. 2011.
[48] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, and G. Han, "Double selection based semi-supervised clustering ensemble for tumor clustering from gene expression profiles," *IEEE/ACM Trans. Comput. Biol. Bioinformat.*, vol. 11, no. 4, pp. 727–740, Jul./Aug. 2014.
[49] S. Basu, A. Banjeree, E. R. Mooney, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proc. 4th SIAM Int. Conf. Data Mining*, 2004, pp. 333–344.
[50] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: A comparative study," *BMC Bioinformatics*, vol. 9, no. 497, 2008.
[51] N.X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, 2010.
[52] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *J. Am. Statist. Assoc.*, vol. 32, no. 200, pp. 675–701, 1937.
[53] S. García and F. Herrera, "An extension on "Statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *J. Mach. Learn. Res.* vol. 9, pp. 2677–2694, 2008.
[54] X. Zhu and A. B. Goldberg, "Introduction to Semi-Supervised Learning," *Synthesis Lectures on Artif. Intell. Mach. Learn.*, Morgan & Claypool, 2009.
[55] D. Greene and P. Cunningham, "Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering,"in *Proc. Conf. Mach. Learn.*, 2007, pp. 140–151.
[56] B. Kulis, S. Basu, I. Dhillon, R. Mooney, "Semi-supervised graph clustering: A kernel approach," *Mach. Learn.*, vol. 74, no. 1, pp. 1-22, 2009.

Zhiwen Yu (S'06-M'08-SM'14) received the PhD degree from the City University of Hong Kong in 2008. He is a professor in the School of Computer Science and Engineering, South China University of Technology, and an adjunct professor at the Sun Yat-Sen University. His research areas focus on data mining, machine learning, bioinformatics, and pattern recognition. He has published more than 90 referred journal papers and international conference papers, including TEVC, TCYB, TMM, TCVST, *TCBB*, *TNB*, PR, IS, *Bioinformatics*, SIGKDD, and so on. He is a senior member of the IEEE, ACM, CCF (China Computer Federation), and CAAI (Chinese Association for Artificial Intelligence). Please refer to the homepage for more details: http://www.hgml.cn/yuzhiwen.htm.

Peinan Luo received the BSc degree from the South China University of Technology, China, and the MPhil degree from the School of Computer Science and Engineering, South China University of Technology. His research interests include pattern recognition, machine learning, and data mining.

Jane You is currently a professor in the Department of Computing, Hong Kong Polytechnic University, and the chair of the Department Research Committee. She has worked extensively in the fields of image processing, medical imaging, computer-aided diagnosis, and pattern recognition. So far, she has more than 190 research papers published with more than 1,000 non-self citations. She has been a principal investigator for one ITF project, three GRF projects, and many other joint grants since she joined PolyU in late 1998. She is also a team member for two successful patents (one HK patent and one US patent) and three awards including Hong Kong Government Industrial Awards. Her current work on retinal imaging has won a Special Prize and Gold Medal with Jury's Commendation at the 39th International Exhibition of Inventions of Geneva (April 2011), and the second place in an international competition (SPIE Medical Imaging'2009 Retinopathy Online Challenge (ROC'2009)). Her research output on retinal imaging has successfully led to technology transfer with clinical applications. She is also an associate editor of the *Pattern Recognition* and other journals.



Hau-San Wong received the BSc and MPhil degrees in electronic engineering from the Chinese University of Hong Kong, and the PhD degree in electrical and information engineering from the University of Sydney. He is currently an associate professor at the Department of Computer Science, City University of Hong Kong. He has also held research positions in the University of Sydney and Hong Kong Baptist University. His research interests include multimedia information processing, multimodal human-computer interaction, and machine learning.

Hareton Leung received the PhD degree in computer science from the University of Alberta. He is an associate professor and director of the Laboratory for Software Development and Management in the Department of Computing, the Hong Kong Polytechnic University. His research interests include software testing, project management, risk management, quality and process improvement, software metrics, and e-health.

Si Wu received the PhD degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in 2013. He is an associate professor in the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. His research interests include computer vision and pattern recognition.

Jun Zhang (M'02-SM'08) received the PhD degree in electrical engineering from the City University of Hong Kong, Kowloon, Hong Kong, in 2002. Since 2004, he has been with the Sun Yat-Sen University, Guangzhou, China, where he is currently a Cheung Kong professor. He has authored seven research books and book chapters, and over 100 technical papers in his research areas. His current research interests include computational intelligence, cloud computing, big data, and wireless sensor networks. He was a recipient of the China National Funds for Distinguished Young Scientists from the National Natural Science Foundation, China, in 2011, and the First-Grade Award in Natural Science Research from the Ministry of Education, China, in 2009. He is currently an associate editor of the *IEEE Transactions on Evolutionary Computation, the IEEE Transactions on Industrial Electronics, and the IEEE Transactions on Cybernetics*. He is the founding and current chair of the IEEE Guangzhou Subsection, and ACM Guangzhou Chapter. He is a senior member of the IEEE.

Guoqiang Han received the BSc degree from the Zhejiang University in 1982, and the master's and PhD degrees from the Sun Yat-Sen University, in 1985 and 1988, respectively. He is a professor in the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is the head of the School of Computer Science and Engineering, SCUT. His research interests include multimedia, computational intelligence, machine learning, and computer graphics. He has published over 100 research papers.