Cluster's Quality Evaluation and Selective Clustering Ensemble

FEIJIANG LI, YUHUA QIAN, and JIETING WANG, Shanxi University
CHUANGYIN DANG, City University of Hong Kong
BING LIU, University of Illinois at Chicago

Clustering ensemble has drawn much attention in recent years due to its ability to generate a high quality and robust partition result. Weighted clustering ensemble and selective clustering ensemble are two general ways to further improve the performance of a clustering ensemble method. Existing weighted clustering ensemble methods assign the same weight to each cluster in a partition of the ensemble. Since the qualities of the clusters in a partition are different, the clusters should be weighted differently. To address this issue, this article proposes a new measure to calculate the similarity between a cluster and a partition. Theoretically, this measure is effective in handling two problems in measuring the quality of a cluster, which are defined as the symmetric problem and the context meaning problem. In addition, some properties of the proposed measure are analyzed. This measure can be easily expanded to a clustering performance measure that calculates the similarity between two partitions. As a result of this measure, we propose a novel selective clustering ensemble framework, which considers the differences between the objective of the ensemble selection stage and the object of the ensemble integration stage in the selective clustering ensemble. To verify the performance of the new measure, we compare the performance of the measure with the two existing measures in weighting clusters. The experiments show that the proposed measure is more effective. To verify the performance of the novel framework, four existing state-of-the-art selective clustering ensemble frameworks are employed as references. The experiments show that the proposed framework is statistically better than the others on 17 UCI benchmark datasets, 8 document datasets, and the Olivetti Face Database.

CCS Concepts: • **Theory of computation** → *Unsupervised learning and clustering*; • **Computing methodologies** → *Ensemble methods*;

Additional Key Words and Phrases: Clustering ensemble, selective clustering ensemble, weighted clustering ensemble, cluster quality

This work was supported in part by the National Natural Science Fund of China (Nos. 61672332, 61322211, 61432011, and U1435212), Program for New Century Excellent Talents in University (No. NCET-12-1031), Program for the Outstanding Innovative Teams of Higher Learning Institutions of Shanxi, and Program for the Young San Jin Scholars of Shanxi. It was partially supported by CityU: 101113 of Hong Kong SAR Government.

Authors' addresses: F. Li, Y. Qian (corresponding author), and J. Wang are with the Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, Shanxi Province, China; emails: {feijiangli, jietingwang}@email.sxu.edu.cn; Y. Qian is also with the Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan, 030006 Shanxi, China; email: jinchengqyh@126.com; C. Dang, Department of Manufacture Engineering and Engineering Management, City University of Hong Kong, Hong Kong; email: mecdang@cityu.edu.hk; B. Liu, Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607; email: liub@cs.uic.edu.

https://doi.org/10.1145/3211872

ACM Reference format:
Feijiang Li, Yuhua Qian, Jieting Wang, Chuangyin Dang, and Bing Liu. 2018. Cluster's Quality Evaluation and Selective Clustering Ensemble. *ACM Trans. Knowl. Discov. Data.* 12, 5, Article 60 (June 2018), 27 pages.
https://doi.org/10.1145/3211872

1 INTRODUCTION

Clustering analysis plays an important role in machine learning. The goal of a clustering method is to discover a group structure of an unsupervised dataset. Without prior-knowledge about a dataset, different clustering methods generally generate different results. It is hard to judge which one is the best. To address this issue, one can integrate multiple clustering results to achieve a high quality and robust clustering result. This technique is called clustering ensemble. The integrated clustering is expected to obtain higher quality and robustness than the result of a single clustering algorithm. In the past decade, clustering ensemble has become a popular technique to deal with the data clustering problem. Due to the good performance and flexible processes, clustering ensemble has been applied in many areas in machine learning, such as document datasets learning (Xu et al. 2016), high dimensional data clustering (Fern and Brodley 2003; Jing et al. 2015; Li et al. 2013), streaming data clustering (Khan et al. 2016; Yang and Chen 2011), noisy data analysis (Yu et al. 2015), and imbalanced data analysis (Chen et al. 2010).

In a clustering ensemble problem, there are two major issues: ensemble generation and ensemble integration. As for the former, the generated multiple clustering results are often known as base clusterings or base partitions. The literature has declared that diversity and accuracy of the base partitions are important to the performance of the consensus clustering (Topchy et al. 2005). Many ensemble generation methods, which satisfy the requirements, have been proposed (Fischer and Buhmann 2003; Jing et al. 2015; Topchy et al. 2005; Wang et al. 2014; Yang et al. 2014). To generate different base clustering results, commonly used strategies include multiple parameter settings of one clustering algorithm, multiple clustering algorithms, and multiple data samplings or projections. For the latter, the literature focuses on designing effective clustering ensemble algorithms. A clustering ensemble algorithm generates a consensus clustering which is most similar to the base clustering results without accessing the original dataset. To design such an algorithm, many partition techniques haven been utilized, in which the set match techniques (Li et al. 2017; Zhou and Tang 2006), the graph partition techniques (Acharya et al. 2014; Fern and Brodley 2004; Huang et al. 2016; Strehl and Ghosh 2002; Zheng et al. 2014), and the clustering methods (Fred and Jain 2005; Gionis et al. 2007; Huang et al. 2015; Qian et al. 2016; Wu et al. 2015; Yu et al. 2015) are three widely used techniques. Obviously, a well-designed algorithm can generate a high quality integrated result. In addition, the Weighted Clustering Ensemble (WCE) (Li and Ding 2008; Yang and Chen 2011; Yousefnezhad and Zhang 2015) and Selective Clustering Ensemble (SCE) (Fern and Brodley 2003) have been proposed to improve the performance of clustering ensemble.

As for WCE, many approaches employ a clustering performance measure to weight each partition in the ensemble. The commonly used measures are Adjusted Rand Index (ARI) (Hubert and Arabie 1985) and Normalized Mutual Information (NMI) (Strehl and Ghosh 2002). These measures evaluate the similarity between two clustering results. Based on these measures, different weights are assigned to the base partitions, while the clusters in a partition share the same weight. However, for a clustering result, the characteristics of different clusters may be different. Therefore, the performance of the consensus clustering can be further improved if the quality of every cluster in the base clustering results is taken into consideration. Because the cluster's quality should reflect the context meaning of the cluster in the entire data, it should be measured based on both the cluster and the whole data points. Thus, the cluster's quality can be measured through comparing it with a reference clustering partition. Generally, this comparison is derived by transforming the cluster to a clustering form and applying a clustering similarity measure. Therefore, in the existing researches (Alizadeh et al. 2014; Law et al. 2004), this comparison between a cluster and a partition is called their similarity. Two measures that calculates the similarity between a cluster and a partition have been designed, which are Binary-NMI (BNMI) (Law et al. 2004) and Alizadeh–Parvin–Moshki–Minaei criterion (APMM) (Alizadeh et al. 2014). However, it has been shown in the literature that these two measures has their own problems (Yousefnezhad et al. 2018). In this article, we describe the two problems that exist in the two measures, which are called the symmetric problem and the context meaning problem. The two problems are caused in reconciling the information asymmetry between a cluster and a partition. To effectively solve these two problems, in this article, a new similarity measure between a cluster and a partition is proposed. We theoretically analyze the ability of the proposed measure in handling the two problems and in weighting clusters. This measure not only can be applied to weighting of each cluster, but also can be expanded to a similarity measure between two partitions. As a result of the proposed measure, we also attain a novel SCE framework.

In machine learning, feature selection is an effective approach to improve the learning performance (Blum and Langley 1997; Jain et al. 2000; Qian et al. 2010, 2015). Inspired by feature selection, SCE is proposed to improve the clustering ensemble performance. Research about the SCE mainly focuses on determination of the influence of diversity and quality to the performance of the ensemble result. Due to the unknown truth label, the quality of the base partitions set could not be evaluated directly. As a compromise, the quality of a set of partitions is represented by stability (Kuncheva and Vetrov 2006), which is the average of all pairwise similarity values between partitions. The diversity of an ensemble is often evaluated by the average dissimilarity between each pair of base partitions. It is easy to see that the diversity and stability are two opposite evaluation criteria. Prior researches about SCE explored whether diversity or stability is the determining factor to the selection of base clusterings. Recently, more researches tend to combine diversity and stability in the selection of a subset of special base partitions. These researches propose a measure that combines stability and diversity to guide the selection, or design a complex process that takes into consideration of both quality and diversity. The difficult choice between diversity and stability is mainly caused by the fact that the objective of ensemble selection and the objective of ensemble integration are different. It is well known that similar base partitions limit the improvement of the clustering ensemble performance; so, diversity is often employed as the selection criterion. However, since the ensemble integration is trying to generate a partition which is most similar with the base partitions, stability may be useful in the process of generating such a clustering. Based on the above discussion, this article introduces a novel SCE framework, which uses diversity to select base partitions and uses stability to weight the selected partitions.

Briefly, the contributions of this article are as follows:

- A measure which evaluates the similarity between a cluster and a partition is proposed. This measure is proved to be effectiveness in handling the symmetric problem and the context meaning problem in existing measures. In addition, this measure has some good properties in weighting clusters.
- A novel SCE framework is proposed. In this framework, diversity is used to select a subset of base partitions and stability is used to weight the importance of the selected partitions. This framework considers different objectives in different stages in the SCE process.
- Experiments are conducted to show the effectiveness of the proposed measure and the effectiveness of the proposed selective framework.

The rest of this article is organized as follows: In Section 2, we introduce the notations used in this article and the previous works about SCE. In Section 3, we describe existing similarity measures between a cluster and a partition, and discuss two problems in these measures. In Section 4, a novel measure that evaluates the similarity between a cluster and a partition is proposed, and its advantages are theoretically analyzed. In Section 5, we propose a SCE framework. In Section 6, experiments are conducted to show the effectiveness of the proposed measure in weighting clusters. In addition, the performance of the novel selective framework is evaluated by experiments in this section. Finally, this article is concluded in Section 7.

2 RELATED WORKS

Clustering ensemble technique solves a date clustering problem through combining multiple clustering results. Let $U = \{x_1, x_2, \ldots, x_n\}$ indicate a dataset with n samples. After the ensemble generation step, a set of base clustering results Π will be obtained, which can be expressed as $\Pi = \{\pi^1, \pi^2, \ldots, \pi^l\}$, where l is the ensemble size. Based on Π, a clustering ensemble method will generate a clustering result π^*, which is similar to each base partition. Without loss of generality, let F be a clustering ensemble method. Then, a clustering ensemble problem can be solved by $\pi^* = F(\Pi)$.

SCE is an effectiveness technique to improve the ensemble performance and reduce the computation cost of a clustering ensemble method. It improves the ensemble performance through improving the quality of base partition set. Given a set of base partitions, a SCE method generates the consensus result based on a subset of partitions which conform to the demands for the base clustering results. The researches about the SCE problem mainly try to explore an effective guidance for the selection of base partitions. Diversity and stability are two important factors in SCE. Given a set of base partitions $\Pi = \{\pi^1, \pi^2, \ldots, \pi^l\}$ and a clustering similarity measure sim, the stability s_i and diversity d_i of partition π^i are calculated by:

$$s_i = \frac{1}{l-1} \sum_{j=1, j \neq i}^{l} \mathrm{sim}(\pi^i, \pi^j), \tag{1}$$

$$d_i = \frac{1}{l-1} \sum_{j=1, j \neq i}^{l} \left(1 - \mathrm{sim}(\pi^i, \pi^j)\right). \tag{2}$$

Primely, the researches compared the influence of diversity and stability of the base partitions to the selective ensemble performance. In Fern and Brodley (2003), the author stated that low diversity limits the improvement of the performance, then high divers subset of partitions should be selected. Kuncheva Kuncheva and Hadjitodorov (2004) further developed the work in Fern and Brodley (2003) and suggested that the number of clusters in each base partition should be chosen randomly and should be larger than the expected number. In Hadjitodorov et al. (2006), the relationships between the diversity level and the ensemble accuracy were analyzed, the results show that a subset of partitions with median diversity obtains good performance. The diversity can be calculated by many measures, each of which is effective in specific cases. However, choosing a suitable measure is challenging. To handle this challenge, in Naldi et al. (2013), the author combined relative measures to select diverse partitions. In Kuncheva and Vetrov (2006), a new measure which combines the pairwise clustering similarity and the ensemble stability was proposed, and this measure has positive correlation with the ensemble accuracy. Azimi and Fern (2009) deemed that the selection of a subset partitions should based on diversity or stability is related to the characteristics of the base clusterings. Based on a diversity measure, the base partitions set can be labeled as stable or non-stable. In Azimi and Fern (2009), the author suggested using stability to select base partitions for a stable ensemble, and using diversity for a non-stable ensemble.

It is clear that diversity and stability are two opposite estimations. Recently, Many methods combine diversity and stability to select a subset of partitions. The most direct method is using a control parameter to balance diversity and stability (Hong et al. 2009). To combine diversity and stability, Fern and Lin (2008) proposed three selective clustering methods, which are called Joint Criterion, Cluster and Select, and Convex Hull, respectively. The Joint Criterion method optimizes a single aggregated objective function, which is a trade-off between diversity and stability. The Cluster and Select method runs a clustering algorithm on the base partitions and selects the partition which has the highest stability in each cluster. The Convex Hull method produces a stability–diversity scatter diagram and selects the partitions which correspond to the convex hull. In Jia et al. (2011), multiple referential partitions are generated through integrating multiple randomly selected base partitions, and the clustering results which are similar to the referential partitions are selected. In this process, multiple referential partitions guarantee the diversity, and the selected similar partitions guarantee the stability. Based on a set of partitions, Rastin and Kanawati (2015) builds a multiplex network, in which each community is obtained by diversity measure and the most stable partition is selected. Akbari et al. (2015) proposed a method called Hierarchical Cluster Ensemble Selection (HCES), which merges divers partitions into multiple groups by hierarchical algorithm and selects the most stable partition from each group to form the ensemble members.

Although a large number of SCE methods have been proposed, few of them investigate the influence of the characteristic of clusters to the ensemble performance. In this article, we propose to evaluate the cluster's quality and improve the ensemble performance through selecting and weighting clusters.

3 EXISTING SIMILARITY MEASURES BETWEEN A CLUSTER AND A PARTITION AND THEIR LIMITATIONS

A clustering result or a partition π^i consists of multi non-intersect clusters, which is $\pi^i = \{c_1^i, c_2^i, \ldots, c_{k_i}^i\}$, where k_i is the number of clusters in π^i. There is no doubt that the qualities of clusters in a clustering result are different. To measure the quality of a cluster, a similarity measure between a cluster and a partition is needed. In this section, the existing measures for estimating the similarity between a cluster and a partition are reviewed. There are two related pieces of research, both of which extend the NMI. The two existing measures are BNMI and APMM. In addition, two problems of the existing measures are defined. In the following of this article, we use $|\bullet|$ to indicate the number of samples in $\bullet$, which can be a cluster, a partition, or a set of clusters.

3.1 The BNMI

One challenge in measuring the similarity between a cluster and a partition is the asymmetric information between them, i.e., the sample size of the cluster and that of the partition are different. To handle this challenge, in Law et al. (2004), the authors treated the cluster c as a two group partition $\pi_c = \{c, U/c\}$ and transforms the reference partition π into a two group partition $\pi_g = \{c_g, U/c_g\}$, where c_g is the set of samples in the clusters which correspond to π_c in π. A cluster is corresponding to another one if their common samples are more than half the number of the samples in the measured cluster. Thus, c_g can be defined by:

$$c_g = \left\{x|x \in c_i^\pi, |c_i^\pi \cap c| > \frac{1}{2}|c_i^\pi|, i = 1, \ldots k_\pi\right\}.$$

Then, the similarity between a cluster c and a partition π can be calculated by $\mathrm{NMI}(\pi_c, \pi_g)$. This measure is noted as BNMI because it can be treated as a binary type of NMI. The NMI is

Fig. 1. Examples of a cluster and two partitions.

Fig. 2. Examples of a cluster and four partitions.

an normalized version of mutual information. The mutual information quantifies the information shared between two partitions. The NMI between two partitions π^b and π^d is calculated by:

$$\mathrm{NMI}(\pi^b, \pi^d) = \frac{\sum_{i=1}^{k_b} \sum_{j=1}^{k_d} n_{ij} \log\left(\frac{n n_{ij}}{|c_i^b||c_j^d|}\right)}{\sqrt{\left(\sum_{i=1}^{k_b} |c_i^b| \log\left(\frac{|c_i^b|}{n}\right)\right)\left(\sum_{j=1}^{k_d} |c_j^d| \log\left(\frac{|c_j^d|}{n}\right)\right)}}, \tag{3}$$

where n_{ij} is the number of shared samples of c_i^b and c_j^d.

The BNMI is calculated by:

$$\mathrm{BNMI}(c, \pi) = \mathrm{NMI}(\pi_c, \pi_g). \tag{4}$$

3.2 The APMM

To handle the asymmetric information challenge, in Alizadeh et al. (2014), the authors proposed a measure called APMM. The APMM measures the similarity between a cluster c and its corresponding sub-partition π_p in π. Specifically, π_p is the partition result of samples in c induced by π, which can be defined by:

$$\pi_p = \{c' | c' = c_i^\pi \cap c, i = 1, \ldots, k_\pi\}.$$

The APMM is defined as:

$$\mathrm{APMM}(c, \pi) = \mathrm{APMM}(c, \pi_p) = \frac{-2|c| \log\left(\frac{n}{|c|}\right)}{|c| \log\left(\frac{|c|}{n}\right) + \sum_{i=1}^{k_p} |c_i^p| \log\left(\frac{|c_i^p|}{n}\right)}. \tag{5}$$

3.3 Two Problems in the Existing Measures

BNMI and APMM have their drawbacks in special situations. To preliminarily show their drawbacks, we employ two set of examples which are shown in Figure 1 and Figure 2.

In Figure 1, a cluster c_1 and two partitions π^1 and π^2 are listed. Based on BNMI, the following results will be obtained:

$$\text{BNMI}(c_1, \pi^1) = \text{BNMI}(c_1, \pi^2) = 1.$$

It is obvious that the two partitions π^1 and π^2 are different, especially the partition results of the samples in cluster c_1. However, the BNMI generates the same similarity values on these two comparisons.

Figure 2 shows a cluster and four partitions. In Figure 2, it is obvious that the four partitions are different. However, with Formula (5), we obtain the following results:

$$\text{APMM}(c_2, \pi^3) = \text{APMM}(c_2, \pi^4),$$
$$\text{APMM}(c_2, \pi^5) = \text{APMM}(c_2, \pi^6).$$

The problem in Figure 1 is called the symmetric problem and the problem in Figure 2 is called the context meaning problem. To give the definitions of the two problems, we first introduce two sub-partitions based on a cluster c and a partition π, which are corresponding partition CP_c^π and extended partition EP_c^π.

Definition 3.1 (Corresponding Partition CP_c^π). Given a cluster c and a partition π, the corresponding partition CP_c^π is the partition result of samples in c induced by π, which is formulated as:

$$\text{CP}_c^\pi = \{\text{cp}_i^\pi | \text{cp}_i^\pi = c_i^\pi \cap c, c_i^\pi \cap c \neq \emptyset, i = 1, \ldots, k_\pi\},$$

where c_i^π is the ith cluster in partition π, and k_π is the number of clusters in π.

Definition 3.2 (Extended Partition EP_c^π). Given a cluster c and a partition π, the extended partition EP_c^π is the union of the clusters in π which have nonempty intersection with c. EP_c^π is:

$$\text{EP}_c^\pi = \{\text{ep}_i^\pi | \text{ep}_i^\pi = c_i^\pi, c_i^\pi \cap c \neq \emptyset, i = 1, \ldots, k_\pi\},$$

where c_i^π is the ith cluster in partition π, and k_π is the number of clusters in π.

It is obvious that the cluster indices in CP_c^π and EP_c^π are the same. For the convenience of indicating the clusters in CP_c^π and EP_c^π, we define the set of their cluster indices as K_c^π, which is:

$$K_c^\pi = \{i | c_i^\pi \cap c \neq \emptyset, i = 1, \ldots, k_\pi\}.$$

Obviously, the set of samples in CP_c^π are the same as the set of samples in c, which is:

$$\text{SCP}_c^\pi = \text{SC} = \{x | x \in c\}.$$

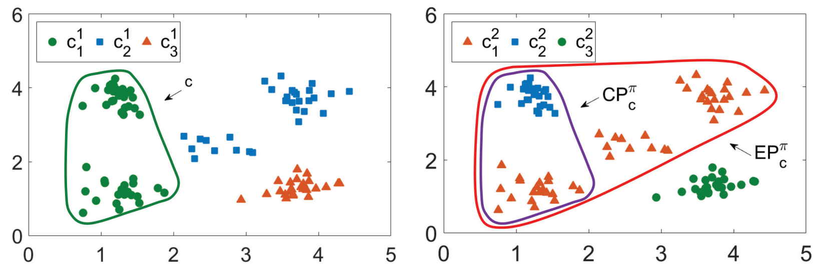
The set of samples in EP_c^π is:

$$\text{SEP}_c^\pi = \{x | x \in \text{ep}_i^\pi, \text{ep}_i^\pi \in P_c^\pi, i \in K_c^\pi\}.$$

To clearly show the corresponding partition CP_c^π and the extended partition EP_c^π, we employ an example with two partitions, which is shown in Figure 3. Without loss of generality, we assume that the measured cluster is $c = c_1^1$, which is the green area in Figure 3(a), and the measured partition is the partition π in Figure 3(b). Based on c and π, Figure 3(b) shows the corresponding partition CP_c^π and the extended partition EP_c^π.

With the above definitions, the symmetric problem and the context meaning problem can be described as follows:

— Problem 1 (The Symmetric Problem). *Given a cluster c and two partitions π^b and π^d, the symmetric problem is that when $CP_c^b \neq CP_c^d$ and $SEP_c^b = SEP_c^d = SC$, a similarity measure sim generates the result with $sim(c, \pi^b) = sim(c, \pi^d)$.*



(a) Example of a cluster c (b) Example of a partition π

Fig. 3. Examples of two partitions.

The condition $\text{CP}_c^b \neq \text{CP}_c^d$ indicates that the partitions π^b and π^d are different, then the similarity values $\text{sim}(c, \pi^b)$ and $\text{sim}(c, \pi^d)$ should be different.

— Problem 2 (The Context Meaning Problem). *Given a cluster c and two partitions π^b and π^d, the context meaning problem is that when $EP_c^b \neq EP_c^d$ and $CP_c^b = CP_c^d$, a similarity measure sim generates the result with $sim(c, \pi^b) = sim(c, \pi^d)$.*

Similarly, $\text{EP}_c^b \neq \text{EP}_c^d$ indicates that $\pi^b \neq \pi^d$. In this situation, an effective similarity measure should generates different similarity values, i.e., $\text{sim}(c, \pi^b) \neq \text{sim}(c, \pi^d)$.

Based on the definition of BNMI and APMM, the following two facts can be obtained:

— Fact 1 *BNMI has the symmetric problem.*

Proof. Following from the definition of BNMI and that of the symmetric problem, one has:

$$\text{SEP}_c^b = \text{SC} \Rightarrow |\text{ep}_i^b \cap c| = |\text{ep}_i^b|, i \in K_c^b.$$

Then, $c_g^b = \bigcup_{i \in K_c^b} \text{ep}_i^b = c$ and $\pi_c = \pi_g^b$. With Formula (4), one has $\text{BNMI}(c, \pi_b) = \text{NMI}(\pi_c, \pi_g^b) = 1$.

In the same way, $\text{BNMI}(c, \pi_d) = 1$. That is $\text{BNMI}(c, \pi_b) = \text{BNMI}(c, \pi_d) = 1$, which means the BNMI has the symmetric problem. □

— Fact 2 *APMM has the context meaning problem.*

Proof. Following from Definition 1 and the definition of APMM, one has:

$$\text{CP}_c^b = \text{CP}_c^d \Rightarrow \pi_p^b = \pi_p^d.$$

Then, $\text{APMM}(c, \pi^b) = \text{APMM}(c, \pi^d)$, which means APMM has the context meaning problem. □

To mitigate the above two problems, we propose a novel measure to calculate the similarity between a cluster and a partition.

4 A NEW SIMILARITY MEASURE BETWEEN A CLUSTER AND A PARTITION

From the discussions in Section 3.3, it is obvious that both the two cluster goodness measures BNMI and APMM are extended types of the NMI measure. The NMI measure requires equal size of the compared partitions. To satisfy this requirement, the BNMI measure transforms the compared partition into a binary type, which causes the symmetric problem, while the APMM extracts the sub-partition corresponding to the compared cluster from the partition, which causes the context meaning problem. It can be concluded that the NMI may be unsuitable for measuring the similarity between a cluster and a partition. In this section, to calculate the similarity between a cluster and a partition, we introduce a novel measure, which is based on matching degree evaluation. Following that, we show the advantages of the measure in handling the symmetric problem and the context meaning problem, and then we analyze some properties of the measure in weighting clusters.

4.1 The Measure Based on Set Matching Degree Evaluation

The new measure calculates the similarity between a cluster and a partition through evaluating the set matching degree between them. We use SME to represent the measure in the following of the article.

To clearly show how SME calculates the similarity between a cluster and a partition, we employ the examples in Figure 3. Our goal is to measure the similarity between the cluster c and the partition π, which is expressed as $\mathrm{SME}(c,\pi)$. The proposed $\mathrm{SME}(c,\pi)$ is composed of two parts, which are the similarity between cluster c and the corresponding partition CP_c^π and the similarity between CP_c^π and the extended partition EP_c^π.

Following the definition of CP_c^π, the CP_c^π can be treated as a partition result of c. The requirement of CP_c^π is not breaking up the cluster c. Therefore, there should exist a main cluster in CP_c^π. The quality of CP_c^π can be measured by the cardinality of the main cluster. Thus, the similarity between c and CP_c^π is calculated by

$$\mathrm{sim}(c,\mathrm{CP}_c^\pi) = \max_{i\in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|}, \tag{6}$$

where cp_i is the ith cluster in the corresponding partition CP_c^π.

The comparison between CP_c^π and EP_c^π should reflect the quality of treating CP_c^π as a cluster. Through this comparison, the context meaning of CP_c^π will be taken into consideration. The similarity between CP_c^π and EP_c^π is calculated by:

$$\mathrm{sim}(\mathrm{CP}_c^\pi,\mathrm{EP}_c^\pi) = \sum_{i\in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|}\frac{|\mathrm{cp}_i|}{|ep_i|}, \tag{7}$$

In Formula (7), $\frac{|\mathrm{cp}_i|}{|\mathrm{ep}_i|}$ calculates the fraction of discovered samples in a cluster, and $\frac{|\mathrm{cp}_i|}{|c|}$ weights the influence of each cluster in CP_c^π.

Combining Formula (6) and Formula (7), the similarity between a cluster c and a partition π is calculated by:

$$\mathrm{SME}(c,\pi) = \max_{i\in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|} \cdot \sum_{i\in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|}\frac{|\mathrm{cp}_i|}{|\mathrm{ep}_i|}. \tag{8}$$

4.2 The Advantages of SME in Handling the Two Problems

To show the advantages of SME in handling the symmetric problem and the context meaning problem, we first calculate the similarity values of the examples in Figure 1 and Figure 2 with Formula (8). For the examples in Figure 1, the results are:

$$\mathrm{SME}(c_1,\pi^1) = \frac{1}{2}, \quad \mathrm{SME}(c_1,\pi^2) = \frac{3}{4}.$$

The partition π^2, which contains an obvious main cluster, obtains a greater similarity value. These results accord with our anticipate.

Property 1. *If* $\max_{i\in K_c^b} \frac{|cp_i^b|}{|c|} \neq \max_{j\in K_c^d} \frac{|cp_j^d|}{|c|}$, *the SME has no symmetric problem.*

Proof. From the definition of the symmetric problem, we have $\mathrm{SEP}_c^b = \mathrm{SEP}_c^d = \mathrm{SC}$. From this condition, it follows that $\mathrm{CP}_c^b = \mathrm{EP}_c^b$ and $\mathrm{CP}_c^d = \mathrm{EP}_c^d$. Then,

$$\sum_{i \in K_c^b} \frac{|\mathrm{cp}_i^b|}{|c|} \frac{|\mathrm{cp}_i^b|}{|\mathrm{ep}_i^b|} = 1,$$

$$\sum_{i \in K_c^d} \frac{|\mathrm{cp}_i^d|}{|c|} \frac{|\mathrm{cp}_i^d|}{|\mathrm{ep}_i^d|} = 1.$$

Therefore,

$$\mathrm{SME}(c, \pi^b) = \max_{i \in K_c^b} \frac{|\mathrm{cp}_i^b|}{|c|},$$

and

$$\mathrm{SME}(c, \pi^d) = \max_{j \in K_c^d} \frac{|\mathrm{cp}_j^d|}{|c|}.$$

With the condition $\max_{i \in K_c^b} \frac{|\mathrm{cp}_i^b|}{|c|} \neq \max_{j \in K_c^d} \frac{|\mathrm{cp}_j^d|}{|c|}$, it holds true that $\mathrm{SME}(c, \pi^b) \neq \mathrm{SME}(c, \pi^d)$, which means that the SME is able to handle the symmetric problem in this situation. □

The results of the examples in Figure 2 are listed as follows:

$$\mathrm{SME}(c_2, \pi^3) = \frac{1}{2}, \quad \mathrm{SME}(c_2, \pi^4) = \frac{1}{3}, \quad \mathrm{SME}(c_2, \pi^5) = \frac{2}{3}, \quad \mathrm{SME}(c_2, \pi^6) = 1.$$

These results show the ability of SME in mitigating the context meaning problem. Comparing with the APMM values on these examples, it is easy to see that APMM generates two groups of equal values on the four different partitions, while the SME reflects these differences. This advantage comes from that SME takes into consideration the expended area.

Property 2. *If the vectors*

$$X_1 = \left[\frac{|\mathrm{cp}_1^b|}{|c|}, \frac{|\mathrm{cp}_i^b|}{|c|}, \ldots \frac{|\mathrm{cp}_{k_b}^b|}{|c|} \right],$$

$$X_2 = \left[\frac{|\mathrm{cp}_1^b|}{|ep_1^b|}, \frac{|\mathrm{cp}_2^b|}{|ep_2^b|}, \ldots \frac{|\mathrm{cp}_{k_b}^b|}{|ep_{k_b}^b|} \right],$$

$$Y_1 = \left[\frac{|\mathrm{cp}_1^d|}{|c|}, \frac{|\mathrm{cp}_i^d|}{|c|}, \ldots \frac{|\mathrm{cp}_{k_d}^d|}{|c|} \right],$$

and

$$Y_2 = \left[\frac{|\mathrm{cp}_1^d|}{|ep_1^d|}, \frac{|\mathrm{cp}_2^d|}{|ep_2^d|}, \ldots \frac{|\mathrm{cp}_{k_d}^d|}{|ep_{k_d}^d|} \right]$$

satisfy $X_1 X_2^\top \neq Y_1 Y_2^\top$*; the* SME *has no context meaning problem.*

Proof. From the definition of the context meaning problem, one has $\mathrm{CP}_c^b = \mathrm{CP}_c^d$. Then,

$$\max_{i \in K_c^b} \frac{|\mathrm{cp}_i^b|}{|c|} = \max_{j \in K_c^d} \frac{|\mathrm{cp}_j^d|}{|c|}.$$

With $\mathrm{CP}_c^b = \mathrm{CP}_c^d$, it can be obtained that the indices set of the clusters in π^b and π^d are the same, i.e., $K_c^b = K_c^d$. For convenience, we let $K = K_c^b = K_c^d$. Then,

$$\frac{|\mathrm{cp}_i^b|}{|c|} = \frac{|\mathrm{cp}_i^d|}{|c|} = \frac{|\mathrm{cp}_i|}{|c|},$$

where $i \in K$.

We let $Z = [\frac{|\mathrm{cp}_1|}{|c|}, \frac{|\mathrm{cp}_i|}{|c|}, \ldots \frac{|\mathrm{cp}_k|}{|c|}]$. Then, $X_1 = Y_1 = Z$. Thus,

$$\mathrm{SME}(c, \pi_b) = \max_{i \in K} \frac{|\mathrm{cp}_i^b|}{|c|} \cdot ZX_2^\top,$$

$$\mathrm{SME}(c, \pi_d) = \max_{i \in K} \frac{|\mathrm{cp}_j^d|}{|c|} \cdot ZY_2^\top.$$

With the condition $X_1 X_2^\top \neq Y_1 Y_2^\top$, $\mathrm{SME}(c, \pi^b) \neq \mathrm{SME}(c, \pi^d)$ will hold, which means that the SME is able to handle the context meaning problem in this situation. □

4.3 Analysis of SME

In this section, several intuitive tendencies in comparing a cluster and a partition are discussed, and the corresponding performances of SME are analyzed.

The estimation of similarity between a cluster and a partition can be treated as measuring the preserved consistency of the partition when treat the corresponding samples in the measured cluster as a group. Without correction for chance, the preserved consistency should be larger than zero. From this consideration, the region of the estimation should be $(0, 1]$.

Property 3. *The range of* SME *is* $(0, 1]$, *and* $\mathrm{SME}(c, \pi) = 1$ *if and only if the cluster c is a cluster in the partition* π.

Proof. Following from the definition of CP_c^π and EP_c^π, it is easy to obtain that $|\mathrm{cp}_i| > 0$ and $|\mathrm{ep}_i| > 0$. Then, $\max_{i \in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|} > 0$, and $\sum_{i \in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|} \frac{|\mathrm{cp}_i|}{|ep_i|} > 0$. With Formula (8), SME > 0 will hold.

From Formula (8), it can be seen that the two parts of SME are no greater than 1. Then, SME will obtain value 1 if and only if both of the two parts of SME get value 1. Considering the former part of SME, $\max_{i \in K_c^\pi} \frac{|\mathrm{cp}_i|}{|c|} = 1$ will hold if and only if $|\mathrm{cp}_i| = |c|$, which means $\mathrm{CP}_c^\pi = c$. In this situation, the later part of SME will be 1 if and only if $|\mathrm{CP}_c^\pi| = |\mathrm{EP}_c^\pi|$. Then, it can be concluded that SME takes on value 1 only when $\mathrm{CP}_c^\pi = \mathrm{EP}_c^\pi = c$, which means that the measured cluster corresponds to a cluster in the measured partition. □

In the following, we discuss the influence of the size of the expanded group and the size of the measured cluster to the similarity value.

Intuitively, for a partition, if the size of the expanded partition is similar to that of the measured cluster, this partition tend to be similar to the cluster. SME defers to this tendency.

Property 4. *For a cluster c and two partitions* π^b *and* π^d, *if* $\mathrm{CP}_c^b = \mathrm{CP}_c^d$ *and* $|\mathrm{ep}_i^b| < |\mathrm{ep}_i^d|$, $i \in K$, *where* $K = K_c^b = K_c^d$, *then* $\mathrm{SME}(c, \pi^b) > SME(c, \pi^d)$.

Proof. The condition $\mathrm{CP}_c^b = \mathrm{CP}_c^d$ indicates that $\mathrm{cp}_i^b = \mathrm{cp}_i^d$, where $i \in K$. Then,

$$\max_{i \in K} \frac{|\mathrm{cp}_i^b|}{|c|} = \max_{j \in K} \frac{|\mathrm{cp}_j^d|}{|c|},$$

and

$$\frac{|\mathrm{cp}_i^b|}{|c|} = \frac{|\mathrm{cp}_i^d|}{|c|},$$

where $i \in K$.

With the condition $|\mathrm{ep}_i^b| < |\mathrm{ep}_i^d|$, one has

$$\frac{|\mathrm{cp}_i^b|}{|\mathrm{ep}_i^b|} > \frac{|\mathrm{cp}_i^d|}{|\mathrm{ep}_i^d|},$$

where $i \in K$.

Based on the above results and Formula (8), $\mathrm{SME}(c, \pi^b) > \mathrm{SME}(c, \pi^d)$ holds. □

In Property 4, $\mathrm{CP}_c^b = \mathrm{CP}_c^d$ and $|\mathrm{ep}_i^b| < |\mathrm{ep}_i^d|$ indicate that the partition π_b has smaller extend partition than partition π_d in the same situation. The result $\mathrm{SME}(c, \pi^b) > \mathrm{SME}(c, \pi^d)$ indicates that the partition which has smaller extend area has higher SME value.

As for the size of the measured cluster, if the expand partitions of two clusters are the same, the cluster which contains more samples should obtain greater similarity value than the other cluster. That is to say, a bigger cluster should obtain higher similarity value when the other situations are the same.

Property 5. *For a partition π and two clusters c_b and c_d, if $EP_b^\pi = EP_d^\pi$ and $\frac{|c_b|}{|c_d|} = \frac{|\mathrm{cp}_i^b|}{|\mathrm{cp}_i^d|} > 1$, $i \in K$, where $K = K_c^b = K_c^d$, then* $\mathrm{SME}(c, \pi_b) > \mathrm{SME}(c, \pi_d)$.

Proof. Following from the condition that $\frac{|c_b|}{|c_d|} = \frac{|\mathrm{cp}_i^b|}{|\mathrm{cp}_i^d|} > 1$, one has

$$\max_{i \in K} \frac{|\mathrm{cp}_i^b|}{|c_b|} = \max_{j \in K} \frac{|\mathrm{cp}_j^d|}{|c_d|},$$

and

$$\frac{|\mathrm{cp}_i^b|}{|c_b|} = \frac{|\mathrm{cp}_i^d|}{|c_d|},$$

where $i \in K$.

Based on the condition $\mathrm{EP}_b^\pi = \mathrm{EP}_d^\pi$ and $\frac{|\mathrm{cp}_i^b|}{|\mathrm{cp}_i^d|} > 1$, one has

$$\frac{|\mathrm{cp}_i^b|}{|\mathrm{ep}_i^b|} > \frac{|\mathrm{cp}_i^d|}{|\mathrm{ep}_i^d|},$$

where $i \in K$.

With the above results and Formula (8), $\mathrm{SME}(c, \pi^b) > \mathrm{SME}(c, \pi^d)$ holds. □

From the above discussions, the proposed SME conforms to the intuitive demands for the measure between a cluster and a partition. Therefore, the proposed SME may be suitable for weighting the quality of the clusters in a set of partitions.

4.4 Using SME to Measure the Similarity Between Two Partitions

Another advantage of SME is that it is easy to be extended to measuring the similarity between two partitions, which is notated as *SMEP*. Suppose the two partitions to be measured are $\pi^b = \{c_1^b, c_2^b, \ldots, c_{k_b}^b\}$ and $\pi^d = \{c_1^d, c_2^d, \ldots, c_{k_d}^d\}$. Referring to the partition π^d, the quality of all the clusters in π^b can be measured by SME. The quality of partition π^b can be reflected by the average quality of its clusters. In the same way, the quality of partition π^d can be measured. The similarity between the two partitions can be quantized by their average quality. Thus, based on SME, the similarity between the partitions π^b and π^d can be calculated as follows:

$$\text{SMEP}(\pi^b, \pi^d) = \frac{1}{2}\left(\frac{1}{k_b}\sum_{i=1}^{k_b}\text{SME}(c_i^b, \pi^d) + \frac{1}{k_d}\sum_{j=1}^{k_d}\text{SME}(c_j^d, \pi^b)\right).$$

As a result of SME, in what follows, we propose a novel SCE framework. This framework combines diversity and stability. As an improvement, the proposed framework meets the different demands in the selecting step and integrating step in the SCE process.

5 A NEW SELECTIVE CLUSTERING ENSEMBLE FRAMEWORK

It has been commonly agreed that both diversity and stability of the base partitions are important to the performance of a clustering ensemble algorithm. Then, both the factors should be utilized in the process of a SCE algorithm. The main challenge of using both diversity and stability in a single selective algorithm is that these two factors are conflicting. To handle this challenge, most of the existing algorithms are very complicated. In this section, we propose a novel SCE framework which responds to this challenge in a simple way.

The requirement of both diversity and stability comes from the fact that diverse base partitions are important in improving the ensemble performance, while the final objective is to discover a stable partition. To meet this requirement, the utilization of diversity and stability can be separated in different stages in the processes of SCE. Generally, a SCE problem is solved in two stages, which are ensemble selection and ensemble integration. Meanwhile, the fundamental objectives of these two stages are quite different. In the ensemble selection stage, the objective is to select diverse base partitions. Thus, in this stage, the diversity should play the most important role. In the ensemble stage, the objective is to discover a partition which shares the most information with the base partitions. That is, the discovered partition can be treated as a stable partition from the view of the base partitions. Therefore, the stability is important in this stage. Based on the above considerations, the proposed framework utilizes diversity to select diverse base partitions and utilizes stability to weight the selected partitions. We call this framework DS for short. The DS framework takes three factors as input − a similarity measure Sim measuring the diversity of each partition by Formula (2), a threshold t_s selecting a set of base partitions, and a clustering ensemble method F generating a consensus partition $\pi^* = F(\Pi)$.

In what follows, we embed the SME into the DS framework to form the DSME method. In Section 3.2, the SME is extended to SMEP, which can be utilized to measure the similarity between two partitions. Then, SMEP can be employed as the similarity measure in the DS framework to select a subset of base partitions.

It should be noted that the weights for base partitions treat the clusters in a partition result equally. However, due to the complex data distribution, the qualities of different clusters in a partition could be different. The measure which quantifies the similarity between two partitions cannot reflect this difference. The advantage of utilizing SME in the weighting process is that each single cluster in the base partitions can be weighted. These weights offer the confidence that two samples are in the same cluster at the cluster level. With a set of cluster weights, a weighted refined cluster matrix (WRA) and a weighted co-association matrix (WCO) can be obtained, which will improve the performance of clustering ensemble methods based on these matrices. The refined cluster matrix ($\text{RA}^{(n\times h)}$) is a binary matrix, i.e., $\text{RA} \in \{0,1\}^{(n\times h)}$. The RA-matrix indicates every cluster in the base partitions. Each column in a RA-matrix corresponds to a cluster, in which the entries will be 1 if the corresponding samples belong to the cluster and the entries will be 0 otherwise. With a set of cluster weights, the weighted RA-matrix wRA is easy to obtain. The co-association matrix

($\mathrm{CO}^{(n\times n)}$) reflects the relation between all pairs of samples, in which each element is the frequency that two samples appear in the same cluster. The weighted co-association matrix $w\mathrm{CO}^{(n\times n)}$ can be obtained based on wRA. With the weights of clusters, the effectiveness of a clustering ensemble method based on the RA or CO could be improved through introducing the wRA or wCO.

The detailed process of the DSME is shown in Algorithm 1. The time complexity of DSME contains three parts, which are selecting base partitions, weighting base partitions, and combining base partitions. The time complexity of selecting base partitions and weighting base partitions are $O(h^2)$, where h is the total number of clusters in the ensemble, i.e., $h = k_1 + k_2 + \cdots + k_l$. Combining base partitions has the same time complexity as the utilized clustering ensemble method F, which is noted as $O(T_F)$. The time complexity of a clustering ensemble method based on RA-matrix is at least $O(nh)$. The time complexity of a clustering ensemble method based on CO-matrix is at least $O(n^2)$. Then the total time complexity of DSME is $O(2h^2 + T_F)$.

ALGORITHM 1: DSME

INPUT: Base partitions $\Pi = \{\pi^1, \pi^2, \ldots, \pi^l\}$,
a selection threshold t_s,
a consensus function F
OUTPUT: A consensus clustering $\pi*$

$l' = 1, \Pi' = \emptyset$
for $i = 1$ to l **do**
  $D_i = \frac{1}{l-1} \sum_{j=1, j\neq i}^{l} (1 - SMEP(\pi^i, \pi^j))$
  if $D_i > t_s$ **then**
    $\pi^{l'} \leftarrow \pi^i, \Pi' = \Pi' \cup \pi^j, l' = l' + 1$
  end if
end for
The selected partitions $\Pi' = \{\pi^{1'}, \pi^{2'}, \ldots \pi^{l'}\}$
$h' = h_{1'} + h_{2'} + \cdots + h_{l'}$
for $i = 1$ to h' **do**
  $S_i = S_p(\pi_i) = \frac{1}{l'-1} \sum_{j=1, j\neq i}^{l'} SME(c_i, \pi^j)$
end for
for $i = 1$ to h' **do**
  $w_i = \frac{S_i}{\sum_{i=1}^{l'} S_i}$
end for
$W = \{w_1, w_2, \ldots, w_{l'}\}$
$\pi* = F(\Pi', W)$

6 EXPERIMENTAL ANALYSIS

In this section, we verify the effectiveness of SME in weighting clusters, the rationality of using SMEP to measure the similarity between two partitions, and the ensemble performance of DSME.

6.1 Datasets

To conduct the experimental analyses, we use 2 groups of datasets, which are 17 real datasets and 8 document datasets. The details of these datasets are outlined in Table 2. In Table 2, n is the data size, a is the number of attributes of a dataset, k is the truth number of clusters in a dataset. The 17 real datasets come from the UCI Machine Learning Repository (UCI, http://archive.ics.uci.edu/ml/). It has been widely accepted that a large variety of used datasets validate better obtained results. Thus, to effectively validate the obtained results, the datasets were chosen in such a way that they

Table 1. The Twenty-five Benchmark Datasets

Number	Datasets	n	Number of attributes (a)	Number of clusters (k)
1	Iris	150	4	3
2	Wine	178	13	3
3	Seeds	210	7	3
4	Glass	214	9	7
5	Protein Localization Sites	272	7	3
6	Ecoli	336	7	8
7	LIBRAS Movement Database	360	91	15
8	User Knowledge Modeling	403	5	4
9	Vote	435	16	2
10	Wisconsin Diagnostic Breast Cancer	569	30	2
11	Synthetic Control Chart Time Series	600	60	6
12	Student	600	5	3
13	Australian Credit Approval	690	14	2
14	Cardiotocography	2126	40	10
15	Wave form Database Generator	5000	21	3
16	Parkinsons Telemonitoring	5875	21	42
17	Statlog Landsat Satellite	6435	36	6
18	Tr12	313	5804	8
19	Tr11	414	6428	9
20	Tr45	690	8261	10
21	Tr41	878	7454	10
22	Tr31	927	10128	7
23	Wap	1560	8460	20
24	Hitech	2301	126321	6
25	Fbis	2463	2000	17

have high diversity in the number of samples, attributes, and true clusters. The samples of these datasets range from 150 to 6435. The attribute number of these data ranges from 4 to 91. The cluster number of these data range from 2 to 41. The 8 benchmark document datasets come with the CLUTO clustering toolkit (Steinbach et al. 2000). These documents datasets are represented using the TF-IDF vector-space model. These datasets are sparse and have much more attributes than the 17 UCI datasets.

6.2 Evaluation Indices

To evaluate the performance of an ensemble result, we employ three widely used indices, which are Accuracy (AC) (Yang 1999), ARI and Normalized Mutual Information (NMI). These three indices are external indices, which evaluate the performance of an ensemble result through estimating the similarity between the result with a reference partition. In the experiments, we treat the truth partition of each datasets as the reference partition. These three indices can be calculated based on the overlap matrix between two partitions. Suppose the two partitions are π^b and π^d, the overlap matrix is shown in Table 2. Here, we only consider a special case, that is, $k_b = k_d = k$. In Table 1, n_{ij} is the number of common samples of cluster c_i^b from partition π^b and cluster c_j^d from partition

Table 2. The Overlap Matrix Between π^b and π^d

$\pi^b \backslash \pi^d$	c_1^d	c_2^d	$\cdots$	c_k^d	Sums
c_1^b	n_{11}	n_{12}	$\cdots$	n_{1k}	n_{1*}
c_2^b	n_{21}	n_{22}	$\cdots$	n_{2k}	n_{2*}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
c_k^b	n_{k1}	n_{k2}	$\cdots$	n_{kk}	n_{k*}
Sums	n_{*1}	n_{*2}	$\cdots$	n_{*k}	n

π^d, n_{i*} is the number of samples in cluster c_i^b from clustering π^b, and n_{*j} is the number of samples in cluster c_j^d from clustering π^d.

The AC index is a set based measure, which matches the clusters in the compared partitions and calculates the fraction of the common samples. Based on Table 2, the AC is calculated by:

$$AC(\pi^b, \pi^d) = \sum_{i=1}^{k} \frac{\max\{n_{ij} : j \le k\}}{n}. \tag{9}$$

The ARI is an form of Rand Index (RI) corrected for chance (Hubert and Arabie 1985). The RI calculates the fraction that two partitions have the same decision on sample pairs. The ARI is calculated by:

$$ARI(\pi^b, \pi^d) = \frac{t_0 - t_3}{\frac{1}{2}(t_1 + r_2) - t_3}, \tag{10}$$

where

$$t_0 = \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2}, \quad t_1 = \sum_{i=1}^{k}\binom{n_{i*}}{2}, \quad t_2 = \sum_{j=1}^{k}\binom{n_{*j}}{2}, \quad t_3 = \frac{2t_1t_2}{n(n-1)}.$$

The NMI is calculated by Formula (3).

6.3 Performance Analysis of SME in Weighting Clusters

To demonstrate the effectiveness of SME in weighting clusters, we compare SME with other two state-of-the-art cluster quality evaluation measures BNMI and APMM, which are introduced in Section 3.1 and Section 3.2. In this experiment, four clustering ensemble algorithms are utilized to integrate the weighted base partitions, which are WCT (Iam-On et al. 2011), WTQ (Iam-On et al. 2011), CSPA (Strehl and Ghosh 2002), and EAC (Fred and Jain 2005). The time complexity of the four methods are $O(nh + hm^2)$ (WCT), $O(nh + h^2m^2)$ (WTQ), $O(n^2lk)$ (CSPA), $O(n^2l)$ (EAC), where m represents the average number of neighbors connecting to one cluster. The first two algorithms are based on the RA-matrix, while the remaining two algorithms are based on the CO-matrix. As introduced in Section 5, the four algorithms are easy to be expanded to a weighted type. Through comparing the ensemble performance, we evaluate the effectiveness of the three measures in weighting clusters.

To eliminate the influence caused by the quality of base partitions, for each dataset, we generate 50 sets of base partitions, and report the average index values of AC, ARI, and NMI, respectively. All the base partitions are generated by k-means algorithm. In detail, the Euclidean distance is used for the UCI real datasets and the cosine similarity is used for the document datasets. As for the number of clusters in the base partitions, literature (Kuncheva and Hadjitodorov 2004) suggests that the base partitions should contain more clusters than the expectation. Then, for the the UCI real datasets, the number of clusters is randomly selected from $[2, \sqrt{n}]$, where n is the number of samples in the corresponding datasets. As for the document datasets, we follow the setting by Xu et al. (2016), and set the number of clusters in each base partition equal to the excepted number of clusters in the final ensemble result.

The results of the three indices are shown in Tables 3–5, respectively. In Tables 3–5, the maximum value in each comparison is underlined. To statistically analyze the experimental results, we conduct independent two-sample t-test with 90% confidence level. For each comparison, we test the top two maximum index values. If the top two maximum values are significantly different from each other, we assign a bullet behind the maximum value, which indicates the corresponding method is statistically better. In the last line of Tables 3—5, we report the times that a method is better and statistically better than the other methods. It can be seen from these tables that the four employed clustering ensemble algorithms based on SME obtain the most marks from the three evaluation indices. For each index and each weighted ensemble method, SME obtains much more higher values and bullets than both the BNMI and the APMM on the twenty-five datasets. The results indicate that the SME is statistically and significantly better than the BNMI and the APMM in weighting base clusters. With these results, it can be concluded that SME is more effective in measuring the similarity between a cluster and a partition.

6.4 Correlation Analysis of SMEP in Measuring Partitions

To demonstrate the rationality of SMEP in measuring the similarity between two partitions, we test the correlation between SMEP, ARI, and NMI. To conduct this test, we construct three variables from a set of partitions based on the three indices, and calculate the correlation values between each pair of variables by Pearson correlation coefficient (Reshef et al. 2011).

The base partitions are generated based on the twenty-five datasets. For each dataset, 50 partitions are generated. Totally, we obtain $L = 1250$ base partitions. With this partition set, a variable $X = \{x_1, x_2, \ldots, x_{\binom{L}{2}}\}$ can be constructed based on SMEP, where x_i is the similarity value of the ith pair of partitions. Based on another index, a variable $Y = \{y_1, y_2, \ldots, y_{\binom{L}{2}}\}$ will be constructed. The Pearson correlation coefficient $\rho(X, Y)$ between variables X and Y is calculated by

$$\rho(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}, \tag{11}$$

where cov is the covariance of X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively.

The correlation analysis results are shown in Figure 4. In Figure 4, each axis indicates the value of similarity between two partitions measured by the corresponding index, and each point is index value between two partitions. From Figure 4, it is obvious that the SMEP is strongly associated with ARI and NMI. The Pearson correlation value is 0.9104 between SMEP and ARI, and is 0.8323 between SMEP and NMI. The correlation value between the widely used measures ARI and NMI is 0.8606. This correlation analysis result show that SMEP has very similar performance to ARI and NMI. It can be concluded that SMEP can be utilized to measure the similarity between two partitions.

6.5 Performance Analysis of DSME in Integrating Clusterings

In order to verify the performance of DSME, we compare DSME with four representative SCE methods, which are adaptive cluster ensemble selection method (AD) (Azimi and Fern 2009), joint method (JO) (Fern and Lin 2008), cluster and select (CAE) (Fern and Lin 2008), and bagging based method (BA) (Jia et al. 2011). Concerning computational time complexity, the proposed

Table 3. Index AC from the Compared Weighted Methods

Data	WCT			WTQ		
	BNMI	APMM	SME	BNMI	APMM	SME
1	0.8093±0.1149	0.7740±0.1539	0.8247±0.1219	0.7440±0.1480	0.8140±0.1324	0.8127±0.1367
2	0.8876±0.1574	0.9287±0.1130	0.9646±0.0026 •	0.9247±0.1154	0.9191±0.1304	0.9663±0.0000 •
3	0.7905±0.1279	0.7405±0.1546	0.8119±0.1431	0.8386±0.1073	0.8062±0.1369	0.8148±0.1201
4	0.5089±0.0352	0.5061±0.0404	0.5047±0.0404	0.4771±0.0214	0.5182±0.0346	0.5224±0.0357
5	0.9132±0.0777	0.8754±0.1168	0.9368±0.0022 •	0.8868±0.1062	0.8610±0.1250	0.9096±0.0949
6	0.5063±0.0608	0.5485±0.0578	0.6060±0.0699 •	0.5387±0.0887	0.5685±0.0939	0.5842±0.1060
7	0.4250±0.0263	0.4167±0.0260	0.4319±0.0183 •	0.4114±0.0235	0.3933±0.0259	0.4147±0.0240
8	0.5360±0.0535	0.5340±0.0221	0.5362±0.0452	0.5164±0.0496	0.5067±0.0584	0.5459±0.0492 •
9	0.8632±0.0093	0.8600±0.0059	0.8616±0.0055	0.8528±0.0158	0.8437±0.0122	0.8497±0.0097
10	0.9339±0.0050	0.9237±0.0077	0.9336±0.0065	0.9329±0.0048	0.9213±0.0351	0.9336±0.0068
11	0.6587±0.0904	0.6853±0.1128	0.6672±0.0588	0.7055±0.1139	0.6998±0.1015	0.7062±0.0809
12	0.5373±0.1123	0.5585±0.0970	0.6138±0.1494 •	0.5412±0.1274	0.5620±0.1415	0.5907±0.1065
13	0.7625±0.1115	0.7658±0.1237	0.8084±0.0767 •	0.7146±0.1670	0.7546±0.1202	0.7561±0.1076
14	0.6854±0.1276	0.7193±0.0740	0.7030±0.1145	0.6720±0.0701	0.7165±0.0788	0.7254±0.0941
15	0.5479±0.1687	0.6453±0.1139	0.6663±0.1085	0.3874±0.0019	0.3961±0.0087	0.4435±0.1326 •
16	0.4208±0.0191	0.4095±0.0176	0.4213±0.0232	0.3972±0.0253 •	0.3923±0.0098	0.3405±0.0149
17	0.4208±0.0191 •	0.4095±0.0176	0.4113±0.0232	0.3972±0.0253	0.3923±0.0098	0.3995±0.0149
18	0.6054±0.0515	0.6169±0.0467	0.6217±0.0416	0.6019±0.0640	0.5920±0.0391	0.6351±0.0179 •
19	0.5684±0.0488	0.5756±0.0389	0.5903±0.0831	0.5440±0.0369	0.5580±0.0548	0.5928±0.0844 •
20	0.5541±0.0496	0.5407±0.0567	0.5871±0.0410 •	0.5249±0.0406	0.5049±0.0548	0.5419±0.0420 •
21	0.5401±0.0580	0.5574±0.0380	0.5866±0.0302 •	0.5825±0.0345	0.5694±0.0490	0.5682±0.0521
22	0.5239±0.0491	0.4752±0.0649	0.5570±0.0357 •	0.4874±0.0717	0.4912±0.0604	0.5282±0.0435 •
23	0.4445±0.0379	0.4785±0.0543	0.4806±0.0447	0.4358±0.0422	0.4286±0.0496	0.4605±0.0254 •
24	0.5111±0.0240	0.4970±0.0235	0.5121±0.0332	0.4932±0.0259	0.4814±0.0202	0.4962±0.0353
25	0.5135±0.0245	0.5623±0.0262 •	0.5400±0.0225	0.4831±0.0273	0.4978±0.0372	0.5005±0.0301
w-sw	4-1	3-1	18-9	4-1	1-0	20-8

Data	CSPA			EAC		
	BNMI	APMM	SME	BNMI	APMM	SME
1	0.7640±0.1349	0.7967±0.1117	0.8227±0.1079	0.8573±0.0140	0.8553±0.0108	0.8680±0.0136 •
2	0.9197±0.1287	0.7663±0.1921	0.9264±0.1159	0.9596±0.0075	0.9652±0.0055	0.9657±0.0047
3	0.8181±0.0995	0.8152±0.1284	0.8005±0.1070	0.8986±0.0122	0.8986±0.0074	0.9033±0.0154
4	0.4874±0.0292	0.4930±0.0307	0.4706±0.0203	0.5285±0.0014	0.5355±0.0067	0.5350±0.0070
5	0.8081±0.1438	0.8456±0.1610	0.9382±0.0040 •	0.9342±0.0031	0.9338±0.0084	0.9364±0.0057 •
6	0.5146±0.0470	0.5432±0.0797	0.5399±0.0616	0.5658±0.0492	0.5583±0.0585	0.5935±0.0474 •
7	0.3731±0.0419	0.3783±0.0464	0.4167±0.0176 •	0.4297±0.0095	0.4303±0.0034	0.4344±0.0065 •
8	0.4603±0.0560	0.4253±0.0205	0.4643±0.0537	0.5447±0.0517	0.5280±0.0470	0.5449±0.0311
9	0.8508±0.0140	0.8595±0.0131	0.8607±0.0068	0.8274±0.0125	0.8315±0.0307	0.8322±0.0308
10	0.7868±0.1819	0.8403±0.0656	0.8323±0.1003	0.8949±0.0327	0.8895±0.0295	0.8996±0.0308
11	0.6417±0.0704	0.7028±0.0554	0.7070±0.0936	0.7165±0.0616	0.7352±0.0747	0.7535±0.0791
12	0.6432±0.1596	0.5857±0.1199	0.6613±0.0463 •	0.4507±0.0809	0.4823±0.1219	0.5133±0.1228
13	0.7016±0.1068	0.6484±0.1088	0.6675±0.1230	0.8433±0.0147	0.8201±0.0471	0.8383±0.0220
14	0.6111±0.0560	0.6206±0.0977	0.6622±0.1140 •	0.8646±0.0803	0.8534±0.0669	0.8862±0.0479 •
15	0.5406±0.0799	0.4954±0.1027	0.5346±0.0711	0.6682±0.0857	0.5866±0.1402	0.6852±0.0683
16	0.3704±0.0240	0.3730±0.0214	0.3828±0.0152 •	0.4211±0.0093	0.4112±0.0096	0.4276±0.0128 •
17	0.3704±0.0240	0.3730±0.0214	0.3728±0.0152	0.4211±0.0093	0.4112±0.0096	0.4276±0.0128 •
18	0.6160±0.0704	0.5815±0.0765	0.6409±0.0134 •	0.6288±0.0147	0.6409±0.0142 •	0.6358±0.0126
19	0.5372±0.0664	0.4889±0.0607	0.5418±0.0620	0.6075±0.0375	0.6027±0.0231	0.6118±0.0236
20	0.4874±0.0452	0.5433±0.0487	0.5729±0.0340 •	0.6035±0.0287	0.6158±0.0069	0.6154±0.0063
21	0.5649±0.0371	0.5282±0.0617	0.5732±0.0230 •	0.5757±0.0485	0.5523±0.0354	0.5859±0.0414
22	0.5357±0.0495	0.4710±0.0585	0.5258±0.0309	0.5498±0.0258	0.5753±0.0229	0.5768±0.0240
23	0.4068±0.0377	0.3961±0.0346	0.4488±0.0305 •	0.4491±0.0119 •	0.4460±0.0149	0.4461±0.0089
24	0.5003±0.0197 •	0.4901±0.0220	0.4850±0.0207	0.5252±0.0242	0.5276±0.0172	0.5363±0.0173 •
25	0.4712±0.0453	0.4981±0.0182	0.4934±0.0176	0.5657±0.0237	0.5841±0.0088	0.5843±0.0132
w-sw	5-1	4-1	16-9	2-1	2-1	21-8

Table 4. Index ARI from the Compared Weighted Methods

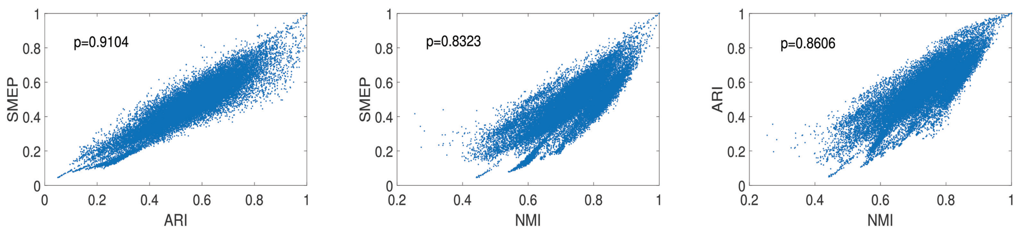
	WCT			WTQ		
Data	BNMI	APMM	SME	BNMI	APMM	SME
1	0.6415±0.0864	0.6256±0.1317	0.6646±0.1192	0.5922±0.1086	0.6596±0.1220	0.6597±0.1247
2	0.8007±0.1916	0.8513±0.1362	0.8914±0.0077 ●	0.8417±0.1386	0.8372±0.1456	0.8965±0.0001 ●
3	0.5928±0.1394	0.5649±0.1559	0.6446±0.1546 ●	0.6491±0.1141	0.6220±0.1296	0.6212±0.1315
4	0.2475±0.0252	0.2480±0.0285	0.2416±0.0338	0.2162±0.0151	0.2434±0.0248	0.2455±0.0224
5	0.7885±0.1150	0.7175±0.1876	0.8174±0.0070 ●	0.7494±0.1555	0.7184±0.1804	0.7821±0.1465
6	0.3769±0.0541	0.3963±0.0657	0.4526±0.1260 ●	0.4493±0.1103	0.4326±0.1101	0.4642±0.1419
7	0.3108±0.0341	0.3047±0.0384	0.3171±0.0229	0.3103±0.0278	0.2903±0.0298	0.3128±0.0321
8	0.2760±0.0733	0.2814±0.0531	0.2856±0.0383	0.2793±0.0629	0.2471±0.0664	0.3051±0.0340 ●
9	0.5266±0.0267	0.5172±0.0167	0.5177±0.0290	0.4963±0.0466	0.4719±0.0340	0.4883±0.0270
10	0.7516±0.0176 ●	0.7165±0.0261	0.7401±0.0226	0.7472±0.0170	0.7110±0.1096	0.7499±0.0240
11	0.6088±0.0795	0.6281±0.0599	0.5757±0.0638	0.6275±0.1001	0.6356±0.0883	0.6601±0.0523 ●
12	0.2585±0.1609	0.2184±0.0915	0.3241±0.1974	0.2656±0.1802	0.2722±0.1558	0.2888±0.1026
13	0.3240±0.1879	0.3428±0.2052	0.4029±0.1550 ●	0.2944±0.2364	0.3156±0.1743	0.3068±0.2000
14	0.6938±0.1550	0.7069±0.1152	0.6990±0.1196	0.6550±0.0782	0.6735±0.1012	0.6624±0.1357
15	0.3084±0.0691	0.3323±0.0632	0.3411±0.0895	0.2556±0.0005	0.2616±0.0074	0.2721±0.0375
15	0.3523±0.0108	0.3440±0.0233	0.3632±0.0152 ●	0.3395±0.0167	0.3297±0.0136	0.3476±0.0168 ●
17	0.3523±0.0108	0.3440±0.0233	0.3532±0.0152	0.3395±0.0167	0.3297±0.0136	0.3376±0.0168
18	0.4113±0.0475	0.4214±0.0359	0.4326±0.0242 ●	0.3958±0.0461	0.3759±0.0534	0.4316±0.0362 ●
19	0.4779±0.0513	0.4714±0.0358	0.5115±0.0936 ●	0.4304±0.0519	0.4755±0.0955	0.5172±0.1071 ●
20	0.3856±0.0542	0.3911±0.0525	0.4288±0.0421 ●	0.3617±0.0483	0.3572±0.0548	0.3819±0.0431 ●
21	0.3749±0.0668	0.3926±0.0562	0.4354±0.0276 ●	0.4199±0.0492	0.4160±0.0539	0.4083±0.0689
22	0.3837±0.0837	0.3154±0.1002	0.4563±0.0684 ●	0.3367±0.1130	0.3340±0.0958	0.3920±0.0758 ●
22	0.3391±0.0575	0.3832±0.1042	0.3665±0.0993	0.3053±0.1115	0.3258±0.1013	0.3123±0.0507
24	0.2910±0.0220	0.2806±0.0212	0.2936±0.0269	0.2705±0.0285	0.2620±0.0270	0.2554±0.0331
25	0.3807±0.0302	0.4196±0.0227 ●	0.3926±0.0287	0.3506±0.0213	0.3611±0.0268	0.3606±0.0301
w-sw	2-1	5-1	18-11	5-0	4-0	16-8

	WCT			WTQ		
Data	BNMI	APMM	SME	BNMI	APMM	SME
1	0.5867±0.1110	0.5898±0.1476	0.6484±0.0802 ●	0.6706±0.0218	0.6671±0.0162	0.6876±0.0223 ●
2	0.8375±0.1435	0.6408±0.2293	0.8430±0.1495	0.8765±0.0219	0.8931±0.0165	0.8948±0.0143
3	0.5984±0.1299	0.5818±0.1624	0.5745±0.1407	0.7318±0.0277	0.7311±0.0152	0.7424±0.0339
4	0.2339±0.0218	0.2395±0.0239	0.2187±0.0196	0.2584±0.0016	0.2646±0.0058	0.2640±0.0062
5	0.6403±0.1995	0.7055±0.1784	0.8232±0.0101 ●	0.8135±0.0120	0.8119±0.0229	0.8203±0.0170 ●
6	0.3375±0.0525	0.3837±0.1011	0.3681±0.0807	0.4129±0.0265	0.4081±0.0351	0.4775±0.0529 ●
7	0.2217±0.0547	0.2222±0.0638	0.2927±0.0283 ●	0.3185±0.0084	0.3166±0.0074	0.3220±0.0101 ●
8	0.1756±0.0776	0.1433±0.0498	0.1688±0.1085	0.2667±0.0424	0.2689±0.0489	0.2702±0.0463
9	0.4909±0.0377	0.5158±0.0359	0.5182±0.0199	0.4276±0.0333	0.4417±0.0822	0.4437±0.0830
10	0.4510±0.3061	0.4808±0.1783	0.4758±0.2536	0.6240±0.1014	0.6063±0.0897	0.6390±0.0954
11	0.6036±0.0895	0.6229±0.0765	0.5983±0.0801	0.6591±0.0315	0.6534±0.0322	0.6672±0.0391
12	0.3090±0.2567	0.2957±0.1664	0.2947±0.1643	0.1570±0.0042	0.2151±0.1195	0.2500±0.1434
13	0.2059±0.1671	0.1312±0.1466	0.1687±0.1603	0.4715±0.0400	0.4177±0.1043	0.4587±0.0574
14	0.5042±0.0436	0.5559±0.1604	0.5664±0.1479	0.8463±0.0851	0.8324±0.0687	0.8599±0.0536
15	0.2329±0.0639	0.2347±0.0721	0.2300±0.0856	0.3645±0.0679	0.3247±0.0620	0.3764±0.0491
16	0.2025±0.0485	0.2139±0.0613	0.2202±0.0463	0.3605±0.0057	0.3566±0.0075	0.3685±0.0086 ●
17	0.2025±0.0485	0.2139±0.0613	0.2102±0.0463	0.3605±0.0057	0.3566±0.0075	0.3665±0.0086 ●
18	0.4057±0.0593	0.3912±0.0840	0.4411±0.0270 ●	0.4368±0.0175	0.4414±0.0212	0.4434±0.0187
19	0.4291±0.0642	0.3628±0.0731	0.4450±0.0903	0.5193±0.0344	0.5094±0.0283	0.5254±0.0220
20	0.3594±0.0447	0.3926±0.0561	0.4181±0.0430 ●	0.4411±0.0391	0.4625±0.0090	0.4591±0.0120
21	0.4101±0.0475	0.3678±0.0753	0.4190±0.0283	0.4326±0.0454	0.4011±0.0337	0.4357±0.0422
22	0.4034±0.0867 ●	0.2935±0.0915	0.3857±0.0506	0.4270±0.0502	0.4732±0.0418	0.4771±0.0451
23	0.3188±0.0744	0.2868±0.0579	0.3546±0.0666 ●	0.3101±0.0230 ●	0.3071±0.0186	0.2949±0.0080
24	0.2753±0.0230	0.2717±0.0250	0.2685±0.0245	0.2956±0.0191	0.2983±0.0095	0.2974±0.0180
25	0.3334±0.0462	0.3520±0.0248	0.3538±0.0213	0.4081±0.0215	0.4245±0.0117	0.4280±0.0122
w-sw	6-1	6-0	13-6	2-1	3-0	20-6

Table 5. Index NMI from the Compared Weighted Methods

	WCT			WTQ		
Data	BNMI	APMM	SME	BNMI	APMM	SME
1	0.7425±0.0382	0.7316±0.0623	0.7419±0.0768	0.7188±0.0469	0.7525±0.0595	0.7524±0.0633
2	0.8060±0.1175	0.8394±0.0768	0.8646±0.0143	0.8390±0.0852	0.8293±0.0834	0.8754±0.0041 •
3	0.6391±0.0730	0.6234±0.0768	0.6657±0.0776 •	0.6689±0.0580	0.6566±0.0687	0.6512±0.0650
4	0.3906±0.0296	0.4085±0.0331 •	0.3936±0.0399	0.3701±0.0183	0.3998±0.0247	0.4050±0.0251
5	0.7304±0.0670	0.6772±0.1203	0.7483±0.0074 •	0.7091±0.0924	0.6936±0.1073	0.7279±0.0981
6	0.5783±0.0332	0.5835±0.0273	0.6024±0.0446 •	0.5871±0.0299	0.5844±0.0295	0.6010±0.0512
7	0.5974±0.0284	0.5883±0.0302	0.6043±0.0203	0.5973±0.0214	0.5817±0.0246	0.6064±0.0215 •
8	0.3626±0.0800	0.3689±0.0687	0.3753±0.0465	0.3556±0.0673	0.3395±0.0680	0.3925±0.0386 •
9	0.4410±0.0197	0.4399±0.0266	0.4562±0.0150 •	0.4601±0.0328 •	0.4350±0.0235	0.4502±0.0239
10	0.6363±0.0207 •	0.5972±0.0287	0.6225±0.0256	0.6512±0.0224	0.6316±0.0811	0.6533±0.0331
11	0.7750±0.0449	0.7716±0.0327	0.7581±0.0424	0.7867±0.0552	0.7922±0.0522	0.8071±0.0318 •
12	0.4037±0.1594	0.3692±0.0858	0.4545±0.1741	0.4088±0.1680	0.4262±0.1525	0.4328±0.1041
13	0.2729±0.1528	0.2896±0.1749	0.3408±0.1248 •	0.2437±0.1932	0.2589±0.1443	0.2543±0.1571
14	0.8437±0.0682	0.8616±0.0437	0.8591±0.0520	0.8272±0.0385	0.8515±0.0363	0.8497±0.0574
15	0.3950±0.0319	0.4084±0.0326	0.4096±0.0508	0.3720±0.0007	0.3755±0.0041	0.3767±0.0133
16	0.6953±0.0056	0.6914±0.0106	0.7062±0.0082 •	0.6923±0.0073	0.6885±0.0060	0.7094±0.0058 •
17	0.6953±0.0056	0.6914±0.0106	0.6962±0.0082	0.6923±0.0073 •	0.6885±0.0060	0.6894±0.0058
18	0.5860±0.0291	0.5886±0.0256	0.5941±0.0177	0.5934±0.0325	0.5641±0.0253	0.6031±0.0198 •
19	0.6260±0.0212	0.6137±0.0234	0.6271±0.0324	0.6048±0.0259	0.6196±0.0181	0.6388±0.0392 •
20	0.5254±0.0251	0.5286±0.0186	0.5435±0.0194 •	0.5021±0.0292	0.4982±0.0347	0.5439±0.0237 •
21	0.5457±0.0409	0.5532±0.0270	0.5755±0.0240 •	0.5831±0.0275	0.5747±0.0249	0.5782±0.0338
22	0.3930±0.0354	0.3891±0.0241	0.4226±0.0232 •	0.4042±0.0327	0.4028±0.0226	0.4104±0.0204
23	0.5827±0.0133	0.5841±0.0144	0.5876±0.0148	0.5708±0.0184	0.5764±0.0201	0.5793±0.0145
24	0.3398±0.0087	0.3406±0.0062	0.3426±0.0095	0.3500±0.0069	0.3494±0.0078	0.3470±0.0085
25	0.5812±0.0089	0.5888±0.0088	0.5857±0.0119	0.5704±0.0128	0.5734±0.0161	0.5773±0.0154
w-sw	3-1	3-1	19-9	5-2	3-0	17-8

	WCT			WTQ		
Data	BNMI	APMM	SME	BNMI	APMM	SME
1	0.6919±0.0783	0.6924±0.1225	0.7348±0.0449 •	0.7582±0.0112	0.7563±0.0083	0.7670±0.0117 •
2	0.8286±0.0789	0.7076±0.1292	0.8376±0.0917	0.8446±0.0289	0.8590±0.0217	0.8603±0.0178
3	0.6343±0.0774	0.6202±0.1000	0.6094±0.1015	0.7184±0.0180	0.7140±0.0069	0.7218±0.0173
4	0.3944±0.0181	0.3903±0.0211	0.3823±0.0197	0.3995±0.0023	0.4148±0.0147	0.4140±0.0153
5	0.6363±0.1296	0.6739±0.1115	0.7516±0.0114 •	0.7463±0.0099	0.7399±0.0223	0.7481±0.0149
6	0.5486±0.0375	0.5566±0.0483	0.5511±0.0428	0.6096±0.0113	0.6089±0.0163	0.6392±0.0247 •
7	0.5390±0.0310	0.5430±0.0430	0.5819±0.0199 •	0.6092±0.0078	0.6068±0.0064	0.6111±0.0087
8	0.2733±0.0684	0.2641±0.0444	0.2731±0.0770	0.3449±0.0550	0.3576±0.0597	0.3452±0.0559
9	0.3952±0.0106	0.4192±0.0249 •	0.4021±0.0192	0.4030±0.0179	0.4061±0.0496	0.4254±0.0487 •
10	0.4379±0.2004	0.4539±0.1451	0.4455±0.1600	0.5236±0.0769	0.5001±0.0584	0.5325±0.0611
11	0.7725±0.0571	0.7797±0.0488	0.7572±0.0467	0.8091±0.0155 •	0.7861±0.0048	0.7987±0.0174
12	0.4597±0.2167	0.4429±0.1559	0.4485±0.0726	0.3006±0.0045	0.3501±0.1012	0.3789±0.1196
13	0.1972±0.1301	0.1463±0.1199	0.1832±0.1225	0.3973±0.0400	0.3468±0.0885	0.3787±0.0607
14	0.7731±0.0194	0.7830±0.0699	0.8074±0.0765	0.9392±0.0350	0.9341±0.0290	0.9459±0.0217
15	0.3352±0.0690	0.3283±0.0541	0.3264±0.0707	0.4249±0.0333	0.4092±0.0335	0.4327±0.0282
16	0.6396±0.0188	0.6385±0.0240	0.6480±0.0165 •	0.7044±0.0032	0.6989±0.0032	0.7133±0.0038 •
17	0.6396±0.0188	0.6385±0.0240	0.6380±0.0165	0.7044±0.0032	0.6989±0.0032	0.7033±0.0038
18	0.5768±0.0448	0.5673±0.0451	0.6005±0.0125 •	0.6060±0.0130	0.6030±0.0141	0.6062±0.0129
19	0.6027±0.0221	0.5688±0.0382	0.6044±0.0349	0.6331±0.0169	0.6183±0.0226	0.6342±0.0097
20	0.5101±0.0258	0.5151±0.0323	0.5510±0.0066 •	0.5622±0.0132	0.5623±0.0076	0.5611±0.0082
21	0.5582±0.0293	0.5260±0.0440	0.5666±0.0167 •	0.5807±0.0282	0.5507±0.0215	0.5753±0.0254
22	0.4038±0.0206	0.3906±0.0354	0.4074±0.0158	0.4139±0.0099	0.4122±0.0068	0.4137±0.0101
23	0.5486±0.0137	0.5413±0.0207	0.5676±0.0147 •	0.5909±0.0059 •	0.5861±0.0050	0.5834±0.0016
24	0.3358±0.0084	0.3344±0.0106	0.3305±0.0063	0.3444±0.0098	0.3432±0.0084	0.3435±0.0058
25	0.5544±0.0226	0.5678±0.0119	0.5704±0.0104	0.5975±0.0069	0.5972±0.0072	0.6069±0.0063 •
w-sw	8-0	4-1	13-8	7-2	3-0	15-5



(a) SMEP and ARI (b) SMEP and NMI (c) ARI and NMI

Fig. 4. Correlation analysis between SMEP, NMI and ARI.

Table 6. Index AC from the Five Selective Methods

Data sets	AD	JO	CAE	BA	DSME
1	0.8587±0.0013	0.8600±0.0000	0.8560±0.0045	0.8620±0.0062	0.8613±0.0183
2	0.9607±0.0019	0.9640±0.0010	0.9517±0.0026	0.9573±0.0010	0.9663±0.0014 ●
3	0.8943±0.0028	0.8967±0.0041	0.8981±0.0030	0.9057±0.0030	0.9062±0.0050
4	0.4304±0.0023	0.4322±0.0018	0.4313±0.0034	0.4248±0.0031	0.4481±0.0019 ●
5	0.9294±0.0030	0.9287±0.0032	0.9353±0.0014	0.9353±0.0024	0.9370±0.0006 ●
6	0.5530±0.0144	0.5616±0.0118	0.5932±0.0091	0.5813±0.0228	0.6095±0.0297 ●
7	0.4356±0.0028	0.4322±0.0022	0.4117±0.0050	0.4217±0.0029	0.4347±0.0043
8	0.5511±0.0182	0.5462±0.0094	0.5315±0.0102	0.5360±0.0139	0.5653±0.0083 ●
9	0.8163±0.0217	0.8448±0.0055	0.8278±0.0060	0.8138±0.0221	0.8634±0.0020 ●
10	0.8961±0.0314	0.8740±0.0118	0.8861±0.0086	0.8815±0.0116	0.9207±0.0065 ●
11	0.7135±0.0179	0.7325±0.0205	0.7090±0.0217	0.6867±0.0137	0.7430±0.0263 ●
12	0.4972±0.0270	0.4972±0.0270	0.4813±0.0304	0.5443±0.0378	0.6642±0.0233 ●
13	0.7341±0.0302	0.8062±0.0305	0.8071±0.0304	0.8126±0.0046	0.8196±0.0254 ●
14	0.8702±0.0212	0.8696±0.0211	0.8529±0.0223	0.8677±0.0131	0.9513±0.0189 ●
15	0.5934±0.0393	0.6091±0.0294	0.6606±0.0278	0.5814±0.0318	0.7200±0.0067 ●
16	0.4236±0.0019	0.4169±0.0022	0.4156±0.0026	0.4226±0.0033	0.4295±0.0038 ●
17	0.7110±0.0078	0.7069±0.0076	0.7199±0.0053	0.7131±0.0068	0.8075±0.0029 ●
18	0.6236±0.0068	0.6256±0.0047	0.6227±0.0060	0.6259±0.0041	0.6403±0.0124 ●
19	0.5659±0.0140	0.5853±0.0161	0.6109±0.0125	0.5993±0.0148	0.6345±0.0068 ●
20	0.5943±0.0117	0.5978±0.0120	0.5923±0.0111	0.5916±0.0119	0.6093±0.0077 ●
21	0.5984±0.0091	0.6025±0.0112	0.6091±0.0123	0.5974±0.0094	0.6163±0.0096 ●
22	0.5186±0.0108	0.5371±0.0132	0.5554±0.0096	0.5294±0.0128	0.5828±0.0120 ●
23	0.5374±0.0043	0.5385±0.0054	0.5349±0.0060	0.5382±0.0057	0.5370±0.0088
24	0.4556±0.0044	0.4429±0.0073	0.4601±0.0078	0.4325±0.0080	0.4651±0.0113 ●
25	0.5456±0.0059	0.5453±0.0058	0.5495±0.0072	0.5580±0.0047	0.5851±0.0038 ●

method DSME is $O(2h^2 + T_F)$, compared to AD (O($l^2 + T_F$)), JO ($O(l^2 + T_F)$), CAE ($T_C + T_F$), BA ($O(bl^2 + T_F)$), where b is the bagging times in the BA method, T_C is the time complexity of the employed clustering method in CAE, and T_F is the time complexity of the employed clustering ensemble method in the five compared methods. To be fair, we employ the average-link hierarchical clustering algorithm (Johnson 1967) based on the CO-matrix as the integrating method in all the five algorithms. The hierarchical clustering algorithm highly depends on the CO-matrix and it is helpful in reflecting the performance of the selected partitions.

Table 7. Index ARI from the Five Selective Methods

Datasets	AD	JO	CAE	BA	DSME
1	0.6717±0.0019	0.6737±0.0000	0.6686±0.0071	0.6790±0.0105	<u>0.6942±0.0297</u> •
2	0.8807±0.0053	0.8908±0.0025	0.8536±0.0075	0.8698±0.0029	<u>0.8967±0.0042</u> •
3	0.7224±0.0058	0.7276±0.0089	0.7306±0.0067	0.7461±0.0065	<u>0.7481±0.0111</u>
4	0.1630±0.0019	0.1621±0.0018	0.1633±0.0043	0.1636±0.0009	<u>0.1818±0.0028</u> •
5	0.8180±0.0083	0.8053±0.0090	0.8256±0.0052	0.8003±0.0068	<u>0.8261±0.0062</u>
6	0.4080±0.0082	0.4069±0.0115	0.4262±0.0083	0.4356±0.0220	<u>0.5039±0.0294</u> •
7	<u>0.3185±0.0017</u> •	0.3142±0.0026	0.2906±0.0059	0.3113±0.0060	0.3117±0.0090
8	0.2987±0.0204	0.3031±0.0173	0.3117±0.0191	0.3111±0.0151	<u>0.3208±0.0050</u> •
9	0.4131±0.0458	0.4747±0.0146	0.4293±0.0154	0.4075±0.0453	<u>0.5261±0.0057</u> •
10	0.6593±0.0742	0.5583±0.0363	0.5949±0.0263	0.5822±0.0354	<u>0.7082±0.0218</u> •
11	0.6283±0.0116	0.6369±0.0142	0.6291±0.0111	0.6235±0.0073	<u>0.6493±0.0169</u> •
12	0.1838±0.0226	0.1838±0.0226	0.1935±0.0322	<u>0.2851±0.0495</u>	0.2745±0.0427
13	0.2542±0.0474	0.4111±0.0449	0.4131±0.0444	<u>0.4395±0.0122</u>	0.4332±0.0446
14	0.8480±0.0213	0.8467±0.0211	0.8311±0.0225	0.8406±0.0116	<u>0.9397±0.0234</u> •
15	0.3351±0.0154	0.3293±0.0179	0.3488±0.0245	0.3128±0.0166	<u>0.3960±0.0070</u> •
16	0.3608±0.0014	0.3555±0.0011	0.3580±0.0029	0.3598±0.0028	<u>0.3684±0.0038</u> •
17	0.5722±0.0072	0.5663±0.0070	0.5876±0.0055	0.5748±0.0093	<u>0.6715±0.0060</u> •
18	0.4315±0.0089	0.4337±0.0087	0.4285±0.0079	0.4285±0.0051	<u>0.4730±0.0113</u> •
19	0.4818±0.0119	0.4993±0.0141	0.5214±0.0119	0.5084±0.0140	<u>0.5459±0.0059</u> •
20	0.4308±0.0144	0.4390±0.0155	0.4300±0.0145	0.4270±0.0151	<u>0.4548±0.0077</u> •
21	0.4530±0.0094	0.4538±0.0134	0.4618±0.0134	0.4518±0.0095	<u>0.4689±0.0115</u> •
22	0.3818±0.0136	0.4127±0.0198	0.4395±0.0160	0.4043±0.0199	<u>0.4864±0.0169</u> •
23	0.3013±0.0053	0.2999±0.0049	0.2962±0.0050	0.2996±0.0053	<u>0.3179±0.0045</u> •
24	0.2974±0.0093	0.2862±0.0115	0.3065±0.0064	0.2777±0.0082	<u>0.3634±0.0205</u> •
25	0.3908±0.0055	0.3940±0.0054	0.3986±0.0079	0.4066±0.0049	<u>0.4267±0.0043</u> •

Similar to Section 6.3, 50 sets of base partitions are generated for each dataset to eliminate influence caused by the uncertainty of the base partitions, and the ensemble performance is quantified by the indices AC, ARI, and NMI. For a single experiment, each method selects 25 base partitions from the ensemble. The values of the three indices from the five SCE methods are shown in Table 6 to Table 8.

In Table 6 to Table 8, the maximum value for each dataset is underlined. If the maximum value is significantly different from the others based on t-test, a bullet is assigned behind it. From Table 6 to Table 8, it is easy to see that DSME obtains in the most time the highest value of the three evaluation indices for clustering the twenty-five datasets. Concretely, DSME obtains average 21 times the highest value, which is much more than the sum of the other methods. In addition, DSME is significantly better than the other four methods on average 20 datasets. The results indicate that DSME is statistically better than the other four methods on the view of the three indices.

In what follows, we explore the effect of the number of selected partitions on ensemble performance in terms of AC. In this experiment, the number of selected partitions is gradually increased, which is set as $[5, 10, 15, \ldots, 45]$. The other experiment settings are the same as previous. The results are shown in Figure 5. It is obvious in Figure 5 that the curve of DSME is mostly lies above the other methods on the twenty-five datasets. In particular, the peak of AC vales on each dataset is obtained by DSME. The experiments show that the DSME is an effective method to handle the SCE problem.

Table 8. Index NMI from the Five Selective Methods

Datasets	AD	JO	CAE	BA	DSME
1	0.7586±0.0010	0.7596±0.0000	0.7571±0.0036	0.7627±0.0056	0.7689±0.0139 •
2	0.8541±0.0043	0.8651±0.0018	0.8177±0.0075	0.8346±0.0041	0.8656±0.0051
3	0.7132±0.0019	0.7158±0.0051	0.7174±0.0047	0.7143±0.0030	0.7244±0.0079 •
4	0.3251±0.0031	0.3233±0.0030	0.3171±0.0049	0.3249±0.0022	0.3469±0.0036 •
5	0.7430±0.0089	0.7290±0.0096	0.7498±0.0044	0.7358±0.0065	0.7566±0.0025 •
6	0.6046±0.0040	0.5998±0.0053	0.6083±0.0031	0.6099±0.0076	0.6333±0.0091 •
7	0.6081±0.0015 •	0.6041±0.0030	0.5807±0.0056	0.6018±0.0046	0.5950±0.0063
8	0.3897±0.0255	0.4008±0.0173	0.4172±0.0207	0.4210±0.0160 •	0.4117±0.0065
9	0.3586±0.0369	0.3777±0.0086	0.3870±0.0074	0.3574±0.0404	0.4366±0.0098 •
10	0.5811±0.0607	0.4780±0.0195	0.4871±0.0161	0.4914±0.0208	0.6161±0.0224 •
11	0.7738±0.0121	0.7760±0.0126	0.7735±0.0092	0.7794±0.0083	0.7929±0.0124 •
12	0.3256±0.0207	0.3256±0.0207	0.3310±0.0263	0.4081±0.0409	0.4539±0.0322 •
13	0.2164±0.0417	0.3458±0.0376	0.3447±0.0371	0.3737±0.0112	0.3715±0.0355
14	0.9393±0.0088	0.9386±0.0087	0.9329±0.0094	0.9377±0.0053	0.9772±0.0088 •
15	0.4104±0.0079	0.4090±0.0093	0.4168±0.0120	0.3957±0.0091	0.4378±0.0032 •
16	0.7015±0.0008	0.6983±0.0008	0.7000±0.0018	0.7023±0.0012	0.7062±0.0018 •
17	0.6412±0.0040	0.6349±0.0054	0.6500±0.0039	0.6454±0.0051	0.6884±0.0027 •
18	0.6025±0.0063	0.6032±0.0064	0.6013±0.0055	0.5992±0.0035	0.6215±0.0089 •
19	0.6214±0.0083	0.6272±0.0091	0.6310±0.0087	0.6320±0.0091	0.6443±0.0044 •
20	0.5582±0.0046	0.5635±0.0056	0.5576±0.0049	0.5572±0.0051	0.5661±0.0058 •
21	0.5936±0.0061	0.5941±0.0070	0.5978±0.0075	0.5919±0.0078	0.5992±0.0053
22	0.4057±0.0017	0.4084±0.0029	0.4265±0.0023	0.4149±0.0041	0.4569±0.0103 •
23	0.3542±0.0019	0.3567±0.0023	0.3512±0.0021	0.3551±0.0023	0.3622±0.0039 •
24	0.5818±0.0023	0.5792±0.0030	0.5825±0.0014	0.5773±0.0018	0.5815±0.0033
25	0.5917±0.0023	0.5927±0.0030	0.5967±0.0035	0.5942±0.0033	0.6015±0.0030 •

To visually show the performance of DSME, we run the five SCE algorithms on the Olivetti Face Database (Samaria and Harter 1994). This face dataset contains 400 figures of forty persons. For each person, there are ten figures. We employ the Density Peaks (DP) algorithm (Frey and Dueck 2007) to generate the base partitions. Because the DP algorithm will generate stable partition result when the distance matrix and the cluster number are fixed. To obtain diverse base partition results, we set the cluster numbers increase progressively from 20 to 70. The distance matrix of the Olivetti dataset is obtained from Frey and Dueck (2007). The number of selected partitions is also set as 25 in this experiment. The five SCE algorithms and the DP algorithm totally generate three different clustering results. The DP, JO, CAE, and BA generate the same result. The AD and the DSME generate the other two different results. Table 9 shows the three indices values of the three clustering results. It is easy to see from Table 9 that the DSME obtains the highest value of the three indices. The major differences between the three results come from 50 samples. Figure 6 to Figure 8 show the three results on the 50 particular samples. Comparing Figure 6 and Figure 8, it can be found that the figures of two persons in the red cluster in Figure 6 can be recognized by the DSME. From Figure 7 and Figure 8, it is obvious that the two persons in the blue cluster in Figure 7 can be separated by the DSME. Therefore, it can be concluded that the DSME generates more effective clustering result on the Olivetti Face Database than the other five methods.

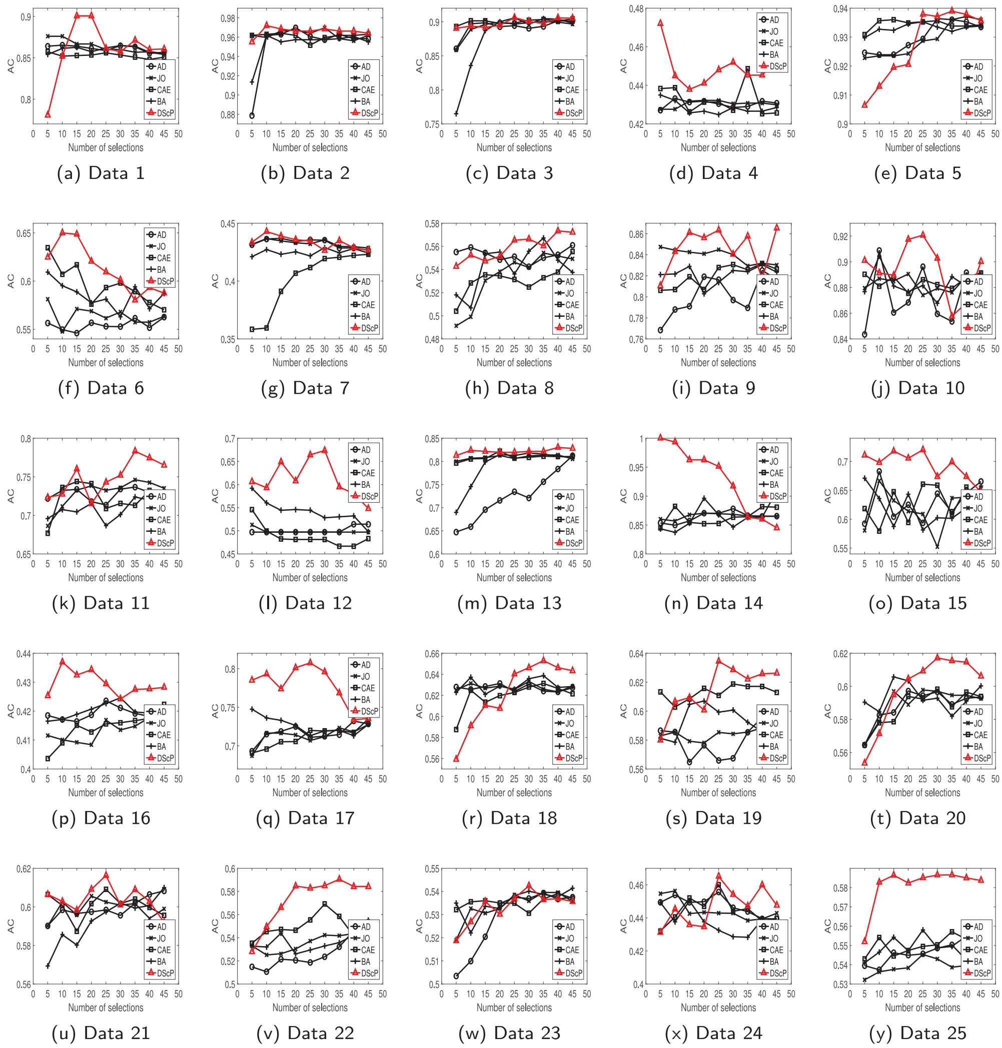


Fig. 5. Effect of the number of selections on performance in terms of AC.

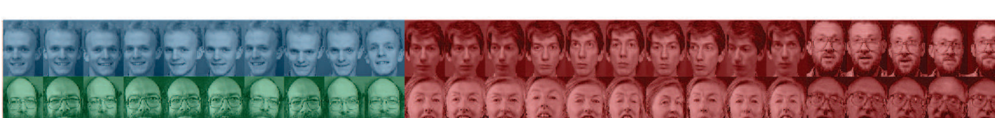


Fig. 6. The clustering result of the 50 particular Olivetti Face figures induced by the DP, JO, CAE, and BA.

7 CONCLUSION

Clustering ensemble, which integrates multiple diverse base clustering results, is an effective approach to improve the quality and the robustness of a single clustering algorithm in discovering the inherent grouping structure of a dataset. The WCE and SCE are two approaches to further improve the performance of a clustering ensemble method. The performance of these two approaches

Table 9. Indices from the DP Algorithm and the Five Selective Ensemble Methods for the Olivetti Face Database

	AC	ARI	NMI
AD	0.7975	0.5344	0.8577
DP, JO, CAE, BA	0.8000	0.5499	0.8580
DSME	**0.8100**	**0.5699**	**0.8641**

Fig. 7. The clustering result of the 50 particular Olivetti Face figures induced by the AD.

Fig. 8. The clustering result of the 50 particular Olivetti Face figures induced by the DSME.

are greatly affected by the employed similarity measure of two partitions. Due to the fact that the qualities of the clusters in a partition are different, the weighted method and selective method can be further improved through employing a measure that calculates the similarity between a cluster and a partition. The existing measures have two main problems, one is the symmetric problem and the other is the context meaning problem. In this article, we proposed a new measure SME. We proved that the SME is able to handle these two problems effectively in theory. Some properties of the SME make it effective in measuring the quality of each cluster in the ensemble. Moreover, we expanded SME to a similarity measure between two partitions, which is called SMEP.

Due to the different demands in the stages of SCE process, most of the existing SCE methods are complicated. To solve the SCE problem in a simple way, we proposed a novel framework DS, which considers the difference between the demand in the ensemble selection stage and the demand in the ensemble integration stage. We then exploited the advantages of SME and embedded it into DS, which forms DSME.

To verify the performances of SME and DSME, respectively, we compared SME with two similarity measures between a cluster and a partition, and compared DSME with four existing SCE methods which combine diversity and stability. The results show that SME is more effective in weighting the clusters in the ensemble, and DSME is more effective in discovering the grouping structure of a dataset. In future, it is interesting to expand SME to an index corrected for chance. Another interesting problem is the determination of the selection size.

REFERENCES

Ayan Acharya, Eduardo R. Hruschka, Joydeep Ghosh, and Sreangsu Acharyya. 2014. An optimization framework for combining ensembles of classifiers and clusterers with applications to nontransductive semisupervised learning and transfer learning. *ACM Transactions on Knowledge Discovery from Data* 9, 1 (2014), 1.

Ebrahim Akbari, Halina Mohamed Dahlan, Roliana Ibrahim, and Hosein Alizadeh. 2015. Hierarchical cluster ensemble selection. *Engineering Applications of Artificial Intelligence* 39 (2015), 146–156.

Hosein Alizadeh, Behrouz Minaei-Bidgoli, and Hamid Parvin. 2014. Cluster ensemble selection based on a new cluster stability measure. *Intelligent Data Analysis* 18, 3 (2014), 389–408.

Javad Azimi and Xiaoli Fern. 2009. Adaptive cluster ensemble selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. Vol. 9, 992–997.

Avrim L. Blum and Pat Langley. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 1 (1997), 245–271.

Si Chen, Gongde Guo, and Lifei Chen. 2010. A new over-sampling method based on cluster ensembles. In *Proceedings of the 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*. IEEE, 599–604.

Xiaoli Zhang Fern and Carla E. Brodley. 2003. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the International Conference on Machine Learning (ICML)*. Vol. 3, 186–193.

Xiaoli Zhang Fern and Carla E. Brodley. 2004. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the 21st International Conference on Machine Learning*. ACM, 36.

Xiaoli Z. Fern and Wei Lin. 2008. Cluster ensemble selection. *Statistical Analysis and Data Mining* 1, 3 (2008), 128–141.

Bernd Fischer and Joachim M. Buhmann. 2003. Bagging for path-based clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 11 (2003), 1411–1415.

Ana L. N. Fred and Anil K. Jain. 2005. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 6 (2005), 835–850.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science* 315, 5814 (2007), 972–976.

Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. 2007. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data* 1, 1 (2007), 4.

Stefan T. Hadjitodorov, Ludmila I. Kuncheva, and Ludmila P. Todorova. 2006. Moderate diversity for better cluster ensembles. *Information Fusion* 7, 3 (2006), 264–275.

Yi Hong, Sam Kwong, Hanli Wang, and Qingsheng Ren. 2009. Resampling-based selective clustering ensembles. *Pattern Recognition Letters* 30, 3 (2009), 298–305.

Dong Huang, Jianhuang Lai, and Chang-Dong Wang. 2016. Ensemble clustering using factor graph. *Pattern Recognition* 50 (2016), 131–142.

Shudong Huang, Hongjun Wang, Dingcheng Li, Yan Yang, and Tianrui Li. 2015. Spectral co-clustering ensemble. *Knowledge-Based Systems* 84 (2015), 46–55.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (1985), 193–218.

Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price. 2011. A link-based approach to the cluster ensemble problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 12 (2011), 2396–2409.

Anil K. Jain, Robert P. W. Duin, and Jianchang Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1 (2000), 4–37.

Jianhua Jia, Xuan Xiao, Bingxiang Liu, and Licheng Jiao. 2011. Bagging-based spectral clustering ensemble selection. *Pattern Recognition Letters* 32, 10 (2011), 1456–1467.

Liping Jing, Kuang Tian, and Joshua Z. Huang. 2015. Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition* 48, 11 (2015), 3688–3702.

Stephen C. Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

Imran Khan, Joshua Z. Huang, and Kamen Ivanov. 2016. Incremental density-based ensemble clustering over evolving data streams. *Neurocomputing* 191 (2016), 34–43.

Ludmila I. Kuncheva and Stefan Todorov Hadjitodorov. 2004. Using diversity in cluster ensembles. In *Proceedings of the 2004 IEEE International Conference On Systems, Man and Cybernetics*. Vol. 2, IEEE, 1214–1219.

Ludmila I. Kuncheva and Dmitry P. Vetrov. 2006. Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 11 (2006), 1798–1808.

Martin H. C. Law, Alexander P. Topchy, and Anil K. Jain. 2004. Multiobjective data clustering. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 2, IEEE, II–II.

Feijiang Li, Yuhua Qian, Jieting Wang, and Jiye Liang. 2017. Multigranulation information fusion: A dempster-shafer evidence theory-based clustering ensemble method. *Information Sciences* 378 (2017), 389–409.

Tao Li and Chris Ding. 2008. Weighted consensus clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*. SIAM, 798–809.

Yan Li, Yunming Ye, Zhaocai Sun, Edward Hung, Joshua Huang, and Yueping Li. 2013. An ensemble of decision cluster crotches for classification of high dimensional data. *Knowledge-Based Systems* 43 (2013), 63–73.

Murilo Coelho Naldi, A. C. P. L. F. Carvalho, and R. J. G. B. Campello. 2013. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery* 27, 2 (2013), 1–31.

Yuhua Qian, Feijiang Li, Jiye Liang, Bing Liu, and Chuangyin Dang. 2016. Space structure and clustering of categorical data. *IEEE Transactions on Neural Networks and Learning Systems* 27, 10 (2016), 2047–2059.

Yuhua Qian, Jiye Liang, Witold Pedrycz, and Chuangyin Dang. 2010. Positive approximation: An accelerator for attribute reduction in rough set theory. *Artificial Intelligence* 174, 9–10 (2010), 597–618.

Yuhua Qian, Hang Xu, Jiye Liang, Bing Liu, and Jieting Wang. 2015. Fusing monotonic decision trees. *IEEE Transactions on Knowledge and Data Engineering* 27, 10 (2015), 2717–2728.

Parisa Rastin and Rushed Kanawati. 2015. A multiplex-network based approach for clustering ensemble selection. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 1332–1339.

David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, and Pardis C. Sabeti. 2011. Detecting novel associations in large data sets. *Science* 334, 6062 (2011), 1518–1524.

Ferdinando S. Samaria and Andy C. Harter. 1994. Parameterisation of a stochastic model for human face identification. In *Proceedings of the 2nd IEEE Workshop on Applications of Computer Vision, 1994*. IEEE, 138–142.

Michael Steinbach, George Karypis, and Vipin Kumar. 2000. A comparison of document clustering techniques. In *Proceedings of the KDD Workshop on Text Mining*. Vol. 400, Boston, 525–526.

Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research* 3 (Dec. 2002), 583–617.

Alexander Topchy, Anil K. Jain, and William Punch. 2005. Clustering ensembles: Models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 12 (2005), 1866–1881.

Jun Wang, Fu-lai Chung, Shitong Wang, and Zhaohong Deng. 2014. Double indices-induced FCM clustering and its integration with fuzzy subspace clustering. *Pattern Analysis and Applications* 17, 3 (2014), 549–566.

Junjie Wu, Hongfu Liu, Hui Xiong, Jie Cao, and Jian Chen. 2015. K-means-based consensus clustering: A unified view. *IEEE Transactions on Knowledge and Data Engineering* 27, 1 (2015), 155–169.

Sen Xu, Kung-Sik Chan, Jun Gao, Xiufang Xu, Xianfeng Li, Xiaopeng Hua, and Jing An. 2016. An integrated K-means–Laplacian cluster ensemble approach for document datasets. *Neurocomputing* 214 (2016), 495–507.

Fan Yang, Xuan Li, Qianmu Li, and Tao Li. 2014. Exploring the diversity in cluster ensemble generation: Random sampling and random projection. *Expert Systems with Applications* 41, 10 (2014), 4844–4866.

Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval* 1, 1 (1999), 69–90.

Yun Yang and Ke Chen. 2011. Temporal data clustering via weighted clustering ensemble with different representations. *IEEE Transactions on Knowledge and Data Engineering* 23, 2 (2011), 307–320.

Muhammad Yousefnezhad, Sheng-Jun Huang, and Daoqiang Zhang. 2018. WoCE: A framework for clustering ensemble by exploiting the wisdom of crowds theory. *IEEE Transactions on Cybernetics* 48, 2 (2018), 486–499.

Muhammad Yousefnezhad and Daoqiang Zhang. 2015. Weighted spectral cluster ensemble. In *Proceedings of the 2015 IEEE International Conference on Data Mining*. IEEE, 549–558.

Zhiwen Yu, Le Li, Jiming Liu, Jun Zhang, and Guoqiang Han. 2015. Adaptive noise immune cluster ensemble using affinity propagation. *IEEE Transactions on Knowledge and Data Engineering* 27, 12 (2015), 3176–3189.

Li Zheng, Tao Li, and Chris Ding. 2014. A framework for hierarchical ensemble clustering. *ACM Transactions on Knowledge Discovery from Data* 9, 2 (2014), 9.

Zhi-Hua Zhou and Wei Tang. 2006. Clusterer ensemble. *Knowledge-Based Systems* 19, 1 (2006), 77–83.

Received October 2017; revised April 2018; accepted April 2018