Adv. Eng. Tec. Appl. Vol. 1 No.1 (2012) 1-7

Advanced Engineering Technology and Application

Eliminating Effect of Complement Cluster

[1]Zahra Rezaei, [2]Sajad Parvin, [3]Mirsaeid Hosseini Shirvani

[1]Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran
[2]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
[3]Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran

Received: Jul 8, 2011; Revised Oct. 4, 2011; Accepted Jan. 6, 2012
Published online: 1 May 2012

Abstract: In this paper a new criterion for clusters validation is proposed. This new cluster validation criterion is used to approximate the goodness of a cluster. A clustering ensmble framework based on the new metric is proposed. The main idea behind the framework is to extract the most stable clusters in terms of the defined criteria. After extracting a large number of clusters some of them are selected for final ensmble. 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. 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

Keywords: Clustering Ensemble, Stability Measure, Extended EAC, Cluster Evaluation, Selecting Scheme.

1 Introduction

There are many applications which use clustering techniques for discovering structure in data, such as data mining (Fred and Jain, 2006), information retrieval (Ayad and Kamel, 2008), image segmentation (Fred and Jain, 2005), and machine learning. In real-world problems, clusters can appear with different shapes, sizes, data sparseness, 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 (Law et al, 2004). 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 (Fred and Jain, 2002).

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. There are many applications which use clustering techniques for discovering structure in data, such as data mining, information retrieval, image segmentation, and machine learning. In real-world problems,

clusters can appear with different shapes, sizes, data sparseness, 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. Clustering has been considered a very challenging problem in Data Mining due to its lack of supervision. It is desired to partition data in such a way that the data points that belong to a cluster have maximum similarities while the data points that belong to different clusters have minimal similarities (Faceli et al, 2006). 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 (Roth et al, 2002) Because of the difficulty of the problem and the weaknesses of primary clustering, the researches' direction has turned to clustering ensemble. Cluster ensemble methods attempt to find a better and more robust clustering solution by fusing information from several primary data partitionings (Fred and Lourenco, 2008).

Fern and Lin (Fern and Lin, 2008)) 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 (Strehl and Ghosh, 2002) 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 multi objective data clustering method based on the selection of individual clusters produced by several clustering algorithms through an optimization procedure. 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 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.

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 (Alizadeh et al, 2011).

To evaluate a cluster, the NMI method has many weaknesses that are described in. 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 (Alizadeh et al, 2011).

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.

A large number real standard dataset from UCI repository (Newman et al, 1998) are used as benchmarks and it is shown that the proposed metric is very effective.

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 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 cluster stability is defined according to NMI.

The proposed framework 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 $c_1^i, c_2^i, \dots \; and \; c_{k(i)}^i$ 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 $stab_j^i$. 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 cannot truly identify the pairwise similarities between dataitems when there is only a subset of clusters, we use a method explained in (Alizadeh et al, 2011) to construct the co-association matrix from the base selected clusters. This method is called EEAC. The hierarchical single-link clustering is done along with the extraction of the co-association matrix extract the consensus clusters.

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).

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 (Alizadeh et al, 2011; Lange et al, 2003. 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.

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 (Strehl and Ghosh, 2002), however, the stability used in 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.

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^*. Figure 1 shows the method explained above which is named Edited APM, EAPM.

NMI_h in Figure 2 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:

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

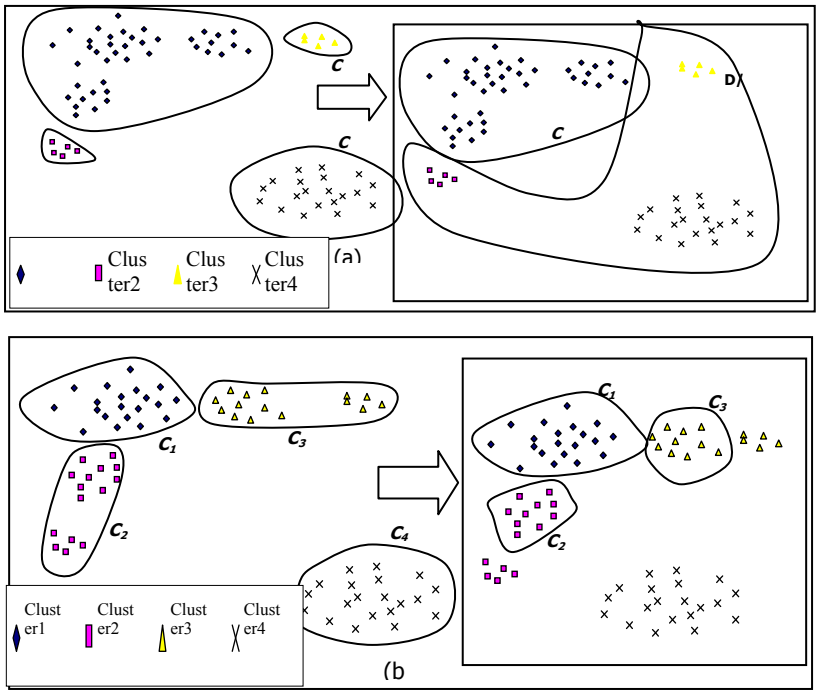

Figure 1. Computing the stability of Cluster 1 of the partition in Figure 1

(a) considering the partition in the Figure 1 (b) of the reference set using EAPM method

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.

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

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

In Consensus Function and Obtaining Final Partition 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

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 (Newman et al, 1998).

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.

The results show that although each of the metrics can obtain a good result over a specific dataset, it does not perform well over other datasets. For example, according to Table 1 the ensmble based on NMI obtains a good clustering result over Glass dataset. But, it has lower performance in comparison to results of ensmble based on other metrics in the case of Bupa dataset. The results of the ensemble methods are the results of an ensemble of 100 K-means which are 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 top 33% stable clusters are employed in constructing co-association matrix.

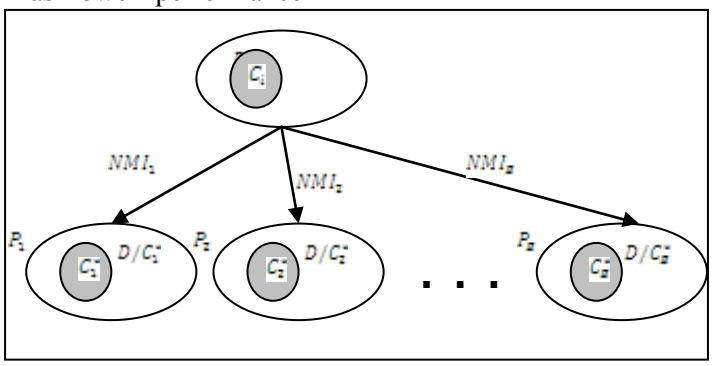

Figure 2. Computing the Stability of Cluster C_i considering a reference set

Table 1. Experimental results.

Metric Evaluation	Dataset										
	N. Breast Cancer	Iris	N. Bupa	N. SAHeart	Ionosphere	N. Glass	Halfrings	N. Galaxy	N. Yeast	Wine	N. 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**

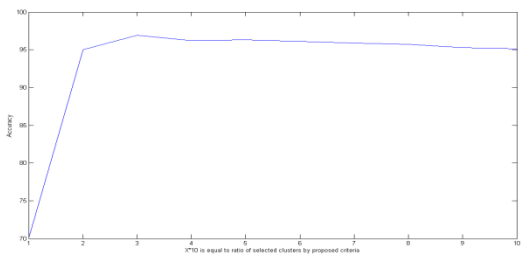

Figure 3. Accuracy in terms of different ratios of selected clusters by proposed criteria over Breast-C dataset.

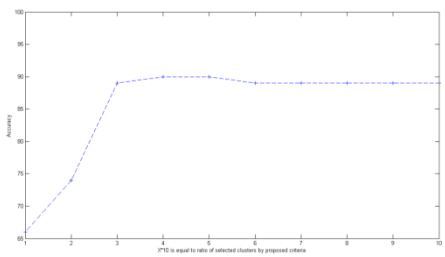

Figure 4. Accuracy in terms of different ratios of selected clusters by proposed criteria over Iris dataset.

To better understand the effect of proposed clustering ensemble framework, consider Figure 3 which is different accuracies of the consensus partitions obtained out of different ratios of the most stable clusters in Breast-C dataset. In Figure 3, the different size of the most stable clusters in terms of max metric are selected to participate in final ensemble. The accuracy of consensus partition extracted out of the selected clusters is presented in vertical axis. As it is obvious participating 20~30% of total clusters in the final ensemble is a very promising option. Also participation all clusters is not a good option. Figure 4 is the same results of Figure 3, but for Iris dataset.

7. Conclusion and Discussion

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 methods 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] Alizadeh, H., Minaei-Bidgoli, B. and Parvin, H. (2011). An Asymmetric Criterion for Cluster Validation. 16th Iberoamerican Congress on Pattern Recognition (CIARP 2011), LNCS, ISSN: 0302-9743. Springer, Heidelberg.

[2] 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, 30(1), pp: 160-173.

[3] Alizadeh, H., Minaei-Bidgoli, B. and Parvin, H. (2011). A New Criterion for Clusters Validation. Artificial Intelligence Applications and Innovations (AIAI 2011), LNCS, ISSN: 0302-9743. Springer, Heidelberg.

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

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

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

[7] 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.

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

[9] 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).

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

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

[12] 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.

[13] 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, 2, p: 424–430.

[14] Newman, C.B.D.J., Hettich, S., Merz ,C. (1998). UCI repository of machine learning databases.