A New Asymmetric Criterion for Cluster Validation

Hosein Alizadeh, Behrouz Minaei, and Hamid Parvin

Islamic Azad University, Mahdishahr Branch, Mahdishahr, Iran
{halizadeh,b_minaei,parvin}@iust.ac.ir

Abstract. In this paper a new criterion for clusters validation is proposed. Many stability measures to validate a cluster have been proposed such as Normalized Mutual Information. We propose a new criterion for clusters validation. The drawback of the common approach is discussed in this paper and then a new asymmetric criterion is proposed to assess the association between a cluster and a partition which is called Alizadeh-Parvin-Minaei criterion, APM. The APM criterion compensates the drawback of the common Normalized Mutual Information (NMI) measure. 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. The empirical studies show that the proposed method outperforms other ones.

Keywords: Clustering Ensemble, Stability Measure, Cluster Evaluation.

1 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 [6]. 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 [17]. 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 partitionings [11].

Fern and Lin [7] 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 diversity. The Sum of Normalized Mutual Information, SNMI [8]-[10] and [18] is used to measure the quality of an individual partition with respect to other partitions. Also, the Normalized Mutual Information, NMI, is employed for measuring 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

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 320–330, 2011.

multi objective data clustering method based on the selection of individual clusters produced by several clustering algorithms through an optimization procedure [14]. 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 [10] have offered a new clustering ensemble method which 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 these clusters. This method which is based on cluster stability evaluates the primary clustering results instead of final clustering.

Moller and Radke [16] 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 [5], 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 [3]. Inokuchi et al. [12] 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 [19]. This validity measure is applied to the determination of the number of clusters and also the evaluation of robustness of different partitionings. Das and Sil [4] 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.

We discuss the drawbacks of the common approaches and then have proposed a new asymmetric criterion to assess the association between a cluster and a partition which is called Alizadeh-Parvin-Minaei criterion, APM. The APM criterion compensates the drawbacks of the common method. Also, a clustering ensemble method is proposed which is based on aggregating a subset of primary clusters. This method uses the Average APM as fitness measure to select a number of clusters. The clusters which satisfy a predefined threshold of the mentioned measure are selected to participate in the clustering ensemble. To combine the chosen clusters, a co-association based consensus function is employed.

2 Proposed Method

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

The main idea of the proposed clustering ensemble method is to utilize a subset of the best performing primary clusters in the ensemble instead of all of them. It seems that every cluster does not have a good quality. So, in this method just those clusters which satisfy enough stability to participate in the combination are chosen. The cluster selection is done based on cluster stability which is defined according to Normalized Mutual Information, NMI.

The manner of computing stability is described in the following sections in detail. As seen in Fig 1, a subset of the most stable clusters is first selected for combination. This is simply done by applying a stability-threshold to each cluster. In the next step, the selected clusters are used to construct the co-association matrix. Several methods have been proposed for combination of the primary results [2] and [18]. In our work, some clusters in the primary partitions may be absent (having been eliminated by the stability criterion). Since the original EAC method [8] 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 the hierarchical single-link clustering to extract the final clusters from this matrix.

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 [13]. 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 [14].

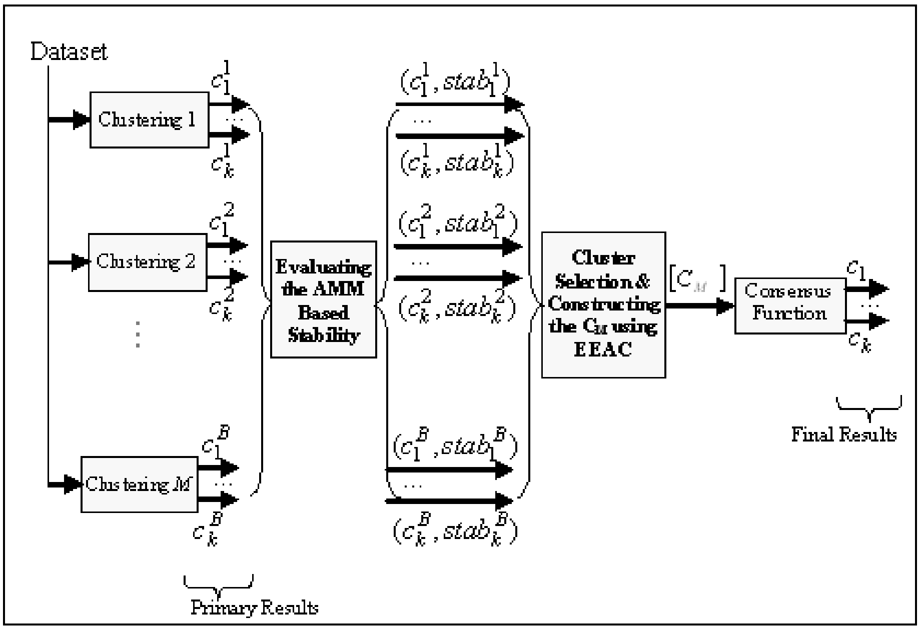

Fig. 1. Training phase of the Bagging method

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* [18]. However, the stability used in [14] 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 [9]:

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

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*.

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

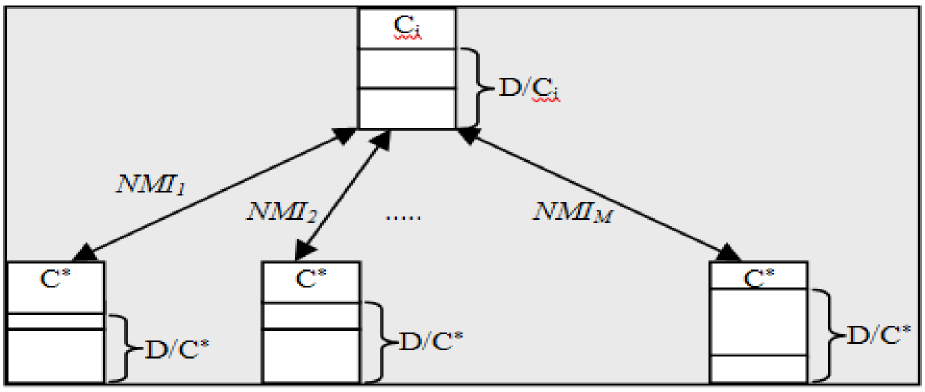

Fig. 2. Computing the Stability of Cluster C_i

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:

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

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

Here 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 3b. Fig 3a shows a true clustering. Since the well separated cluster in the top left corner is repeated several times (90% repetition) in partitionings 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.

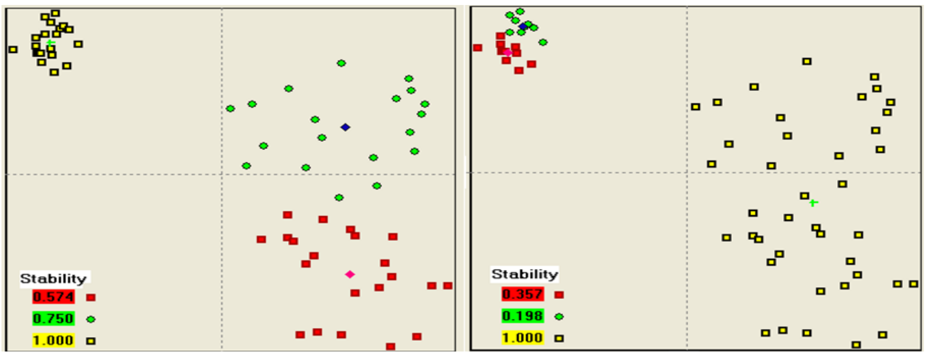

Fig. 3. Two primary partitions with k=3. (a) True clustering. (b) Spurious clustering.

Here, a new criterion is proposed which can solve this problem. Assume that the problem is evaluating the APM criterion for cluster C_1 in Fig. 4a with respect to clustering obtained in Fig. 4b.

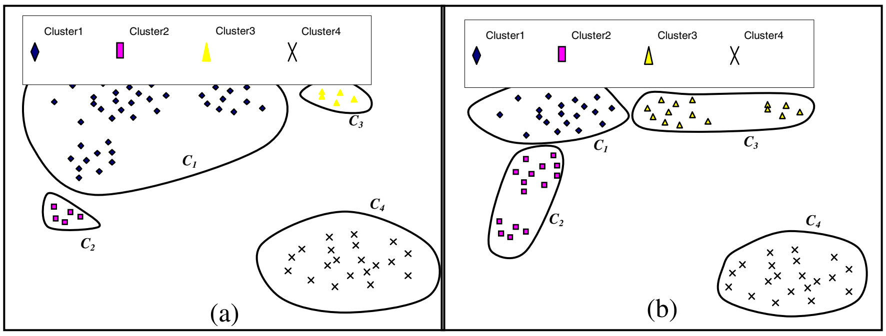

Fig. 4. evaluating the APM criterion for cluster C_1 from clustering (a) with respect to clustering (b), with k=4

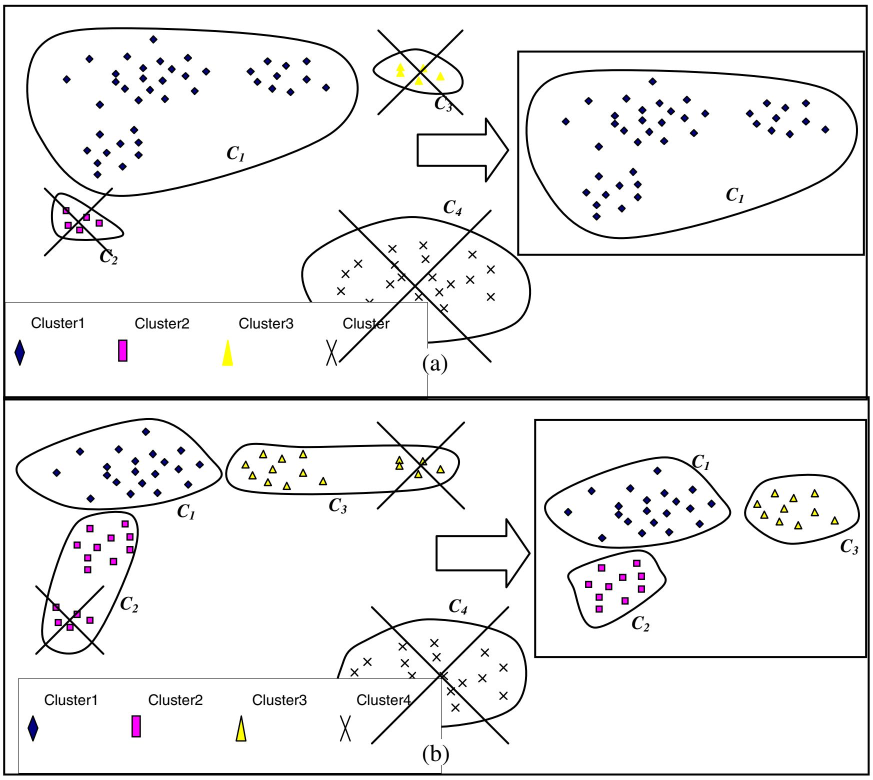

Fig. 5. Providing data for evaluating the APM criterion. (a) Deleting all other clusters except C_1 from P^a. (b) deriving P^{b*}, the corresponding samples of C_1 in P^b

The main idea in this method is to eliminate the symmetricalness which exists in NMI equation. In this approach, except the cluster C_1 all other clusters in P^a are taken out. Also, all clusters in P^b which are not included the samples of this cluster are eliminated. In the next step, the other samples which are not in C_1 of P^a, are removed from clusters in P^b (from the clusters which include some of these samples). This process is depicted in Fig. 5.

Now, the entropy between remained clusters in two partitions P^a and P^b is computed (see Fig. 6). On account of the other involved samples are eliminated, this criterion is not symmetric.

All the previous works are based on the NMI definition as equation 1. Even for evaluating the occurrence of a cluster in a partition, the problem is modified in some way to become the comparing problem between two partitions and then the NMI equation is used. In this paper, the problem is not changed according to definition of NMI; instead, the NMI equation is modified so that the occurrence of a cluster in a partition is computed. It is done by evaluating the entropy between the considered cluster and other pseudo clusters in the corresponding partition. In this paper the Alizadeh-Parvin-Moshki-Minaei criterion, APM, is defined between a cluster C_i from P^a and the partition P^{b*} from P^b, as below equation:

$$APMM(C_i^a, P^{b*}) = \frac{-2\log\left(\frac{n}{n_i^a}\right)\sum_{j=1}^{k_{b*}} n_j^{b*}}{n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}}{n}\right)} \quad (3)$$

where n is number of samples available in the cluster C_i and n_{ij}^{ab*} denotes the number of shared samples between the clusters $C_i^a \in P^a$ and $C_j^{b*} \in P^{b*}$. Also k_{b*} is the number of clusters in P^{b*}.

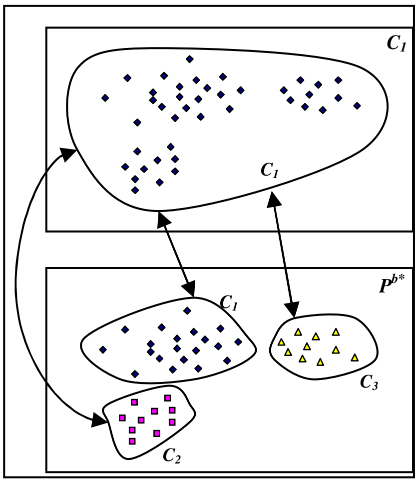

Fig. 6. Computing the entropy between the cluster C_1 from P^a and P^{b*} from P^b

Here, the Average APM, AAPM is proposed as a measure of stability of a primary cluster C_i with respect to the partitions available in the reference set as equation 4:

$$AAPMM(C_i) = \frac{1}{M} \sum_{j=1}^{M} APMM(C_i^a, P_j^{b*}) \tag{4}$$

where P_j^{b*} is from j-th partition of the reference set.

In the following step, the selected clusters are 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:

$$C(i, j) = \frac{n_{i,j}}{m_{i,j}} \tag{5}$$

where n_{ij} counts the number of clusters shared by objects with indices i and j in the partitions over the primary 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:

$$C(i, j) = \frac{n_{i,j}}{\max(n_i, n_j)} \tag{6}$$

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.

3 Experimental Results

This section reports and discusses the empirical studies. The proposed method is examined over 5 different standard datasets. 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 [15].

Table 1. Brief information about the used datasets

	Class	Features	Samples
Glass	6	9	214
Breast-C	2	9	683
Wine	3	13	178
Bupa	2	6	345
Yeast	10	8	1484

All experiments are done over the normalized features. It means each feature is normalized with mean of 0 and variance of 1, N(0, 1). All of them are reported over means of 10 independent runs of algorithm. The final performance of the clustering algorithms is evaluated by re-labeling between obtained clusters and the ground truth labels and then counting the percentage of the true classified samples. Table 2 shows the performance of the proposed method comparing with most common base and ensemble methods.

Table 2. Experimental results

	Simple Methods (%)				Ensemble Methods (%)			
Dataset	Single Linkage	Average Linkage	Complete Linkage	Kmeans	Kmeans Ensemble	Full Ensemble	Cluster Selection by NMI Method	Cluster Selection by max Method
Wine	37.64	38.76	83.71	96.63	96.63	97.08	97.75	**98.31**
Breast-C	65.15	70.13	94.73	95.37	95.46	95.10	95.75	**98.33**
Yeast	34.38	35.11	38.91	40.20	45.46	47.17	**47.17**	**47.17**
Glass	36.45	37.85	40.65	45.28	47.01	47.83	48.13	**50.47**
Bupa	57.68	57.10	55.94	54.64	54.49	55.83	58.09	**58.40**

The first four columns of Table 2 are the results of some base clustering algorithms. The results show that although each of these algorithms can obtain a good result over a specific dataset, it does not perform well over other datasets. For example, according to Table 2 the K-means algorithm has a good clustering result over Wine dataset in comparison with linkage methods. But, it has lower performance in comparison to linkage methods in the case of Bupa dataset. Also, the complete linkage has a good performance in Breast-Cancer dataset in comparison with others; however it is not in the case of all datasets. The four last columns show the performance of some ensemble methods in comparison with the proposed one. Taking a glance at the last four columns in comparison with the first four columns shows that the ensemble methods do better than the simple based algorithms in the case of performance and robustness along with different datasets. The first column of the ensemble methods is the results of an ensemble of 100 K-means which is fused by EAC method. The 90% sampling from dataset is used for creating diversity in primary results. The sub-sampling (without replacement) is used as the sampling method. Also the random initialization of the seed points of K-means algorithm helps them to be more diverse. The single linkage algorithm is applied as consensus function for deriving the final clusters from co-association matrix. The second column from ensemble methods is the full ensemble which uses several clustering algorithms for generating the primary results. Here, 70 K-means with the above mentioned parameters in addition to 30 linkage methods provide the primary results. The third column of *Ensemble Methods* is consensus partitioning using EEAC algorithm of top 33% stable clusters, employing NMI method as measure of stability. The fourth column of the ensemble methods is also consensus partitioning using EEAC algorithm of top 33% stable clusters, employing max method as measure of stability.

4 Conclusion and Future Works

In this paper a new clustering ensemble method is proposed which is based on a subset of total primary spurious clusters. 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 an alternative criterion is suggested which is named Alizadeh-Parvin-Moshki-Minaei, APM. The experiments show that the APM criterion does slightly better than NMI criterion generally; however it significantly outperforms the NMI criterion in the case of synthetic data sets. 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 data set. Another innovation of this paper 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. The empirical studies over several data sets robustly show that the quality of the proposed method is usually better than other ones.

References

1. Alizadeh, H., Minaei-Bidgoli, B., Parvin, H., Mohsen, M.: An Asymmetric Criterion for Cluster Validation. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2011). LNCS. Springer, Heidelberg (in press, 2011) ISSN: 0302-9743
2. Ayad, H., Kamel, M.S.: Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters. IEEE Trans. on Pattern Analysis and Machine Intelligence 30(1), 160–173 (2008)
3. Brandsma, T., Buishand, T.A.: Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling. Hydrology and Earth System Sciences 2, 195–209 (1998)
4. Das, A.K., Sil, J.: Cluster Validation using Splitting and Merging Technique. In: Int. Conf. on Computational Intelligence and Multimedia Applications, ICCIMA (2007)
5. Davison, A.C., Hinkley, D.V., Young, G.A.: Recent developments in bootstrap methodology. Statistical Science 18, 141–157 (2003)
6. Faceli, K., Marcilio, C.P., Souto, D.: Multi-objective Clustering Ensemble. In: Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS 2006) (2006)
7. Fern, X.Z., Lin, W.: Cluster Ensemble Selection. In: SIAM International Conference on Data Mining, SDM 2008 (2008)
8. Fred, A., Jain, A.K.: Data Clustering Using Evidence Accumulation. In: Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR 2002, Quebec City, pp. 276–280 (2002)
9. Fred, A., Jain, A.K.: Combining Multiple Clusterings Using Evidence Accumulation. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(6), 835–850 (2005)
10. Fred, A., Jain, A.K.: Learning Pairwise Similarity for Data Clustering. In: Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR 2006) (2006)
11. Fred, A., Lourenco, A.: Cluster Ensemble Methods: from Single Clusterings to Combined Solutions. Studies in Computational Intelligence (SCI), vol. 126, pp. 3–30 (2008)
12. Inokuchi, R., Nakamura, T., Miyamoto, S.: Kernelized Cluster Validity Measures and Application to Evaluation of Different Clustering Algorithms. In: IEEE Int. Conf. on Fuzzy Systems, Canada (July 16-21, 2006)
13. Lange, T., Braun, M.L., Roth, V., Buhmann, J.M.: Stability-based model selection. In: Advances in Neural Information Processing Systems, vol. 15. MIT Press (2003)

14. Law, M.H.C., Topchy, A.P., Jain, A.K.: Multiobjective data clustering. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 424–430 (2004)
15. Newman, C.B.D.J., Hettich, S., Merz, C.: UCI repository of machine learning databases (1998), http://www.ics.uci.edu/~mlearn/MLSummary.html
16. Möller, U., Radke, D.: Performance of data resampling methods based on clustering. Intelligent Data Analysis 10(2) (2006)
17. Roth, V., Lange, T., Braun, M., Buhmann, J.: A Resampling Approach to Cluster Validation. In: Intl. Conf. on Computational Statistics, COMPSTAT (2002)
18. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (2002)
19. Xie, X.L., Beni, G.: A Validity measure for Fuzzy Clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence 13(4), 841–846 (1991)