A Framework for Cluster Ensemble Based on a Max Metric as Cluster Evaluator

Hosein Alizadeh, Hamid Parvin, and Sajad Parvin

***Abstract*—A new criterion for clusters validation is proposed in the paper and based on the new cluster validation criterion a clustering ensemble framework is proposed. The main idea behind the framework is to extract the most stable clusters in terms of the defined criteria. To combine a set of partitions into one consensus partition, hierarchical clustering algorithms can be employed where first the EAC method is applied over the output partitions to convert them into a co-association matrix and then considering it as a new data space bring a consensus partition out of them. But in proposed method due to having a set of clusters instead of a set of partitions, to extract the best representative consensus partition out of the set of chosen clusters the EAC method cannot be employed, and then we turn to a new EAC based method which is called Extended EAC, EEAC. EEAC is applied to construct the co-association matrix from the subset of clusters. Finally employing a simple hierarchical clustering algorithm as final consensus function the final representative partition is produced. Employing this new cluster validation criterion, the obtained ensemble is evaluated on some well-known and standard data sets. The empirical studies show promising results for the ensemble obtained using the proposed criterion comparing with the ensemble obtained using the standard clusters validation criterion.**

***Index Terms*— Clustering Ensemble, Stability Measure, Extended EAC, Co-association Matrix, Cluster Evaluation**

I. Introduction

DATA clustering or unsupervised learning is an important and very difficult problem. The objective of clustering is to partition a set of unlabeled objects into homogeneous groups or clusters [9], [11], [20] and [34]. There are many applications that use clustering techniques to discover latent structures of data, such as data mining [21], [32], information retrieval [4], face recognition [33], job scheduling [35], image segmentation [18], and machine learning. In real-world problems, clusters can appear with different shapes, sizes, data sparseness’s, and degrees of separation. Clustering techniques require the definition of a similarity measure between patterns. Since there is no prior knowledge about cluster shapes, choosing a specific clustering method is not easy [29]. Studies in the last few years have tended to combinational methods. Cluster ensemble methods attempt to find better and more robust clustering solutions by fusing information from several primary data partitions [16].

We propose a new criterion for clusters validation. Then we employ this criterion to select the more robust clusters in the final ensemble. We also propose a new method named Extended Evidence Accumulation Clustering, EEAC, to construct the matrix of similarity from these selected clusters. Finally, we apply a hierarchical method over the obtained matrix to extract the final partition.

Fern and Lin [16] have suggested a clustering ensemble approach which selects a subset of solutions to form a smaller but better-performing cluster ensemble than using all primary solutions. The ensemble selection method is designed based on quality and diversity, the two factors that have been shown to influence cluster ensemble performance. This method attempts to select a subset of primary partitions which simultaneously has both the highest quality and the most diversity. The Sum of Normalized Mutual Information, SNMI [13], [14] and [30], is used to measure the quality of each individual partition with respect to other partitions. Also, the Normalized Mutual Information, NMI, is employed to measure the diversity among partitions. Although the ensemble size in this method is relatively small, this method achieves significant performance improvement over full ensembles. Law et al. proposed a multi-objective data clustering method based on the selection of individual clusters produced by several clustering algorithms through an optimization procedure [25]. This technique chooses the best set of objective functions for different parts of the feature space from the results of base clustering algorithms. Fred and Jain [15] have offered a new clustering ensemble method which learns the pairwise similarities between points in order to facilitate a proper partition of the data without the a priori knowledge of the number and the shape of the clusters. This method which is based on cluster stability evaluates the primary clustering results instead of final clustering.

Rest of this paper is organized as follows. Section 2 is related works. In section 3, we explain the proposed method. Section 4 demonstrates results of our proposed method against traditional comparatively. Finally, we conclude in section 5.

II. Related Works

A. Review Stage

The clustering ensemble which is based on a subset of selected primary clusters or partitions has a main problem which is the manner of evaluating clusters or partitions. As the data clustering is an unsupervised problem, its validation process is the most troublesome task. Baumgartner et al. [2]

Manuscript received September 5, 2011 and revised December 2, 2011.

Hosein Alizadeh is with Nourabad Mamasani Branch, Islamic Azad University, Mamasani, Iran (phone: +989388930517; e-mail: halizadeh@iust.ac.ir).

Hamid Parvin is with Nourabad Mamasani Branch, Islamic Azad University, Mamasani, Iran (e-mail: hamidparvin@mamasaniiau.ac.ir).

Sajad Parvin is with Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran (e-mail: s.parvin@iust.ac.ir).

(Advance online publication: 27 February 2012)

have presented a resampling based technique to validate the results of exploratory fuzzy clustering analysis. Since the concept of cluster stability was introduced as a means to assess the validity of data partitions, it has been incrementally used in the literature [14]. This idea which is based on resampling method is initially described in [6] and later generalized in different ways in [17]. Roth et al. [28] have proposed a resampling based technique to validate a cluster. The basic element in their method which is a complementary version of the previous methods is the cluster stability. The stability measures the association between obtained partitions from two individual clustering algorithms. The great values of the stability measure mean that applying the clustering algorithm several times on a data set probably yields the fixed results [27]. Roth and Lange [29] have presented a new algorithm for data clustering which is based on feature selection. In their method the resampling based stability measure is used to set the algorithm parameters. There are several cluster validation methods which are based on the stability concept [24]. Ben-Hur et al. [3] have proposed a technique to exploit the stability measurements of the clustering solutions obtained by perturbing a data set. In their approach, the stability is characterized by the distribution of the pairwise similarities between clusterings obtained from sub samples of the data. First, the co-association matrix is acquired using the resampling method. Then, Jaccard coefficient is extracted from this matrix as the stability measure. Also, Estivill-Castro and Yang [10] have offered a method by which Support Vector Machines are used to evaluate the separation of the clustering results. By filtering noise and outliers, this method can identify the robust and potentially meaningful clustering result.

Moller and Radke [22] have introduced an approach to validate a clustering results based on partition stability. This method uses a perturbation which is produced by adding some noise to the data. An empirical study robustly indicates that the perturbation usually outperforms bootstrapping and subsampling. Whereas the empirical choice of the subsampling size is often difficult [8], the choosing of the perturbation strength is not so crucial. This method uses a Nearest Neighbor Resampling approach (NNR) that offers a solution to both problems of information loss and empirical control of the change degree made to the original data. The NNR techniques were first used for time series analysis [5]. Inokuchi et al. [19] have proposed a kernelized validity measures where a kernel means the kernel function used in support vector machines. Two measures are considered in this measure. One is the sum of the traces of the fuzzy covariances within clusters and the second is a kernelized Xie-Beni's measure [31]. This validity measure is applied to the determination of the number of clusters and also the evaluation of robustness of different partitions. Das and Sil [7] have proposed a method to determine the number of clusters which validates the clusters using splitting and merging technique in order to obtain optimal set of clusters.

Fern and Lin [12] have suggested a clustering ensemble approach which selects a subset of solutions to form a smaller but better performing cluster ensemble than using all primary solutions. The ensemble selection method is designed based on quality and diversity, the two factors that

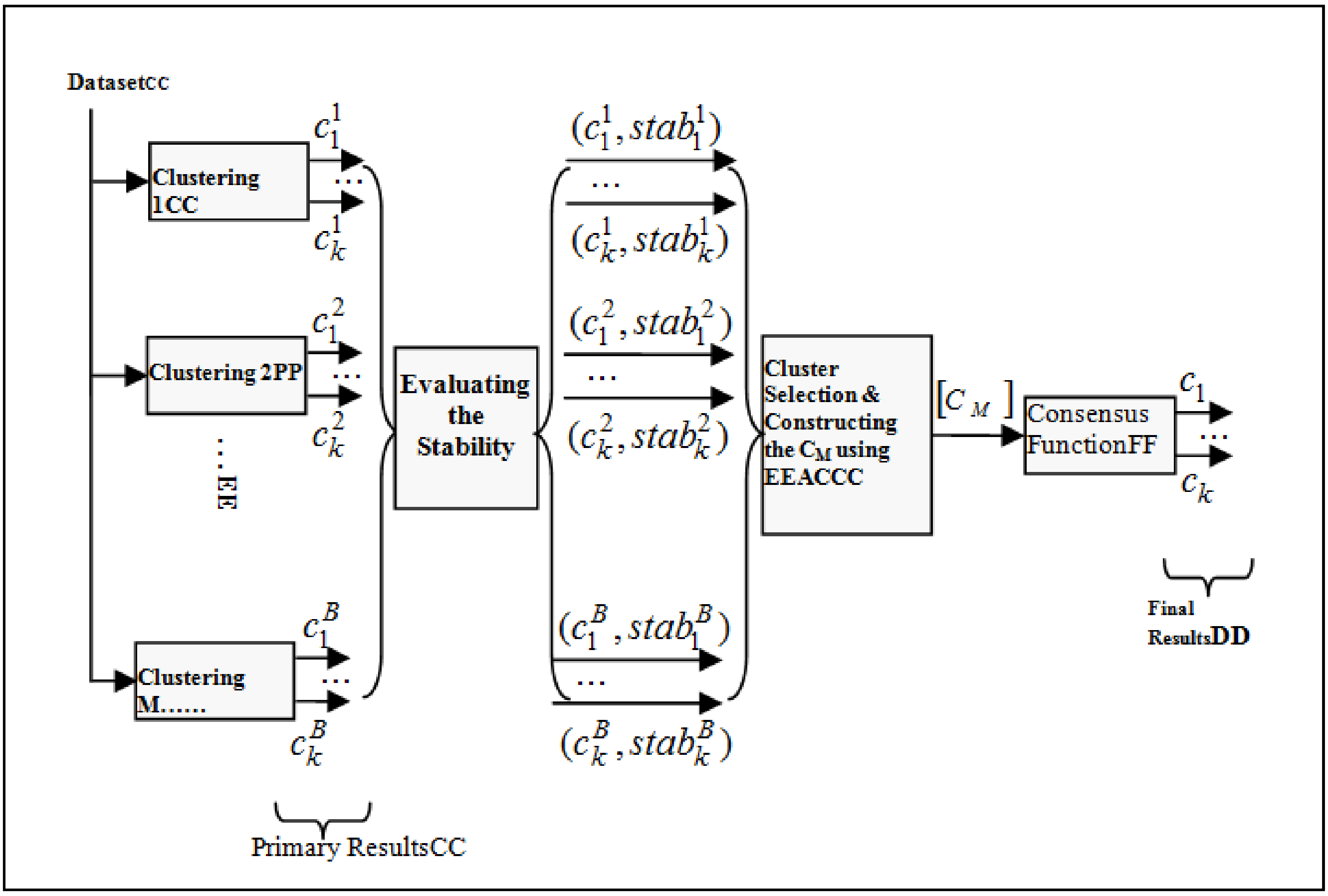


Fig. 1. Framework of clustering ensemble.

have been shown to influence the cluster ensemble performance. This method attempts to select a subset of primary partitions that simultaneously has both the highest quality and diversity. The Sum of Normalized Mutual Information, SNMI [13], [14] and [30], is used to measure the quality of individual partition with respect to other ones. Also, the Normalized Mutual Information, NMI, is employed to measure the diversity between partitions. Although the ensemble size in their method is relatively small, this method can achieve a significant performance improvement over full ensembles. Law et al. propose a multi objective data clustering method based on the selection of individual clusters produced by several clustering algorithms, through an optimization procedure [25]. This technique chooses the best set of objective functions for different parts of the feature space from the results of base clustering algorithms. Fred and Jain [15] have offered a new clustering ensemble method that learns the pairwise similarity between points in order to facilitate a proper partitioning of the data without the a priori knowledge of the number of clusters and of the shape of the clusters. This method which is based on cluster stability evaluates the primary clustering results instead of final clustering.

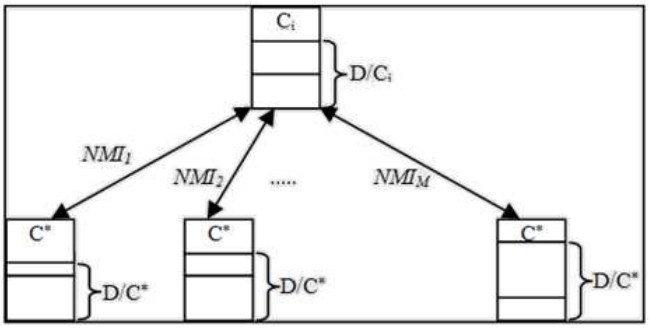


Fig. 2. Computing the Stability of Cluster C_i.

III. Proposed Method

In this section, first our proposed clustering ensemble method is briefly outlined, and then its phases are described in detail.

The main idea of our proposed clustering ensemble framework is utilizing a subset of best performing primary clusters in the ensemble, rather than using all of clusters. Only the clusters that satisfy a stability criterion can participate in the combination. The cluster stability is defined according to Normalized Mutual Information, NMI. Figure 1 depicts the proposed clustering ensemble procedure.

The manner of computing stability is described in the following sections in detail. To select a subset with the most stable clusters for combination, we apply a stability-threshold to each cluster. Different sizes of the most stable clusters are explored to find the best option. After selection phase, the selected clusters are used to construct the co-association matrix. Several methods have been proposed for combination of the primary results [1] and [30]. In this work, some clusters in the primary partitions may be absent (having been eliminated by the stability criterion). Since the original EAC method [13] cannot truly identify the pairwise similarity while there is only a subset of clusters, we present a new method for constructing the co-association matrix. We call this method: Extended Evidence Accumulation Clustering method, EEAC. Finally, we use a hierarchical clustering algorithm, like single-link method, to extract the final clusters out of this matrix. For more generality, some heuristic consensus functions are also used as aggregators of selected clusters [30]. These heuristic consensus functions that are based on hypergraph partitioning and have first introduced by Strehl and Ghosh, are HperGraph Partitioning Algorithm (HGPA), Meta-Clustering Algorithm (MCLA) and Cluster-based Similarity Partitioning Algorithm (CSPA) [30].

A. *Cluster Evaluation*

Since goodness of a cluster is determined by all the data points, the goodness function $g_j(C_i,D)$ depends on both the cluster C_i and the entire dataset D, instead of C_i alone. The stability as measure of cluster goodness is used in [24]. Cluster stability reflects the variation in the clustering results under perturbation of the data by resampling.

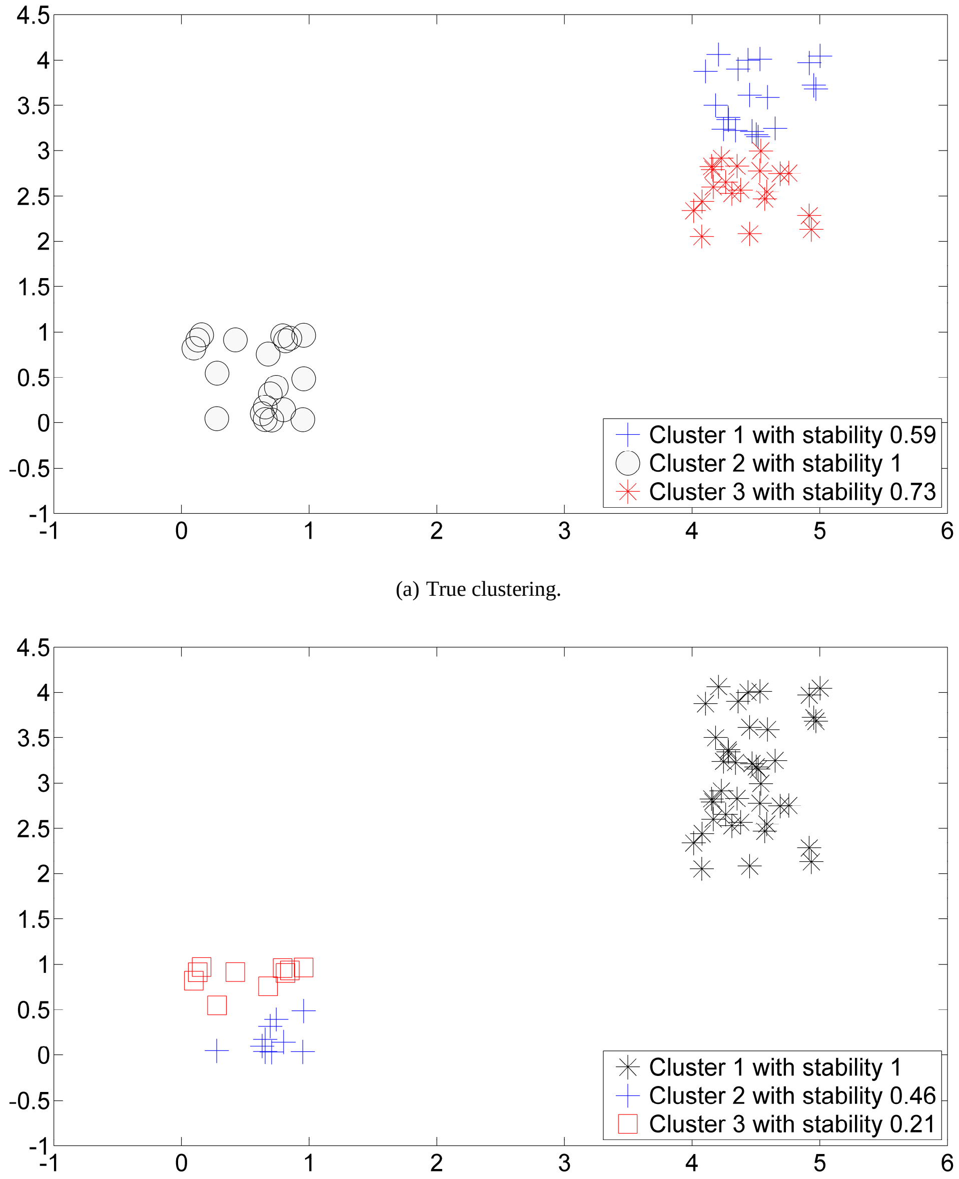
A stable cluster is one that has a high likelihood of recurrence across multiple applications of the clustering method. Stable clusters are usually preferable, since they are robust with respect to minor changes in the dataset [25].

Now assume that we want to compute the stability of cluster C_i. In this method first a set of partitionings over resampled datasets is provided which is called the reference set. In this notation D is resampled data and $P(D)$ is a partitioning over D. Now, the problem is: “How many times is the cluster C_i repeated in the reference partitions?” Denote by NMI(C_i,$P(D)$), the Normalized Mutual Information between the cluster C_i and a reference partition $P(D)$. Most previous works only compare a *partition with another partition* [30]. However, the stability used in [25] evaluates the similarity between a *cluster and a partition* by transforming the cluster C_i to a partition and employing common partition to partition methods. To illustrate this method let $P_1 = P^a = \{C_i,D/C_i\}$ be a partition with two clusters, where D/C_i denotes the set of data points in D that are not in C_i.

Then we may compute a second partition $P_2 = P^b = \{C^*,D/C^*\}$, where C^* denotes the union of all “positive” clusters in $P(D)$ and others are in D/C^*. A cluster C_j in $P(D)$ is positive if more than half of its data points are in C_i. Now, define NMI(C_i,$P(D)$) by NMI(P^a,P^b) which is calculated as [14]:

$$NMI(P^a,P^b)=\frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b} n_{ij}^{ab}\log\left(\frac{n_{ij}^{ab}.n}{n_i^a.n_j^b}\right)}{\sum_{i=1}^{k_a} n_i^a\log\left(\frac{n_i^a}{n}\right)+\sum_{j=1}^{k_b} n_j^b\log\left(\frac{n_j^b}{n}\right)} \qquad 1$$

where n is the total number of samples and n_{ij}^{ab} denotes the number of shared patterns between clusters $C_i^a \in P^a$ and $C_j^b \in P^b$; n_i^a is the number of patterns in the cluster i of partition a; also n_j^b are the number of patterns in the cluster j of partition b.



(a) True clustering.

(b) Spurious clustering.

Fig. 3. Two primary partitions with k=3.

This computation is done between the cluster C_i and all partitions available in the reference set. Fig. 2 shows this method.

	π_1	π_2	π_3	π_4
$\mathbf{x}_1$	2	1	2	1
$\mathbf{x}_2$	2	1	2	1
$\mathbf{x}_3$	2	1	2	2
$\mathbf{x}_4$	2	1	2	2
$\mathbf{x}_5$	3	2	1	2
$\mathbf{x}_6$	3	2	2	2
$\mathbf{x}_7$	1	3	1	1
$\mathbf{x}_8$	1	3	1	1
$\mathbf{x}_9$	1	4	1	2
$\mathbf{x}_{10}$	1	3	1	1
$\mathbf{x}_{11}$	1	4	1	1
$\mathbf{x}_{12}$	1	4	1	1

Fig. 4. Four partitions π_1 - π_4 are extracted from a simple dataset with 12 data points and two real clusters with k-means clustering. The k parameters in k-means is set to 3, 4, 2 and 2 respectively.

NMI_i in Fig. 2 shows the stability of cluster C_i with respect to the i-th partition in reference set. The total stability of cluster C_i is defined as:

$$Stability(C_i) = \frac{1}{M}\sum_{i=1}^{M} NMI_i \qquad 2$$

where M is the number of partitions available in reference set. This procedure is applied for each cluster of every primary partition.

B. Max Method

In this section a drawback of computing stability is introduced and an alternative approach is suggested which is named Max method. Fig. 3 shows two primary partitions for which the stability of each cluster is evaluated. In this example K-means is applied as the base clustering algorithm with K=3. For this example the number of all partitions in the reference set is 40. In 36 partitions the result is relatively similar to Fig 3a, but there are four partitions in which the top left cluster is divided into two clusters, as shown in Fig

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}	ϕ_{11}
x_1	0	1	0	1	0	0	0	0	1	1	0
x_2	0	1	0	1	0	0	0	0	1	1	0
x_3	0	1	0	1	0	0	0	0	1	0	1
x_4	0	1	0	1	0	0	0	0	1	0	1
x_5	0	0	1	0	1	0	0	1	0	0	1
x_6	0	0	1	0	1	0	0	0	1	0	1
x_7	1	0	0	0	0	1	0	1	0	1	0
x_8	1	0	0	0	0	1	0	1	0	1	0
x_9	1	0	0	0	0	0	1	1	0	0	1
x_{10}	1	0	0	0	0	1	0	1	0	1	0
x_{11}	1	0	0	0	0	0	1	1	0	1	0
x_{12}	1	0	0	0	0	0	1	1	0	1	0

Fig. 5. The clusters extracted from partitions of Fig. 4.

3b. Fig 3a shows a true clustering. Since the well separated cluster in the top left corner is repeated several times (90% repetition) in partitions of the reference set, it has to acquire a great stability value (but not equal to 1), however it acquires the stability value of 1. Because the two clusters in right hand of Fig 3a are relatively joined and sometimes they are not recognized in the reference set as well, they have less stability value. Fig. 3.b shows a spurious clustering which the two right clusters are incorrectly merged. Since a fixed number of clusters are forced in the base algorithm, the top left cluster is divided into two clusters. Here the drawback of the stability measure is apparent rarely. Although it is obvious that this partition and the corresponding large cluster on the right reference set (10% repetition), the stability of this cluster is evaluated equal to 1. Since the NMI is a symmetric equation, the stability of the top left cluster in fig 3.a is exactly equal to the large right cluster in fig 3.b; however they are repeated 90% and 10%, respectively. In other words, when two clusters are complements of each other, their stabilities are always equal. This drawback is seen when the number of positive clusters in the considered partition of reference set is greater than 1. It means when the cluster C* is obtained by merging two or more clusters, undesirable stability effects occur.

To solve this problem we allow only one cluster in reference set to be considered as the *C** (i.e. only the most similar cluster) and all others are considered as *D/C**. In this method the problem is solved by eliminating the merged clusters.

C. Consensus Function

One way is to consider the selected clusters as inputs of the HGPA, MCLA and CSPA algorithms [30]. The output of the mentioned algorithms is the final partition which is also called consensus partition. For example consider the Fig. 4. Four partitions π_1 - π_4 are extracted from a simple dataset with 12 data points and two real clusters with k-means clustering. The k parameters in k-means is set to 3, 4, 2 and 2 respectively. These partitions are broken into 11 clusters depicted in Fig. 5. The clusters are served as input for the HGPA, MCLA and CSPA algorithms.

For the second way to extract the final partition from the selected clusters, the clusters are considered as new space for data, and a clustering algorithm, like fuzzy k-means, is employed to partition the mapped data. For example again consider the example of Fig. 4. The partitions are broken into 11 clusters depicted in the Fig. 5 as before. Then the clusters of Fig. 4, considered as the mapped data into a new feature space and a fuzzy k-means is extracted consensus partition from them.

Another alternative way to reach the consensus partition

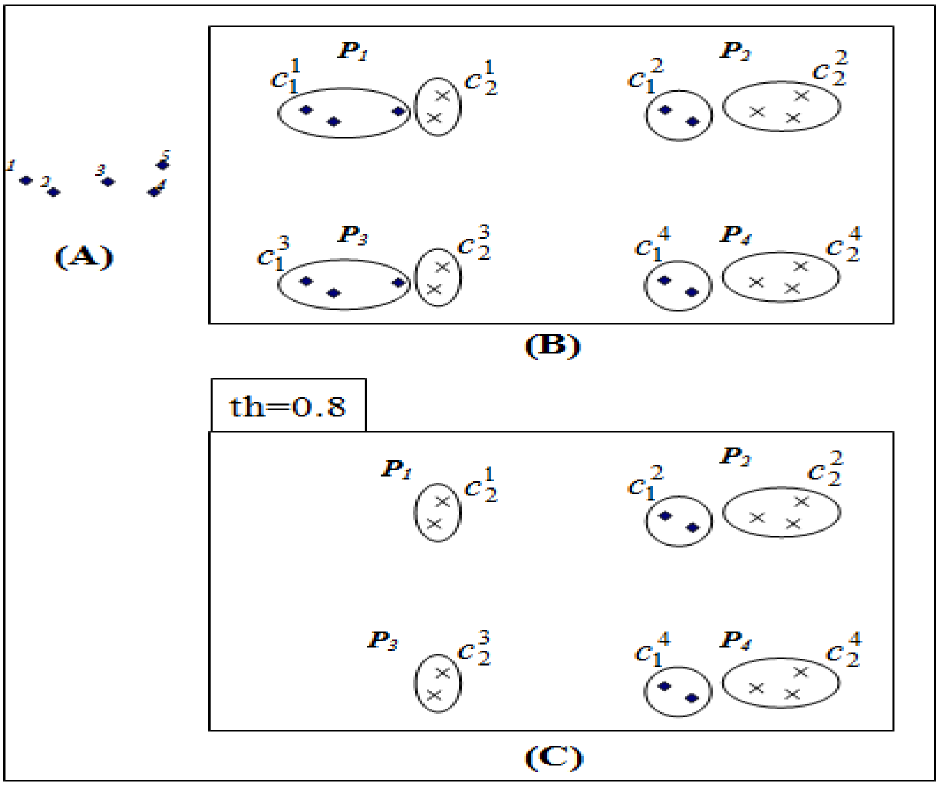


Fig. 6. Computing the co-association matrix by the EEAC method. (A) Data samples. (B) 4 primary clusterings. (C) Remaining clusters after applying threshold, th=0.8.

is to use the co-association based methods. In this method, the selected clusters are first used to construct the co-association matrix. In the EAC method the m primary results from resampled data are accumulated in an $n \times n$ co-association matrix. Each entry in this matrix is computed from this equation:

$$C(i,j) = \frac{n_{i,j}}{m_{i,j}} \qquad 3$$

where n_{ij} counts the number of clusters shared by objects with indices i and j in the partitions over the B clusterings. Also m_{ij} is the number of partitions where this pair of objects is simultaneously present. There are only a fraction of all primary clusters available, after thresholding. So, the common EAC method cannot truly recognize the pairwise similarity for computing the co-association matrix. In our novel method (Extended Evidence Accumulation Clustering, or EEAC) each entry of the co-association matrix is computed by:

$$C(i,j) = \frac{n_{i,j}}{\max(n_i, n_j)} \qquad 4$$

where n_i and n_j are the number present in remaining (after stability thresholding) clusters for the i-th and j-th data points, respectively. Also, n_{ij} counts the number of remaining clusters which are shared by both data points indexed by i and j, respectively. To further explain, consider this example. Assume that we have five samples (Fig 6a), and that four primary clustering are applied (Fig 6b).

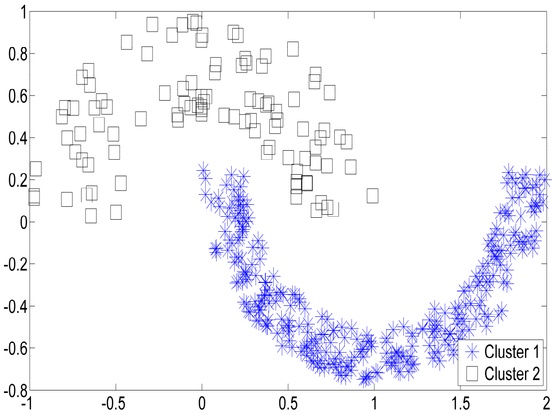


Fig. 7. Half Ring dataset.

Also, suppose that that stability of the clusters of Fig 6b is as given bellow:

$$Stability(c_2^1) = Stability(c_2^3) = 1$$

$$Stability(c_1^2) = Stability(c_1^4) = 1$$

$$Stability(c_2^2) = Stability(c_2^4) = 0.82$$

$$Stability(c_1^1) = Stability(c_1^3) = 0.55$$

By choosing th=0.8 the first clusters from P1 and P3 are deleted (Fig 6c). According to equation 4, each entry of the co-association matrix is:

$$C(1,2) = {}^{2}/_{\max(2,2)} = 1$$

$$C(1,3) = C(2,3) = {}^{0}/_{\max(2,2)} = 0$$

$$C(3,4) = C(3,5) = {}^{2}/_{\max(2,4)} = 0.5$$

$$C(4,5) = {}^{4}/_{\max(4,4)} = 1$$

Table 1. Brief information about the used datasets.

	Dataset Name	*# of Class*	*# of Features*	*# of Samples*
1	Breast-Cancer*	2	9	683
2	Iris*	3	4	150
3	Bupa*	2	6	345
4	SAHeart*	2	9	462
5	Ionosphere	2	34	351
6	Glass*	6	9	214
7	Halfrings	2	2	400
8	Galaxy*	7	4	323
9	Yeast*	10	8	1484
10	Wine	3	13	178

In Fig 6a-c, the data points may be "tracked" by their geometrical arrangement. Example: in computing C(3,4), note that points 3 and 4 both are in cluster 2 of partitions P2 and P4, so that numerator n_{34}=2; also note that n_3=2, since

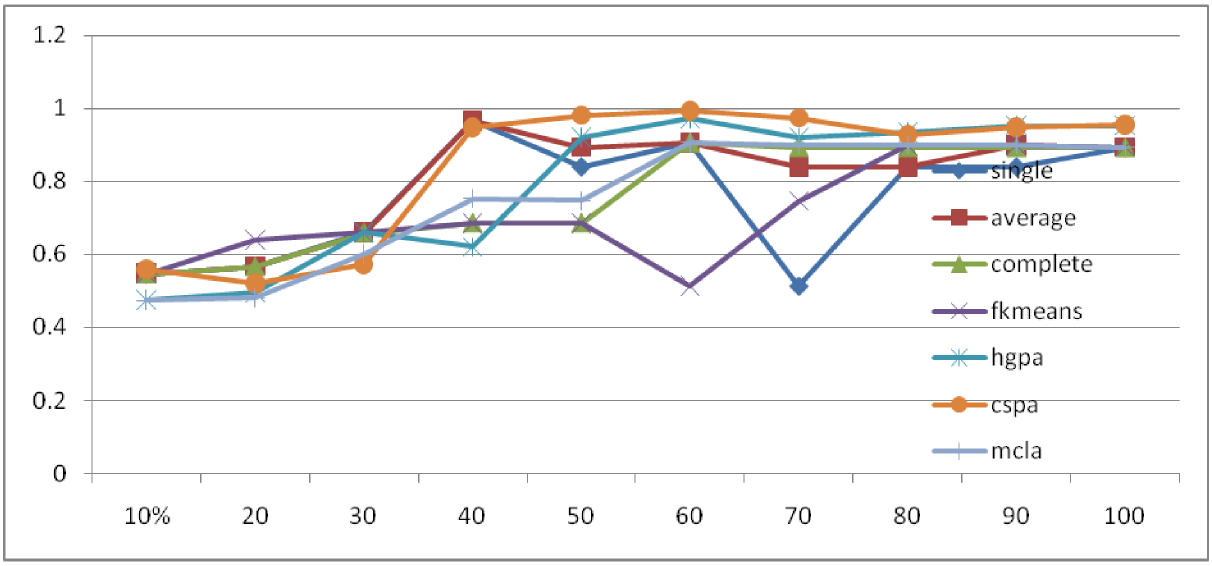


Fig. 8. The horizontal axis stands for the rate of stable clusters that are selected. The vertical axis stands for the NMI values between the labels of Iris dataset and the consensus partitions obtained by different consensus functions over the selected clusters.

point 3 is only in cluster 2 of P2 and P4, but n_4=4 since point 4 is not only in these clusters, but also in cluster 2 of P1 and P3. Before and after applying threshold, the co-association matrix is given by equation 3 and 4, respectively.

$$C_{before} = \begin{bmatrix} 1 & 1 & 0.5 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 \end{bmatrix}$$

In this matrix the 3rd object can be considered as both clusters with an equal probability of 50%. The stability measure adds some information to this matrix by applying the threshold.

$$C_{after} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 \end{bmatrix}$$

By comparing these two matrices and also considering the stability values, it can be seen that deletion of unstable clusters improves the co-association matrix. By eliminating the unstable cluster with samples {1, 2, 3} which is spuriously created by primary clusterings.

After computing the co-association matrix by the EEAC method, a consensus function is employed to extract the final clusters from the matrix. Here, the single-link method is used for this task.

IV. Experimental Study

Evaluation metric based on which a consensus partition is evaluated is discussed in the first subsection of this section. The details of the used datasets are given in the subsequent section. Then the settings of experimentations are given. Finally the experimental results are presented.

A. *Evaluation Metric*

After producing the consensus partition, the most important question is "how good a partition is?". The evaluation of a partition is very important as it is mentioned. Here the NMI between the consensus partition and real labels of the dataset is considered as an evaluation metric of

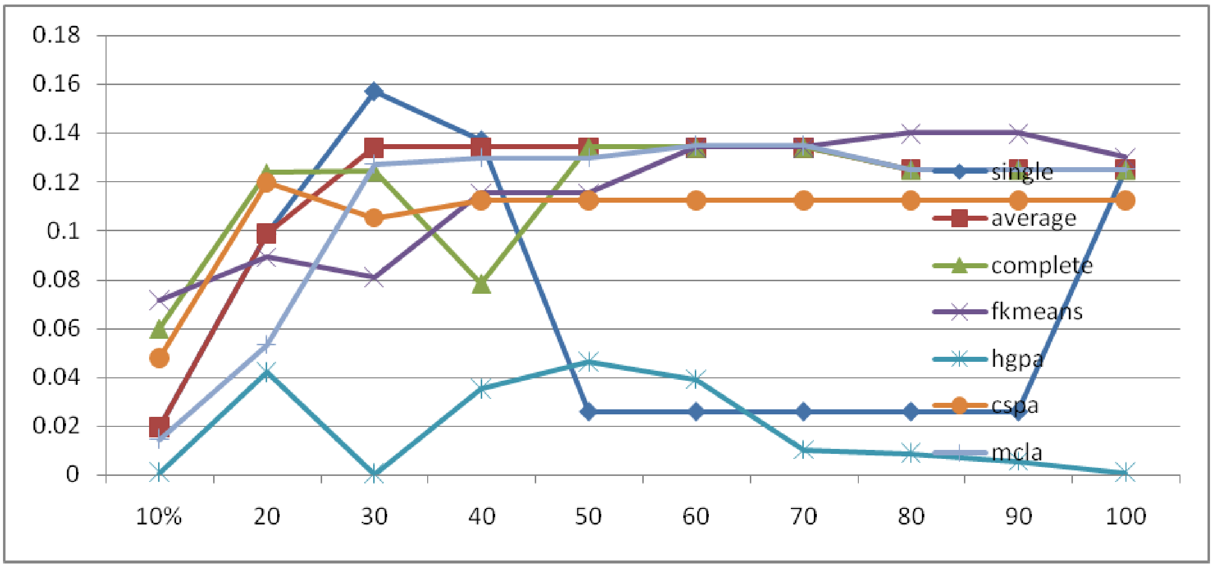


Fig. 9. The horizontal axis stands for the rate of stable clusters that are selected. The vertical axis stands for the NMI values between the labels of Ionosphere dataset and the consensus partitions obtained by different consensus functions over the selected clusters.

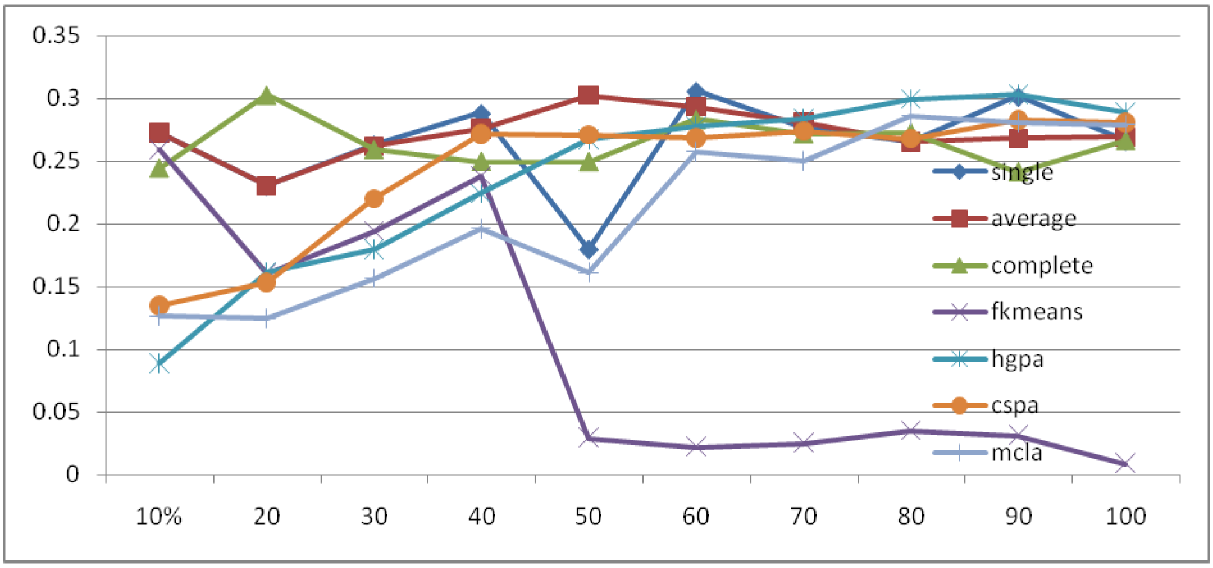


Fig. 10. The horizontal axis stands for the rate of stable clusters that are selected. The vertical axis stands for the NMI values between the labels of Galaxy dataset and the consensus partitions obtained by different consensus functions over the selected clusters.

the consensus partition. Also accuracy between the consensus partition and real labels of the dataset is considered as another metric.

B. Datasets

The proposed method is examined over 9 different standard datasets and one artificial dataset. It is tried for datasets to be diverse in their number of true classes, features and samples. A large variety in used datasets can more validate the obtained results. Brief information about the used datasets is available in Table 1. More information is available in [26].

Note that some of datasets which are marked with star (*) in Table 1 are normalized. All experiments are done over the normalized features in the stared dataset. It means each feature is normalized with mean of 0 and variance of 1, N(0, 1). The artificial Half Ring dataset is depicted in the Fig. 7.

C. Experimental Settings

To be more general and fair, all experiments are averaged over 10 independent runs. In all experimentations there are 120 independent partitions obtained by 120 independent runs of k-means clustering algorithm with different initialized seed points and different k parameter, ranging from k to 2*k.

After selecting a subset of clusters, to extract the final partition from them, the real number of clusters, i.e. the column three of the Table 1, is served by the consensus functions.

As it is known in fuzzy k-means clustering algorithm, each data point belongs to all clusters with different membership values. To extract the final partition from output of fuzzy k-means algorithm as consensus function, each data point is assigned to the most membership value.

D. Experimental Results

To see whether the use of a subset of the most stable clusters can affect the quality of the final cluster or not, consider Fig. 8. This figure depicts the NMI values between the consensus partitions obtained by different consensus functions over the selected clusters and the labels of Iris. As it is inferred from the Fig. 8, the best ratio of selection of the stable clusters is 60% and the best option for consensus function is CSPA for Iris dataset.

Fig. 9 depicts the NMI values between the consensus partitions obtained by different consensus functions over the selected clusters and the labels of Ionosphere. Fig. 9 makes it clear that the best ratio of selection of the stable clusters is 30% and the best option for consensus function is Single-Linkage for Ionosphere dataset.

Fig. 10 depicts the NMI values between the consensus partitions obtained by different consensus functions over the selected clusters and the labels of Galaxy. By choosing the consensus function to Complete-Linkage and the ratio of selection of stable clusters to 20% we reach the best performance for Galaxy dataset. Equivalently the Single-Linkage consensus function over 60% of the most stable clusters reaches the maximum for Galaxy dataset.

To make a general decisive conclusion, the results for all ten datasets of Table 1 are averaged and the final results are illustrated in the Fig. 11.

The Averaged-Linkage consensus function over 50% of the most stable clusters generally reaches the maximum for all dataset.

Table 2. Accuracy of consensus partition produced by cluster selection based on NMI and MAX measures.

Evaluation Method	Dataset Number									
	1	2	3	4	5	6	7	8	9	10
NMI	95.73	76.13	54.33	63.36	**70.60**	**47.76**	74.48	**31.27**	42.93	69.38
MAX	**96.49**	**84.87**	**57.42**	**63.87**	57.75	44.35	**74.55**	29.85	**51.27**	**70.00**

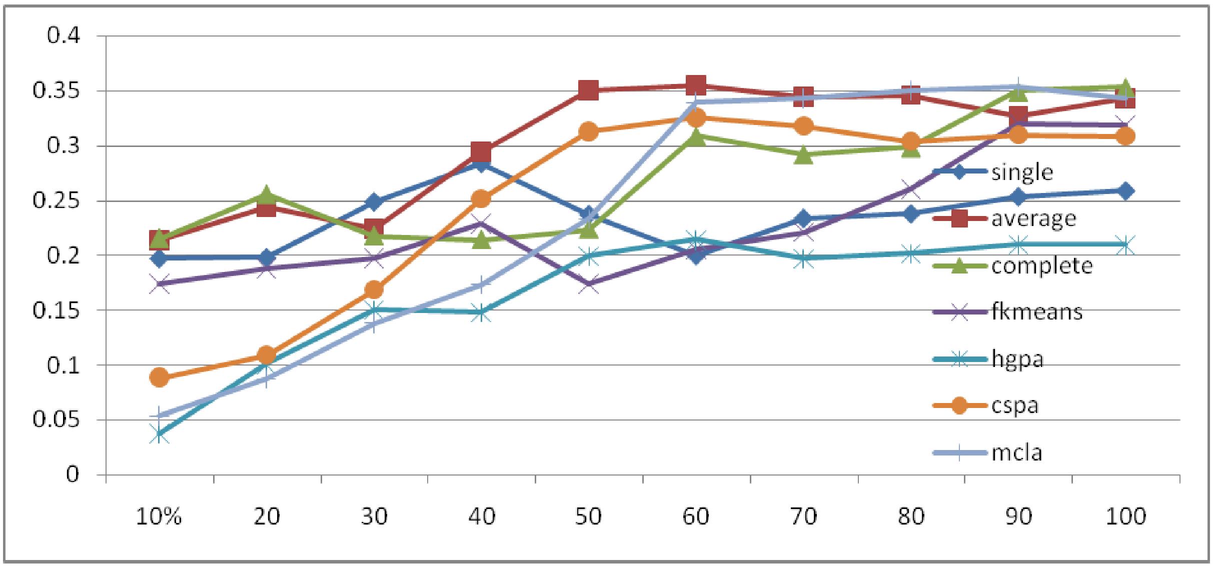


Fig. 11. The horizontal axis stands for the rate of stable clusters that are selected. The vertical axis stands for the averaged NMI values for all ten datasets of Table 1.

Table 2 represents the accuracy between real labels of dataset and consensus partition obtained by cluster selection based on NMI and using complete linkage hierarchical clustering as consensus function. In obtaining the results of Table 2, 33% of the most stable clusters are taken into consideration for participating in final ensemble.

V. Conclusion

In this paper a new clustering ensemble framework is proposed which is based on participating a subset of total primary spurious clusters. Also a new alternative method for common NMI is suggested. Since the quality of the primary clusters are not equal and presence of some of them can even yield to lower performance, here a method to select a subset of more effective clusters is proposed. A common cluster validity criterion which is needed to derive this subset is based on normalized mutual information. In this paper some drawbacks of this criterion is discussed and a method is suggested which is called max mehod. The main idea behind the framework is to extract the most stable clusters in terms of the defined criteria. To combine a set of partitions into one consensus partition, hierarchical clustering algorithms can be employed where first the EAC method is applied over the output partitions to convert them into a co-association matrix and then considering it as a new data space bring a consensus partition out of them. But in proposed method due to having a set of clusters instead of a set of partitions, to extract the best representative consensus partition out of the set of chosen clusters the EAC method cannot be employed, and then we turn to a new EAC based method which is called Extended EAC, EEAC. EEAC is applied to construct the co-association matrix from the subset of clusters. Finally employing a simple hierarchical clustering algorithm as final consensus function the final representative partition is produced. The experiments show that the proposed framework commonly outperforms in comparison with the full ensemble; also participation all clusters in the final ensemble is not a good option; however it uses just 33% of primary clusters. Also the proposed max criterion does slightly better than NMI criterion generally. Because of the symmetry which is concealed in NMI criterion and also in NMI based stability, it yields to lower performance whenever symmetry is also appeared in the dataset. Another innovation of this chapter is a method for constructing the co-association matrix where some of clusters and respectively some of samples do not exist in partitions. This new method is called Extended Evidence Accumulation Clustering, EEAC.

References

[1] Ayad H. and Kamel M.S. (2008), Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters, IEEE Trans. on Pattern Analysis and Machine Intelligence, VOL. 30, NO. 1, 160-173.

[2] Baumgartner, R., Somorjai, R., Summers, R., Richter, W., Ryner, L., and Jarmasz, M.: Resampling as a Cluster Validation Technique in fMRI, JOURNAL OF MAGNETIC RESONANCE IMAGING 11: pp. 228–231, (2000)

[3] Ben-Hur, A., Elisseeff, A., and Guyon, I.: A stability based method for discovering structure in clustered data. Pasific Symposium on Biocomputing, vol. 7, pp. 6-17 (2002)

[4] Bhatia S.K. and Deogun J.S. (1998), “Conceptual Clustering in Information Retrieval,” IEEE Trans. Systems, Man, and Cybernetics, vol. 28, no. 3, pp. 427-536.

[5] Brandsma, T. and Buishand, T.A.: Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling. Hydrology and Earth System Sciences 2, pp. 195-209 (1998)

[6] Breckenridge, J.: Replicating cluster analysis: Method, consistency and validity. Multivariate Behavioral research (1989)

[7] Das, A.K. and Sil, J.: Cluster Validation using Splitting and Merging Technique. Int. Conf. on Computational Intelligence and Multimedia Applications, ICCIMA (2007)

[8] Davison, A.C., Hinkley, D.V., and Young, G.A.: Recent developments in bootstrap methodology. Statistical Science 18, pp. 141-157 (2003)

[9] Dudoit S. and Fridlyand, J. (2003), “Bagging to improve the accuracy of a clustering procedure”, Bioinformatics, 19 (9), pp. 1090-1099.

[10] Estivill-Castro, V. and Yang, J.: Cluster Validity Using Support Vector Machines. DaWaK 2003, LNCS 2737, pp. 244–256 (2003)

[11] Faceli K., Marcilio C.P. Souto d. (2006), Multi-objective Clustering Ensemble, Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06).

[12] Fern X.Z. and Lin W. (2008), Cluster Ensemble Selection, SIAM International Conference on Data Mining (SDM08).

[13] Fred, A. and Jain, A. K. (2002). “Data Clustering Using Evidence Accumulation”, Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280.

[14] Fred A. and Jain A.K. (2005). Combining Multiple Clusterings Using Evidence Accumulation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(6):835–850.

[15] Fred A. and Jain A.K. (2006), Learning Pairwise Similarity for Data Clustering, In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06).

[16] Fred A. and Lourenco A. (2008), Cluster Ensemble Methods: from Single Clusterings to Combined Solutions, Studies in Computational Intelligence (SCI), 126, 3–30.

[17] Fridlyand, J. and Dudoit, S.: Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Stat. Berkeley Tech Report. No. 600 (2001)

[18] Frigui H. and Krishnapuram R. (1999), “A Robust Competitive Clustering Algorithm with Applications in Computer Vision,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 5, pp. 450-466.

[19] Inokuchi, R., Nakamura, T., and Miyamoto, S.: Kernelized Cluster Validity Measures and Application to Evaluation of Different Clustering Algorithms. IEEE Int. Conf. on Fuzzy Systems, Canada, July 16-21 (2006)

[20] Jain A.K., Murty M. N., and Flynn P. (1999), Data clustering: A review. ACM Computing Surveys, 31(3):264–323.

[21] Judd D., Mckinley P., and Jain A.K. (1997) “Large-Scale Parallel Data Clustering,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 153-158.

[22] Möller, U. and Radke, D.: Performance of data resampling methods based on clustering. Intelligent Data Analysis 10(2) (2006)

[23] Lange, T., Roth, V., Braun, M.L., and Buhmann, J.M.: Stability-based validation of clustering solutions. Neural Computation, 16(6):1299–1323 (2004)

[24] Lange, T., Braun M.L., Roth V., and Buhmann J.M. (2003). Stability-based model selection. In Advances in Neural Information Processing Systems 15. MIT Press.

[25] Law M.H.C., Topchy A.P., and Jain A.K. (2004). Multiobjective data clustering. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 424–430, Washington D.C.

[26] Newman C.B.D.J., Hettich S. and Merz C. (1998), UCI repository of machine learning databases, http://www.ics.uci.edu/˜mlearn/MLSummary.html.

[27] Rakhlin, A. and Caponnetto, A.: Stability of k-means clustering. In Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, MA (2007)

[28] Roth, V. and Lange, T.: Feature Selection in Clustering Problems. In Advances in Neural Information Processing Systems (2004)

[29] Roth V., Lange T., Braun M., and Buhmann J. (2002), A Resampling Approach to Cluster Validation, Intl. Conf. on Computational Statistics, COMPSTAT.

[30] Strehl A. and Ghosh J. (2002), Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3(Dec):583–617.

[31] Xie, X.L. and Beni, G.: A Validity measure for Fuzzy Clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.13, No.4, pp. 841–846 (1991)

[32] Sen, S., Narasimhan, S., and Konar, A.: Biological Data Mining for Genomic Clustering Using Unsupervised Neural Learning. Engineering Letters, Vol.14, No.2 (2007)

[33] Mario I. Chacon, M., Pablo Rivas, P., and Graciela Ramirez, A.: A Fuzzy Clustering Approach for Face Recognition Based on Face Feature Lines and Eigenvectors. Engineering Letters, Vol.15, No.1 (2007)

[34] Yang, C.H., Hsiao, C.J. and Chuang, L.Y.: Linearly Decreasing Weight Particle Swarm Optimization with Accelerated Strategy for Data Clustering. IAENG International Journal of Computer Science, Vol.37, No.3 (2010)

[35] Helmy, T. and Rasheed, Z.: Independent Job Scheduling by Fuzzy C-Mean Clustering and an Ant Optimization Algorithm a Computation Grid. IAENG International Journal of Computer Science, Vol.37, No.2 (2009)