A Clustering Ensemble Based on a Modified Normalized Mutual Information Metric

Hamid Parvin, Behzad Maleki, and Sajad Parvin

Islamic Azad University, Nourabad Mamasani Branch, Mamasani Nourabad
Iranhamidparvin@mamasaniiau.ac.ir,
b.maleki@ut.ac.ir, s.parvin@iust.ac.ir

Abstract. It has been proved that ensemble learning is a solid approach to reach more accurate, stable, robust, and novel results in all data mining tasks such as clustering, classification, regression and etc. Clustering ensemble as a sub-field of ensemble learning is a general approach to improve the performance of clustering task. In this paper by defining a new criterion for clusters validation named Modified Normalized Mutual Information (MNMI), a clustering ensemble framework is proposed. In the framework first a large number of clusters are prepared and then some of them are selected for the final ensemble. The clusters which satisfy a threshold of the proposed metric are selected to participate in final clustering ensemble. For combining the chosen clusters, a co-association based consensus function is applied. Since the Evidence Accumulation Clustering (EAC) method can't derive the co-association matrix from a subset of clusters, Extended Evidence Accumulation Clustering (EEAC), is applied for constructing the co-association matrix from the subset of clusters. Employing this new cluster validation criterion, the obtained ensemble is evaluated on some well-known and standard datasets. 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.

Keywords: Clustering Ensemble, Stability Measure, Cluster Evaluation.

1 Introduction

Nowadays, usage of recognition systems has found many applications in almost all fields [15]-[28]. Many researches are done to improve their performance. Most of these algorithms have provided good performance for specific problem, but they have not enough robustness for other problems. Because of the difficulty that these algorithms are faced to, recent researches are directed to the combinational methods. Ensemble learning has been proved to be a solid way to reach more accurate and stable results in data mining. Classifier ensemble as a sub-field of ensemble learning is a general method to improve the performance of classification. At first glance, usage of ensemble learning in clustering sounds similar to the widely prevalent use of combining multiple classifiers to solve difficult classification problems, using techniques such as bagging and boosting.

R. Huang et al. (Eds.): AMT 2012, LNCS 7669, pp. 32–42, 2012.

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 [3], [4] and [10]. There are many applications that use clustering techniques to discover latent structures of data, such as data mining [11], information retrieval [2], image segmentation [9], linkage learning [15], and machine learning. In real-world problems, clusters can appear with different shapes, sizes, degrees of data sparseness, and degrees of separation. Clustering techniques require the definition of a similarity measure between patterns. Some of the most typical clustering algorithms include: (a) Hierarchical clustering algorithms build clusters based on distance connectivity, (b) Centroid clustering algorithms such as k-means algorithm represents each cluster by a single mean vector, and (c) Density clustering algorithms such as DBSCAN define clusters as connected dense regions in the data space.

DBSCAN (stands for Density-Based Spatial Clustering of Applications with Noise) is a data clustering algorithm proposed by Ester et al. [32]. It is named density-based clustering because it searches for some partitions beginning at the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms. Hierarchical clustering is another approach in clustering algorithms that seeks to build a hierarchy of partitions. It uses a number of the merge (or split) operators to reach the goal. The operators are employed in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram [33].

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 [8].

Fern and Lin [8] have offered a clustering ensemble framework that selects a few of the base partitionings to make a thinner but better ensemble than using all primary the base partitionings. The ensemble selection approach is designed based on quality and diversity, the only two factors that have been proven to effect cluster ensemble performance. Their method tries to select a subset of the base partitionings which simultaneously has both the highest quality and the most diversity. The Sum of Normalized Mutual Information, SNMI [5], [6] 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 [13]. 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 [7] 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.

Alizadeh et al. 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 [12], [31].

To evaluate a cluster, the NMI method has many weaknesses that are described in [31]. Alizadeh et al. propose another version of NMI named max method. They also show that the max method also has some drawbacks, so they propose another metric named APMM, which is first of their author names [12].

This paper proposes a new measure to evaluate a cluster in that it is desired to evaluate the average similarity of the cluster with other clusters by eliminating its complement. We employ this criterion to select the more robust clusters in the final ensemble. To aggregate the final partitionings into consensus partitioning, a number of well-known methods are employed to make a decisive conclusion.

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

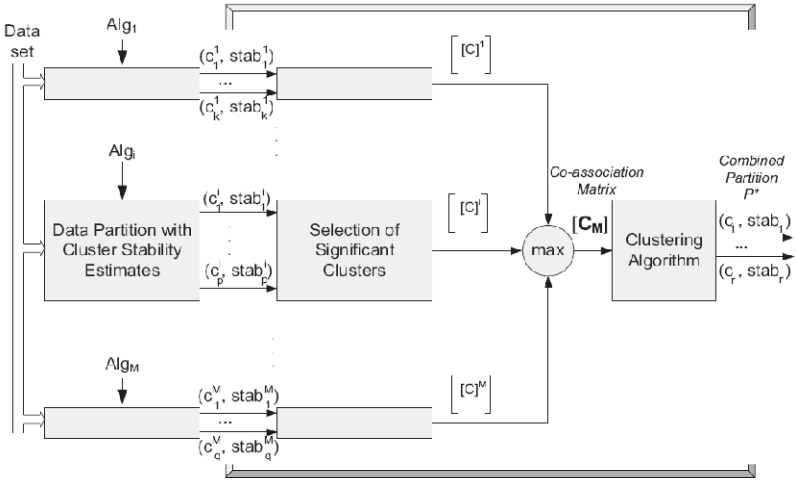

Fig. 1. Clustering Ensemble Framework

2 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 similar to Max and APMM [20] to utilize a subset of the best

performing primary clusters in the ensemble, rather than using all of clusters. Only the clusters which satisfy a stability criterion are better to participate in the consensus function.

The proposed framework is depicted in Fig 1. It has four steps. In the first step B partitionings are extracted out of dataset. The partitioning i is denoted by $partitioning_i$. The $partitioning_i$ is obtained by a k-means algorithm with a new initialization of the seed points. Note that the $partitioning_i$ is to extract $k(i)$ clusters out of dataset. Then each partitioning is broken in some distinct partitions (or clusters). It means $partitioning_i$ converted to $k(i)$ clusters denoted by $c_1^i, c_2^i, \ldots \; and \; c_{k(i)}^i$ respectively. After obtaining a pool of clusters, in the second step, a stability value is computed as a tag for each of them. The stability value of the cluster c_j^i is denoted by $stab_j^i$.

The manner of computing stability for each cluster is described in the sections 2.2 in more detail. A subset of stable clusters having a good diversity is selected by a thresholding scheme in the third step. This step is explained in detail in section 2.3. In the next step, the selected clusters are used to construct the consensus partitioning. This is done in two subparts: (a) to extract a co-association matrix from them (section 2.4) along with (b) a linkage clustering. Since the original EAC method [8] cannot truly identify the pairwise similarities between dataitems when there is only a subset of clusters, we use a method explained in [1] to construct the co-association matrix from the base selected clusters. This method is called EEAC: Extended Evidence Accumulation Clustering method. 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 HyperGraph Partitioning Algorithm (HGPA), Meta-Clustering Algorithm (MCLA) and Cluster-based Similarity Partitioning Algorithm (CSPA) [30].

In the first step B partitionings are extracted out of dataset by B independent runnings of the k-means algorithm. The $partitioning_i$ is obtained by the *i*-th running of the k-means algorithm with a new initialization of the seed points. To produce the diverse cluster as much as possible the k-means algorithms are run, aiming at extracting different number of clusters out of dataset. It means that the $partitioning_i$ extracts $k(i)$ clusters out of dataset. As it is mentioned the proposed method tries to select a subset of well-performing clusters (or equivalently partitions) instead of a subset of clusterings (or equivalently partitionings). So each partitioning is broken in some distinct partitions clusters (or equivalently partitions).

Second step is stability computation. Since the goodness of a cluster C_i is determined by all of 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 a measure of cluster goodness is used in [1], [13] and [20]. A stable cluster is the one that has a high likelihood of recurrence across multiple applications of a clustering algorithm. Stable clusters are usually preferable, since they are robust with respect to minor changes in the dataset [14].

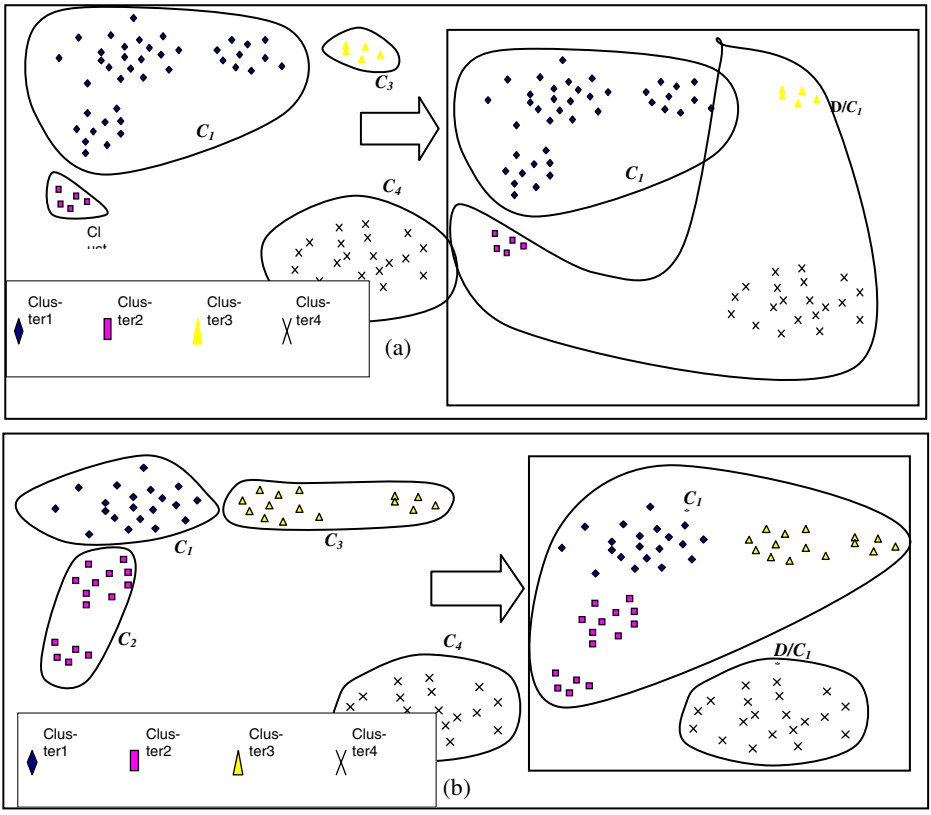

Fig. 2. Computing the stability of Cluster 1 of the partition in Fig. 2 (a) considering the partition in the Fig. 2 (b) of the reference set using NMI method

Now assume that the stability of cluster C_i is to be computed. In this method first a set of partitionings over dataset is provided which is called the reference set. One can consider the partitionings obtained in the first step as reference set for decreasing the runtime. In this notation D is dataset and $P_w(D)$ is a partitioning over D. Now, the problem is: "How many times is the cluster C_i repeated in the reference partitions?" Assume that the NMI between the cluster C_i and a reference partition $P_w(D)$ is denoted by $NMI(C_i, P_w(D))$. While the most of 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 after that by employing the common partition-to-partition *NMI*. 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 assume a second partition $P_2 = P^b = \{C_w^*, D/C_w^*\}$, where C_w^* denotes the union of all "positive" clusters in $P_w(D)$ and others are in D/C_w^*. A cluster C_r in $P_w(D)$ is positive cluster for C_i if more than half of its data points also belongs to C_i.

Now, define $NMI(C_i, P_w(D))$ by $NMI(P^a, P^b)$ which is calculated as [9]:

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

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. This method is named NMI method. Fig. 2 illustrates the NMI method.

After producing P_1, if we assume a second partition $P_2 = P^b = \{C_w^*\} \cup Cs_w^*$, where C_w^* denotes the same clusters in $P_w(D)$ defined by APM [1] and for each of other data we consider a cluster. The set of these clusters is denoted by Cs_w^*. Fig. 3 shows the method explained above which is named Edited APM, EAPM.

NMI_h in the paper shows the stability of cluster C_i with respect to the hth partition in reference set. The total stability of cluster C_i is defined as:

$$Stab(C_i) = \frac{\sum_{j=1}^{B} NMI_j}{B} \quad (2)$$

This procedure is applied for each cluster available in the pool clusters obtained in the first step. It means this procedure must be iterated q times, where q is computed as equation 3.

$$q = \sum_{i=1}^{B} k(i) \quad (3)$$

Third step is simply done be a thresholding. It means that the clusters with higher stability values are selected for next step and other are omitted.

In forth step, the selected clusters are used to produce final clusters in a co-association based model. In the step it is to construct the co-association matrix and then to apply a hierarchical clustering. To construct the co-association matrix from the selected clusters EEAC is employed. In the EAC method the m primary partitions from dataset are accumulated in a $n \times n$ co-association matrix. Each entry in this matrix is computed from equation 4.

$$C_{ij} = \frac{n_{ij}}{m_{ij}} \quad (4)$$

where m_{ij} counts the number of clusters shared by objects with indices i and j in the pool of all clusters obtained in the first step. It is worthy to note that the maximum possible value of m_{ij} computed as equation 3. Also n_{ij} is the number of partitions where this pair of objects is simultaneously present in the selected clusters. Note that the value of n_{ij} is at most as many as the number of selected clusters which is less than the value of m_{ij}.

3 Experimental Study

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

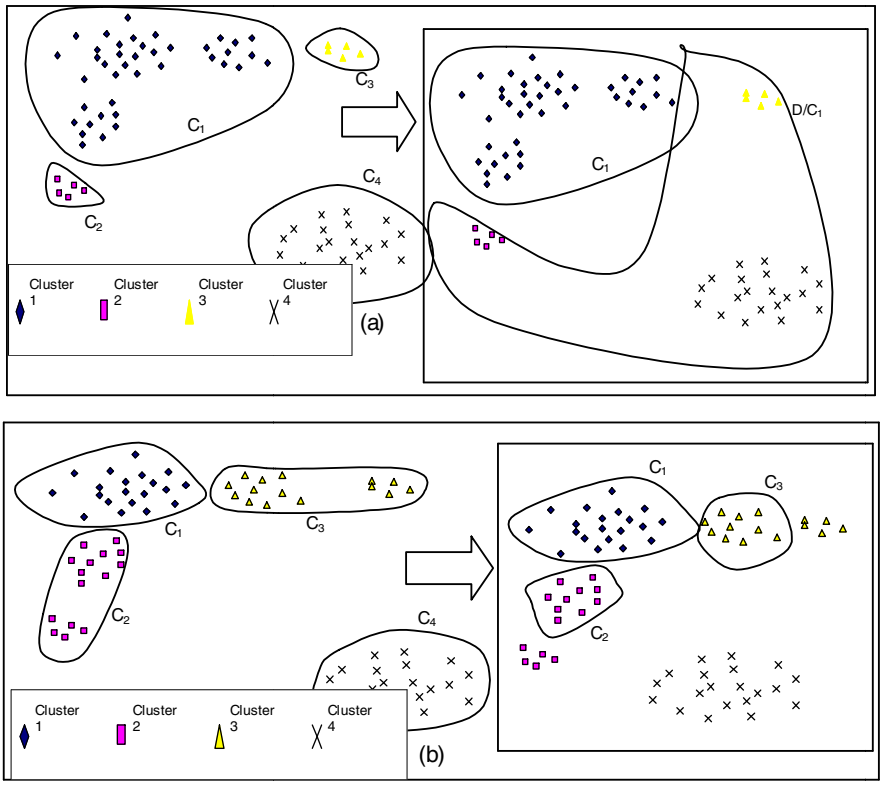

Fig. 3. Computing the stability of Cluster 1 of the partition in Fig. 3 (a) considering the partition in the Fig. 3 (b) of the reference set using EAPM method

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

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. More information is available in [14].

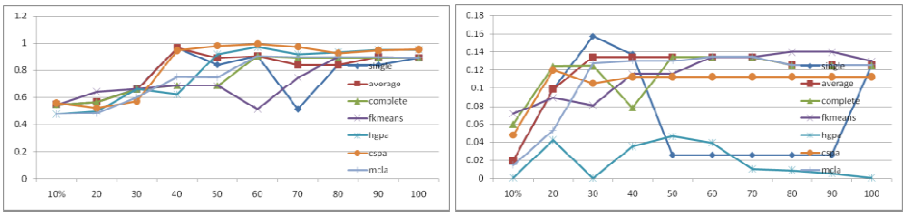

Fig. 4. The horizontal axis stands for the rate of stable clusters that are selected. The vertical axis stands for the NMI values between labels of Iris (left) and Ionosphere (right) datasets and the consensus partitions obtained by different consensus functions over the selected clusters.

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 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.

As it is inferred from the Fig. 4, the best ratio of selection of the stable clusters is 60% and the best option for consensus function is CSPA for Iris dataset.

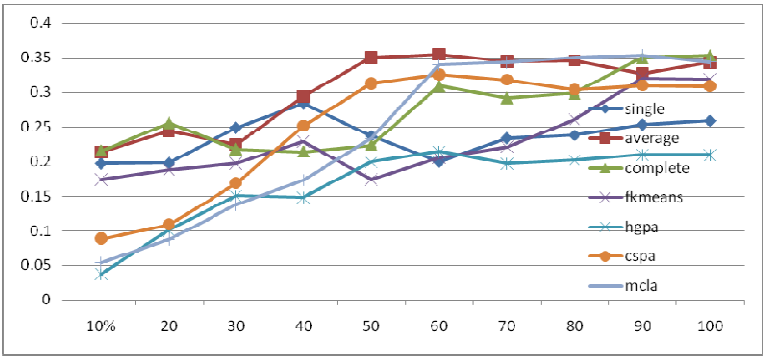

Fig. 5. The horizontal axis stands for the rate of stable clusters that are selected. The vertical axis stands for the averaged NMI values for all ten datasets.

Fig. 4 also 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.

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. 5. To make a general decisive conclusion, the results for all ten datasets are averaged and the final results are illustrated in the Fig. 5. The Averaged-Linkage consensus function over 50% of the most stable clusters generally reaches the maximum for all dataset.

Table 1 shows the performance of the proposed method comparing with most common base and ensemble methods.

Table 1. Experimental results

Metric Evaluation	Dataset										
	Breast Cancer	Iris	Bupa	SAHeart	Ionosphere	Glass	Halfrings	Galaxy	Yeast	Wine	Norm. Wine
NMI	95.73	76.13	54.33	63.36	70.60	**47.76**	74.48	**31.27**	42.93	69.38	85.17
MAX	96.49	84.87	**57.42**	**63.87**	57.75	44.35	74.55	29.85	51.27	70.00	94.44
APM	95.46	**90.00**	55.07	63.85	70.66	45.79	54.00	30.65	**53.10**	**70.23**	96.63
EAPM	**96.93**	88.67	54.78	63.20	**71.23**	43.93	**88.00**	30.65	50.47	**70.23**	**97.19**

4 Conclusion and Future Works

In this paper a new clustering ensemble method that is based on a subset of total primary spurious clusters is offered. Since the worth of the base partitions is not identical and also existence of a subset of them may yet result to a better performance, here an approach to choose a subset of more effective partitions is offered. A common metric based on that this subset is derived is normalized mutual information. Recently some drawbacks of NMI criterion are discussed and some alternative criterions, such as APM and Max, are proposed. In the paper while mentioning some drawbacks unhandled by APM and Max, a new metric that is named EAPM is proposed to solve the new raised drawbacks. The empirical studies over several datasets robustly show that the quality of the proposed method is usually better than other ones. The experiments confirm that the EAPM criterion does slightly better than NMI criterion generally; however it significantly outperforms the NMI criterion in the case of synthetic datasets. 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. The experiments also show that the EAPM criterion does better than Max and APM criterions.

References

1. 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)
2. Bhatia, S.K., Deogun, J.S.: Conceptual Clustering in Information Retrieval. IEEE Trans. Systems, Man, and Cybernetics 28(3), 427–536 (1998)
3. Dudoit, S., Fridlyand, J.: Bagging to improve the accuracy of a clustering procedure. Bioinformatics 19(9), 1090–1099 (2003)
4. Faceli, K., Carvalho, A.C.P.L.F., Souto, M.C.P.: Multi-objective Clustering Ensemble. In: Proceedings of the Sixth International Conference on Hybrid Intelligent Systems, HIS 2006 (2006)
5. 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)
6. Fred, A., Jain, A.K.: Combining Multiple Clusterings Using Evidence Accumulation. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(6), 835–850 (2005)
7. Fred, A., Jain, A.K.: Learning Pairwise Similarity for Data Clustering. In: Proc. of the 18th Int. Conf. on Pattern Recognition, ICPR 2006 (2006)
8. Fred, A., Lourenco, A.: Cluster Ensemble Methods: from Single Clusterings to Combined Solutions. SCI, vol. 126, pp. 3–30 (2008)
9. Frigui, H., Krishnapuram, R.: A Robust Competitive Clustering Algorithm with Applications in Computer Vision. IEEE Trans. Pattern Analysis and Machine Intelligence 21(5), 450–466 (1999)
10. Jain, A.K., Murty, M.N., Flynn, P.: Data clustering: A review. ACM Computing Surveys 31(3), 264–323 (1999)
11. Judd, D., Mckinley, P., Jain, A.K.: Large-Scale Parallel Data Clustering. IEEE Trans. Pattern Analysis and Machine Intelligence 19(2), 153–158 (1997)

12. Alizadeh, H., Minaei-Bidgoli, B., Parvin, H.: A New Asymmetric Criterion for Cluster Validation. In: San Martin, C., Kim, S.-W. (eds.) CIARP 2011. LNCS, vol. 7042, pp. 320–330. Springer, Heidelberg (2011)
13. Law, M.H.C., Topchy, A.P., Jain, A.K.: Multiobjective data clustering. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, Washington, D.C., vol. 2, pp. 424–430 (2004)
14. Newman, C.B.D.J., Hettich, S., Merz, C.: UCI repository of machine learning databases (1998), http://www.ics.uci.edu/~mlearn/MLSummary.html
15. Parvin, H., Minaei-Bidgoli, B., Alinejad, H.: Linkage Learning Based on Differences in Local Optimums of Building Blocks with One Optima. International Journal of the Physical Sciences, IJPS, 3419–3425 (2011)
16. Daryabari, M., Minaei-Bidgoli, B., Parvin, H.: Localizing Program Logical Errors Using Extraction of Knowledge from Invariants. In: Pardalos, P.M., Rebennack, S. (eds.) SEA 2011. LNCS, vol. 6630, pp. 124–135. Springer, Heidelberg (2011)
17. Minaei-Bidgoli, B., Parvin, H., Alinejad-Rokny, H., Alizadeh, H., Punch, W.F.: Effects of resampling method and adaptation on clustering ensemble efficacy, Online (2011)
18. Fouladgar, H., Minaei-Bidgoli, B., Parvin, H.: On Possibility of Conditional Invariant Detection. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part II. LNCS, vol. 6882, pp. 214–224. Springer, Heidelberg (2011)
19. Parvin, H., Minaei-Bidgoli, B.: Linkage Learning Based on Local Optima. In: Jędrzejowicz, P., Nguyen, N.T., Hoang, K. (eds.) ICCCI 2011, Part I. LNCS, vol. 6922, pp. 163–172. Springer, Heidelberg (2011)
20. Parvin, H., Helmi, H., Minaei-Bidgoli, B., Alinejad-Rokny, H., Shirgahi, H.: Linkage Learning Based on Differences in Local Optimums of Building Blocks with One Optima. International Journal of the Physical Sciences 6(14), 3419–3425 (2011)
21. Qodmanan, H.R., Nasiri, M., Minaei-Bidgoli, B.: Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. Expert Systems with Applications 38(1), 288–298 (2011)
22. Parvin, H., Minaei-Bidgoli, B., Alizadeh, H.: A New Clustering Algorithm with the Convergence Proof. In: König, A., Dengel, A., Hinkelmann, K., Kise, K., Howlett, R.J., Jain, L.C. (eds.) KES 2011, Part I. LNCS, vol. 6881, pp. 21–31. Springer, Heidelberg (2011)
23. Parvin, H., Minaei, B., Alizadeh, H., Beigi, A.: A Novel Classifier Ensemble Method Based on Class Weightening in Huge Dataset. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part II. LNCS, vol. 6676, pp. 144–150. Springer, Heidelberg (2011)
24. Parvin, H., Minaei-Bidgoli, B., Alizadeh, H.: Detection of Cancer Patients Using an Innovative Method for Learning at Imbalanced Datasets. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 376–381. Springer, Heidelberg (2011)
25. Parvin, H., Minaei-Bidgoli, B., Ghaffarian, H.: An Innovative Feature Selection Using Fuzzy Entropy. In: Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H. (eds.) ISNN 2011, Part III. LNCS, vol. 6677, pp. 576–585. Springer, Heidelberg (2011)
26. Parvin, H., Minaei, B., Parvin, S.: A Metric to Evaluate a Cluster by Eliminating Effect of Complement Cluster. In: Bach, J., Edelkamp, S. (eds.) KI 2011. LNCS, vol. 7006, pp. 246–254. Springer, Heidelberg (2011)
27. Parvin, H., Minaei-Bidgoli, B., Ghatei, S., Alinejad-Rokny, H.: An Innovative Combination of Particle Swarm Optimization, Learning Automaton and Great Deluge Algorithms for Dynamic Environments. International Journal of the Physical Sciences 6(22), 5121–5127 (2011)

28. Parvin, H., Minaei, B., Karshenas, H., Beigi, A.: A New N-gram Feature Extraction-Selection Method for Malicious Code. In: Dobnikar, A., Lotrič, U., Šter, B. (eds.) ICANNGA 2011, Part II. LNCS, vol. 6594, pp. 98–107. Springer, Heidelberg (2011)
29. Roth, V., Lange, T., Braun, M., Buhmann, J.: A Resampling Approach to Cluster Validation. In: Intl. Conf. on Computational Statistics, COMPSTAT (2002)
30. Strehl, A., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research 3, 583–617 (2002)
31. Alizadeh, H., Minaei, B., Parvin, H.: A New Criterion for Clusters Validation. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) EANN/AIAI 2011, Part II. IFIP AICT, vol. 364, pp. 110–115. Springer, Heidelberg (2011)
32. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press (1996)
33. Sibson, R.: SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal 16(1), 30–34 (1973)