SELECTIVELY LEARNING CLUSTERS IN MULTI-EAC

André Lourenço[1,2] and Ana Fred[2]
[1] *Instituto Superior de Engenharia de Lisboa, Lisboa, Portugal*
[2] *Instituto de Telecomunicações, Instituto Superior Técnico, Lisboa, Portugal*

Keywords: Cluster analysis, Evidence accumulation clustering (EAC), Multiple-criteria evidence accumulation clustering (Multi-EAC).

Abstract: The Multiple-Criteria Evidence Accumulation Clustering (Multi-EAC) method, is a clustering ensemble approach with an integrated cluster stability criterion used to selectively learn the similarity from a collection of different clustering algorithms. In this work we analyze the original Multi-EAC criterion in the context of the classical relative validation criteria, and propose alternative cluster validation indices for the selection of clusters based on pairwise similarities. Taking several clustering ensemble construction strategies as context, we compare the adequacy of each criterion and provide guidelines for its application. Experimental results on benchmark data sets show the proposed concepts.

1 INTRODUCTION

A recent trend in clustering, that constitutes the state-of-the art in the area, are clustering combination techniques (also called clustering ensemble methods) (Fred, 2001; Fred and Jain, 2005; Strehl and Ghosh, 2002; Fern and Brodley, 2004; Topchy et al., 2005; Ayad and Kamel, 2008). These methods attempt to find better and more robust partitioning of the data by combining the information of a set of N different partitions, the *clustering ensemble* - $\mathbb{P}$.

In (Fred and Jain, 2006) the authors proposed the Multi-Criteria Evidence Accumulation Clustering (Multi-EAC) method as an extension to the EAC framework (Fred, 2001; Fred and Jain, 2005), filtering the cluster combination process using a cluster stability criterion. Instead of using the information of the different partitions, it is assumed that, since algorithms can have different levels of performance in different regions of the space, only certain clusters should be considered. This algorithm selectively decides on the expertise level of each algorithm over the several regions of the space, highest local expertise overriding less confident decisions, and low expertise decisions being totally ignored in the combination process.

In this paper we focus on the shift of paradigm - from partition to cluster level validation. We propose to further explore the stability criteria proposed in the framework of the Multi-EAC, contextualizing it on the classical relative validation criteria.

Using this motivation we propose some cluster validation criteria, which validate each of the clusters in a data partition. With the proposed criteria we go beyond the classical problem of partition validation, focusing our goal in evaluating the quality of individual clusters. Furthermore we adapt these criteria to pairwise similarities, evaluating them in the context of the Multi-EAC algorithm.

The remainder of this document is organized as follows. In the next section we review the classical clustering validity criteria used to assess the quality of partitions relative to each other, contextualizing the present work on previous works. In section 3 we briefly review the clustering combination techniques and the Multi-EAC paradigm. We propose new criteria for cluster validation, in section 4, shifting the paradigm from partition to individual cluster analysis. The proposed formulation is based on pairwise similarities instead of feature-based object representations. In section 5 we apply the proposed indices in the context of Multi-EAC method. In section 6 we present the experimental setup and results over benchmark and real data-sets. Finally in section 7 we draw some conclusions and present future work.

Lourenço A. and Fred A..
SELECTIVELY LEARNING CLUSTERS IN MULTI-EAC.
DOI: 10.5220/0003099904910499
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 491-499
ISBN: 978-989-8425-28-7

2 RELATED WORK

The problem of evaluation/comparison of clustering results as well as deciding the number of clusters better fitting the data is fundamental in cluster analysis and it has been subject of many research studies (Jain and Dubes, 1988; Dubes and Jain, 1979; Levine and Domany, 2000; Halkidi et al., 2002a; Halkidi et al., 2002b; Verma and Meila, 2003).

In the literature there are various methods for quantitative evaluation of clustering results known under the name of *clustering validation* or *evaluation techniques*. Measures or indices of cluster validity are classically divided into three types: A) External; B) Internal; C) Relative. For more details consult (Jain and Dubes, 1988; Sergios Theodoridis, 1999; Halkidi et al., 2002a).

In this paper we evaluate the clustering structure, focusing on *Relative criteria* and cross-validation.

The basic idea of *Relative criteria* is the evaluation of the clustering structure by comparing it to other clustering schemes, resulting from the same algorithm but with different parameter values, or from other algorithms. Many different indices have been proposed in the literature, such as the Silhouette (Rousseeuw, 1987), Dunn's Index, Davies and Bouldin (Jain and Dubes, 1988) (Halkidi et al., 2001). Most of them are geometrical motivated and estimate how compact and well separated the clusters are.

Our approach builds over this indices but instead of validating all the partition, we consider each cluster independently.

Other approaches to clustering validation include variants of cross-validation (Jain and Moreau, 1987; Levine and Domany, 2000; Ben-Hur et al., 2002; Roth et al., 2002), where, unlike classical cross-validation (used in supervised classification), no class information is required. Several related procedures have also been proposed for inferring the number of clusters (Lange et al., 2002). These techniques focus on analyzing the stability of the solution and can be categorized in the internal validity criteria for clustering validation, not requiring any kind of *a priori* information. Methods adopting this approach use different resampling schemes to simulate perturbations over the original data-set, so as to assess the stability of the clustering results with respect to sampling variability. Partitioning results that are more robust (i.e., with minor cluster assignment changes) with respect to the data perturbation are chosen as the most consistent and better solutions.

The Multi-EAC uses this last class of techniques, using sub-sampling to assess the stability of pairwise co-associations.

3 CLUSTERING ENSEMBLE METHODS AND MULTI-EAC

Clustering ensemble methods can be decomposed into a cluster generation mechanism and a partition integration process, both influencing the quality of the combination results. Different approaches have been followed for the production of the clustering ensemble, involving different proximity measures, algorithms, initializations, and features spaces. For the combination process there are also diverse approaches, from graph-based (Strehl and Ghosh, 2002; Fern and Brodley, 2004), voting or statistical perspectives (Fred, 2001; Fred and Jain, 2005; Topchy et al., 2005; Ayad and Kamel, 2008). Overall, the several methods equally weight clusters belonging to an individual partition, as schematically plotted in figure 1.

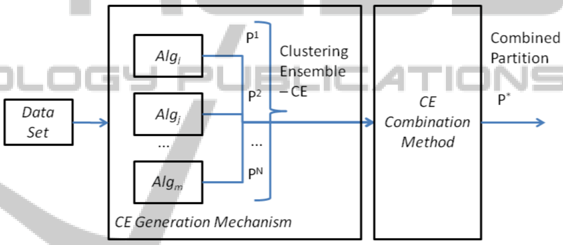


Figure 1: Schematic description of a clustering ensemble method.

Fred and Jain (Fred and Jain, 2005; Fred, 2001) proposed a statistical method of combination, the Evidence Accumulation Clustering (EAC), based on co-occurrences of pairs of objects in the same clusters, which can be interpreted as pairwise votes, over the N different partitions of the clustering ensemble. This is a robust combination method that additionally to the combined partition, P^*, produces as intermediate result a learned pairwise similarity, summarizing maximum likelihood estimates of pairwise co-occurence probabilities, as assessed from the clustering ensemble (see figure 2).

In (Fred and Jain, 2006) the authors proposed Multi-Criteria Evidence Accumulation Clustering (Multi-EAC) method as an extension of the EAC framework, filtering the cluster combination process (see figure 3) using a cluster stability criterion. Instead of using all the pairwise co-associations, it is assumed that, since algorithms can have different levels of performance in different regions of the space, only certain pairwise co-associations should be considered. Its selection is assessed through a stability criterion based on subsampling.

The method relies on two stages: the first stage

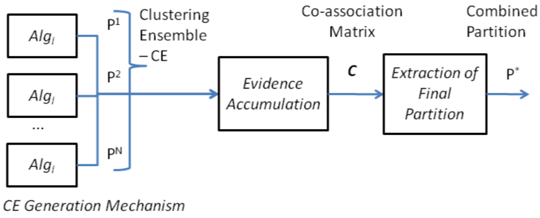


Figure 2: The main phases in the EAC method. The data partitions of the Clustering Ensemble (CE) are combined accumulating the pairwise co-associations over the co-association matrix, $\mathcal{C}$.

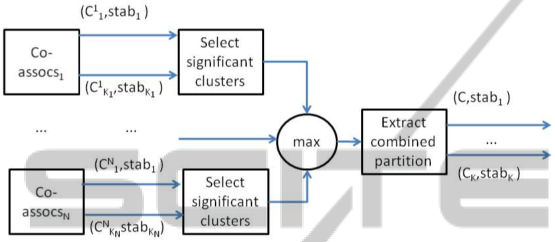


Figure 3: Evidence Accumulation under the Multi-EAC framework: filtering the cluster combination process.

aims at learning, for each algorithm, what are the regions where it performs well, that is, it selective learns clusters based on a sub-sampling and ensemble combination approach; the second stage integrates this information, combining the selected clusters.

Lets us define *clustering algorithm instantiation* the tuple formed by a clustering algorithm and a particular instantiation of the corresponding algorithmic parameters. Hereafter, for simplicity, we refer by algorithm l, denoted by Alg^l, one such clustering algorithm instantiation.

The identification of the regions where a particular algorithm, Alg^l, performs well is based on stability analysis over combination results obtained by applying the EAC method (Fred and Jain, 2005) on perturbed versions of the data-set, obtained through sub-sampling.

Let $X = \{x_1, x_2, \ldots, x_{n_s}\}$ represent a data-set with n_s observations, and X^{sub_i} be a sub-sampling version of the data set, by randomly selecting (without reposition) b samples from X, ($b < n_s$). In total, there are $\binom{n_s}{b}$ different sub-sampling versions of the data-set.

Each X^{sub_i} is clustered using Alg^l, yielding the partition P_i^l. Given N such perturbed versions of the data set, through algorithm Alg^l we obtain a clustering ensemble CE^l. These partitions are combined using the EAC method, accumulating the pairwise evidence over the co-association matrix, $\mathcal{C}^l$, whose entries are defined by:

$$\mathcal{C}^l(i,j) = n_{ij}/m_{ij}, \qquad (1)$$

where n_{ij} is the number of co-associations of the pair (i,j), and m_{ij} the number of subsampling experiments where this pair is present. Note that the subscript l indicates that this matrix was obtained according to algorithm Alg^l. For extracting clusters obtained by this algorithm, any clustering algorithm that accepts a similarity matrix (the co-association matrix) as input can be applied. Typically the Single Link or Average Link agglomerative hierarchical methods (Jain and Dubes, 1988) are used.

In order to identify the different levels of local performance of the algorithm, the stability of each of the extracted cluster is computed using the co-association values of that cluster (denoted by C_k), according to:

$$stab_{C_k} = \sum_{i,j \in C_k, j \neq i} \frac{\mathcal{C}(i,j)}{(n_k)(n_k - 1)}, \qquad (2)$$

where n_k is the number of objects in C_k, corresponding to average pairwise stability of cluster C_k.

Only clusters having stability values higher than a specified threshold, *th*, are **selected**; thus the method selectively learns pairwise similarities, over each ensemble, that is over each Alg^l.

In the second stage of the algorithm, the selected clusters are further combined into a global similarity matrix using a max rule. Each selected cluster, represented in a $n_k \times n_k$ co-association matrix, is combined with the other clusters selecting the sub-matrices with more stable values (max rule), thus joining all the individual contributions.

4 THE PROPOSED FRAMEWORK: CLUSTER VS PARTITION VALIDITY CRITERIA

While classical validity indexes are designed to measure the overall quality of data partitions, our goal is to define indices that measure the quality of individual clusters within a partition, and amongst distinct partitions. Herein we propose several indices to address this problem reviewing first the classical ones.

Most of the relative indices use geometrical considerations, and are based on a dissimilarity matrix computed over the original feature space, such as the Euclidean distance. Let $d(i,j)$ represent the dissimilarity between objects i and j.

The Silhouette index(Rousseeuw, 1987) judges the quality of a clustering solution by quantifying the compactness/separability of clusters. For each object in a cluster, the ”Silhouette value”, measures the degree to which a sample belongs to its current cluster relative to the other K-1 clusters. For each, $x_i \in C_a$, let $a(i)$ denote the average distance between the object and all other objects in C_a, and $b(i)$ denote the average distance between x_i and all objects in the nearest competing cluster C_b:

$$a(i) = \frac{1}{|C_a|-1} \sum_{j \in C_a, i \neq j} d(i,j) \tag{3}$$

$$b(i) = \min_{b \neq i} \left\{ \frac{1}{|C_b|} \sum_{j \in C_b} d(i,j) \right\} \tag{4}$$

The silhouette width for each x_i is computed using:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{5}$$

A global silhouette width can be computed taking the mean of the silhouette width for all samples in the data-set:

$$S = \frac{1}{n_s} \sum_{i=1}^{ns} s(i) \tag{6}$$

To measure the compactness of a cluster, C_k, we propose to measure the average dissimilarity of the pairs of samples within that cluster:

$$A_k = \frac{1}{(n_k)(n_k-1)} \sum_{i,j \in C_k, j>i} d(i,j) \tag{7}$$

To measure the separability between C_k and its nearest cluster C_l, we compute the average dissimilarity between pairs of objects in C_k and the objects in the competing cluster C_l:

$$B_k = \min_{1 \leq l \leq K\ ,\ l \neq k} \frac{1}{(n_k)(n_l)} \sum_{i \in C_k} \sum_{j \in C_l} d(i,j) \tag{8}$$

We define the **Cluster Silhouette** (Bolshakova and Azuaje, 2003), as:

$$S_k = \frac{B_k - A_k}{\max\{A_k, B_k\}} \tag{9}$$

This produces a score in the range $[-1,1]$, indicating how good an individual cluster is within its partition. A value close to 1 indicates that the cluster is compact and separated from the other clusters; a value closer to 0 suggest that the cluster is not so compact or separated from the nearest clusters, while a negative value suggest that is likely that the clusters has been incorrectly assigned.

Dunn's index (Dunn, 1974) (Bezdek and Pal, 1995) quantifies how compact and well separated clusters are, being defined as:

$$D = \min_{1 \leq q \leq K} \left\{ \min_{q+1 \leq r \leq K} \left(\frac{dist(C_q, C_r)}{\max_{1 \leq p \leq K} diam(C_p)} \right) \right\} \tag{10}$$

where $dist(C_q, C_r)$ represents the distance between the q-th and the r-th cluster, and $diam(C_p)$ is the p-th cluster diameter, as defined by:

$$dist(C_q, C_r) = \min_{i \in C_q, j \in C_r} d(i,j), \tag{11}$$

$$diam(C_p) = \max_{i,j \in C_p} d(i,j), \tag{12}$$

In order to compare different clusters we propose to compute the **Dunn's cluster index**, for cluster C_k, which can be defined as:

$$D_k = \frac{\min_{1 \leq r \leq K, r \neq k} dist(C_k, C_r)}{diam(C_k)} \tag{13}$$

This validation index can be used to compare different clusters and if $D_k > 1$ this is indicative that the cluster C_k is compact and separated from the other clusters.

5 CLUSTER SELECTION CRITERIA IN MULTI-EAC

The Multi-EAC algorithm relies on the identification of clusters selected from the extracted solutions of the ensembles produced by the different algorithms, based on co-association values, $\mathcal{C}$. In (Fred and Jain, 2006) the selection of clusters was performed according to equation 2. The idea behind this test is to measure the mean evidence of co-association of pairs of samples over the subsampling experiments performed over the data-set, using one particular algorithm. If pairs of samples have been co-associated by the clustering algorithm over these subsampling versions of the data-set, then these samples should be selected as stable associations. This value can be interpreted as an intra-cluster stability criteria, since only samples belonging to the same cluster are used. We will reformulate the previous index in the context of the cluster validity criteria.

In previous section, indexes are defined based on dissimilarities, the following are adaptations using the similarities represented in the co-association matrix $\mathcal{C}$.

To measure the compactness of a cluster C_k, the A_k term (equation 7) can be reformulated, using the intra-cluster similarity, as:

$$A_k^s = \frac{1}{(n_k)(n_k-1)} \sum_{i,j \in C_k, j>i} \mathcal{C}(i,j) \tag{14}$$

To quantify the separability between clusters, the B_k term (equation 8), can be reformulated with the inter-cluster similarity, measured by the maximum similarity between clusters:

$$B_k^s = \max_{1\leq l\leq K\ ,\ l\neq k} \frac{1}{(n_k)(n_l)} \sum_{i\in C_k}\sum_{j\in C_l} \mathcal{C}(i,j) \tag{15}$$

Notice that, when using sub-sampling for producing clustering ensembles, the co-association matrix, $\mathcal{C}$, represents both pairwise similarity and pairwise stability, due to the perturbation on the data set. In this later situation, the intra-cluster similarity given by equation 14 corresponds to the previously proposed cluster stability validity index, $stab_{C_k}$, as in equation 2.

Both A_k^s and B_k^s vary in the range $[0,1]$ (since $\mathcal{C}(i,j)$ is in the interval $[0,1]$) indicating for the upper limit of the interval, high intra-cluster similarity, and high inter-cluster similarity.

To integrate both the intra-cluster and the inter-cluster information, we reformulate the **Cluster Silhouette** (in equation 9) as:

$$S_k^s = \frac{A_k^s - B_k^s}{\max\{A_k^s, B_k^s\}} \tag{16}$$

Notice that in this case we would like to have a intra-cluster value (A_k^s) larger than the inter-cluster similarity (B_k^s), and therefore the numerator is redefined in reverse order. S_k^s takes values the same interval [-1,1], following the indications of the previous defined S_k: values close to 1 indicate that cluster C_k is compact and well separated.

The second proposed index is a reformulation of the **Dunn's cluster index** (equation 13) using similarities:

$$D_k^s = \frac{diam(C_p)}{\max_{1\leq r\leq K, r\neq k} sim(C_k,C_r)} \tag{17}$$

where $sim(C_k,C_r)$ represents the similarity between the k-th and the r-th cluster, and $diam(C_k)$ is the k-th cluster diameter, defined in terms of pairwise similarities by:

$$sim(C_k,C_r) = \max_{i\in C_k, j\in C_r} \mathcal{C}(i,j) \tag{18}$$

$$diam(C_k) = \min_{i,j\in C_k} \mathcal{C}(i,j) \tag{19}$$

Both $diam(C_k)$ and $sim(C_k,C_r)$ vary in the range $[0,1]$ (since $\mathcal{C}(i,j)$ is in the interval $[0,1]$) indicating for the upper limit of the interval, high intra-cluster similarity, and high inter-cluster similarity.

Notice in the definition of D_k^s, the numerator and the denominator are exchanged (compared with equation 13), allowing that when $D_k^s > 1$ is also an indication of compact and well separated cluster.

We propose to evaluate the efficiency of the previous measures for the selection of the clusters in the Multi-EAC Algorithm. We compare the selection of clusters using the intra-cluster and the inter-cluster pairwise similarity in order to determine which cluster should be selected. Clusters with high value of intra-cluster similarity should be considered; clusters with inter-cluster similarity lower than a given threshold th can also be considered well separated. Moreover, we study the integration of both concepts (intra and inter-cluster similarities), using the Cluster Silhouette and the Dunn's cluster index, trying to improve the robustness of the selection.

In the rest of the paper consider the following notation:

intrasum: The original selection index used as baseline quality measure - uses the intra-cluster average similarity (equation 14). Cluster selection criterion: $intrasum > th_{intrasum}$;

intersum: Uses the inter-cluster average similarity (equation 15). Selection criterion: $intersum < th_1 \Leftrightarrow (1 - intersum) > th_{intersum}$;

silh: Global silhouette of a cluster (equation 16). Selection criterion: $silh > th_{silh}$;

intramin: Computes the minimum intra-cluster similarity, related to cluster diameter, according to 19. Selection criterion: $intramin > th_{intramin}$;

intermax: Computes the maximum inter-cluster similarity, according to 18. Selection criterion: $intermax < th \Leftrightarrow (1 - intermax) > th_{intermax}$;

Dunn: Dunn's cluster index, according to 17. Selection criterion: $Dunn > th_{Dunn}$.

The selection of clusters depends on the comparison with pre-determined thresholds. To select compact and well separated clusters, in the case of $th_{intrasum}$, $th_{intersum}$, $th_{intramin}$ and $th_{intermin}$ the threshold should be close to 1; in the case of the th_{Dunn} the threshold should be higher that 1.

The estimation of the most adequate threshold is out of the scope of the present paper. In this paper we fixed $th_{Dunn} = 10$ and the remaining thresholds were set to $th = 0.9$.

6 EXPERIMENTAL ANALYSIS

We base our evaluation of the proposed criteria on several synthetic and real-world benchmark data-sets from the UCI repository (Asuncion and Newman, 2007).

To produce the ensembles we use different algorithms, enabling the selection of clusters based on different clustering criteria: three agglomerative hierarchical clustering methods (Jain and Dubes, 1988)-

Single Link, Complete Link, Average Link (either fixing the number of clusters or using the life-time criteria (Fred and Jain, 2005)); one partitional clustering method - K-means (Jain and Dubes, 1988); and a Spectral Graph Partioning Algorithm - NJW Spectral Clustering (Ng et al., 2002).

In table 1 we present the data-set characteristics and the parameter values used with different clustering algorithms. For the NJW Spectral Clustering, the scaling parameter σ in the Gaussian affinity matrix, representing the similarity matrix, was taken in the intervals $\sigma_1 = [0.08, 0.1 : 0.05 : 0.95, 1 : .5 : 10]$, $\sigma_2 = [0.08, 0.1 : 0.1 : 0.9, 1 : 1 : 10]$ and $\sigma_3 = [1 : 0.5 : 10]$, where $[\sigma_{min} : inc : \sigma_{max}]$, represents the set of all σ values beginning with σ_{min}, terminating with σ_{max}, with increments of *inc*.

Table 1: Benchmark data-sets characteristics and parameter values used with different clustering algorithms.

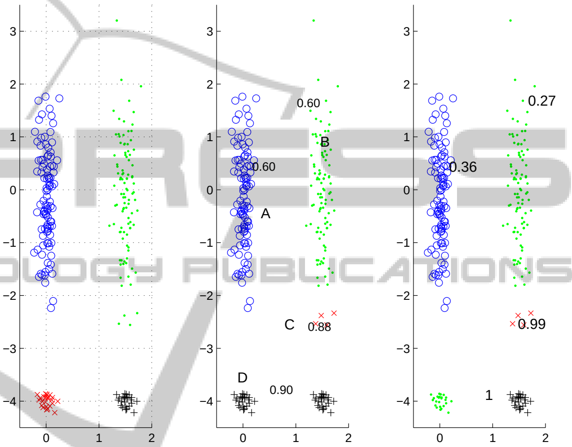
Data-Sets	K	n_s	Ensemble	
			Ks	σ
spiral	2	200	2-8	σ_1
cigar	4	250	2-8	σ_1
rings	3	450	2-6	σ_1
breast-cancer	2	683	2-10,15,20	σ_3
iris	3	150	3-10,15,20	σ_1
image-1-Martin	7	1000	7-15,20,30, 37	σ_2

We applied each *clustering algorithm instantiation*, Alg^l (characterized by a set of parameter values), to N different subsampled versions, X^{sub_i}, of the data-set, fixing N=100. The obtained clustering ensemble CE^l is combined over the co-association matrix $\mathcal{C}^l$. For extracting the combined partitions, we used the Single Link and the Average Link agglomerative hierarchical methods, fixing the number of clusters equal to the real number of clusters, and using the life-time criteria (Fred and Jain, 2005).

We applied each of the stability criteria to the extracted solutions. We selected a cluster when its stability was higher than a given threshold. The applied thresholds were $th = 0.9$ for intrasum, silh, intramin, 1-intramax, 1-intersum. For the Dunn we selected $th = 10$.

The accuracy of the results is evaluated using the Consistency Index - CI (Fred, 2001), obtained by matching the clusters in the obtained partition with the ground truth class labels. If a cluster groups two natural classes, CI matches cluster with the class with higher number of objects. When the clustering partitions have the same number of clusters as the ground truth, CI corresponds to the percentage of correct labeling, that is $(1 - P_e)$, where P_e is the error probability. In this work, when extracting the final partition, we fix the number of clusters equal to the ground truth, so the CI gives the percentage of correct labeling.

To obtain more information about the selection criteria, we analyzed if the selected clusters represent ”good” or ”bad” clusters. This categorization is based on the following: a ”good” cluster represents a grouping of objects belonging to only one natural cluster; on the other hand a ”bad” cluster represents a cluster that joins objects belonging to two (or more) different natural clusters, that is we prefer clusters that cannot include the overlapping between natural clusters.



(a) Ground truth. (b) K-Means ensemble. (c) Spectral clustering ensemble.

Figure 4: Examples of obtained partitions when clustering algorithm instantiation is inadequate.

We illustrate the problem of bad clusters with the synthetic 2D cigar data set presented in figure 4. The data-set is composed of four distinct gaussians (ground truth illustrated in figure 4(a)). Figures 4(b) and 4(c) are two examples of a partitions obtained using SL as extraction method (fixing the number of clusters to $K = 4$) over the ensembles produced for two different clustering algorithm instantiations: the first using K-Means with $K = 4$ and the second using Spectral Clustering with $K = 10$ and $\sigma = 0.1$. The numbers adjacent to each cluster represent the original **intrasum** stability index and the letters are used to identify each of the clusters.

In both examples we have wrong partitionings of the data set (when compared with the ground truth 4(a)). In example 4(b) we see that the black colored clusters (named ’D’ and marked with ’+’) have been wrongly joined; in example 4(c) the green painted clusters (with ’.’ marker) have also been wrongly joined.

In the first case the stability obtained with the intrasum selection index is high (0.9), but on the latter

Table 2: Selection criteria values for example of figure 4.

C	ns_k	Match	intrasum	1-intersum	silh	intramin	1-intermax	Dunn
A	100	100	0.60	0.87	0.78	0.01	0.36	0.02
B	96	96	0.60	0.87	0.78	0.02	0.36	0.04
C	4	4	0.88	0.46	0.40	0.81	0.21	1.02
D	50	0	**0.90**	0.46	0.41	0.77	0.21	0.97

Table 3: Clustering results, in terms of Consistency Index, CI, using the Single Link (SL) and the Average Link hierarchical (AL) methods. The ensembles result from combining the selected clusters from all the clustering algorithm instantiations.

Data-Sets	(SL)						(AL)					
	intrasum	intersum	silh	intramin	intermax	Dunn	intrasum	intersum	silh	intramin	intermax	Dunn
cigar	0.80	0.80	0.80	0.80	1.00	1.00	0.99	0.99	0.99	0.99	1.00	1.00
spiral	0.51	0.51	0.51	0.51	1.00	1.00	0.59	0.61	0.62	0.51	1.00	1.00
iris	0.67	0.67	0.67	0.67	0.34	0.34	0.90	0.90	0.90	0.67	0.34	0.34
rings	0.45	0.56	1.00	0.45	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
breast	0.65	0.65	0.65	0.65	0.65	0.65	0.96	0.96	0.96	0.96	0.65	0.65
image	0.68	0.68	0.68	0.33	1.00	1.00	0.98	0.98	1.00	0.33	1.00	1.00

is lower (0.279), allowing the selection of a wrong cluster in one of the situations. Table 2 presents for this example the values of the other proposed cluster selection criteria. Each row represents each of the obtained clusters (labelled from 'A' to 'D') and columns represent: ns_k -the number of samples of each cluster; Match - the number of samples that matches the objects in the natural clusters of figure 4(a) considering the definition of "good" clusters; **intrasum** to **Dunn** the selection criterion.

Considering the global selection threshold 0.9 for $th_{intrasum}$, $th_{intersum}$, $th_{intramin}$ and $th_{intermin}$, and 10 for th_{Dunn} , the cluster 'D' (last row of the table) would only be selected by the *intrasum* criteria. That is, the other proposed cluster validity indexes are more selective than the previous criterion.

To better understand the differences, consider the co-association matrix of figure 5 which corresponds to the example in figure 4(b). The different stability indexes use differently the co-association values:

- The **intrasum** uses only the intra cluster similarities. The obtained value is an average of the similarities of the objects within the cluster in analysis;
- When using the **intramin** instead of the average between the similarities, the minimum similarity is the value that is extracted (thus being more restrictive);
- When considering to use the inter-cluster information, the **intersum** criteria averages the similarities of the objects of the analysed cluster with the ones on the closer cluster, and the **intermax** uses the maximum value, being more restrictive.

In figure 5 , the color scheme ranges from blue ($\mathcal{C}(i,j) = 0$) to red ($\mathcal{C}(i,j) = 1$), corresponding to the magnitude of similarity, and the axis represent the samples of the data-set organized such that samples belonging to the same cluster are displayed contiguously. The two lower block diagonal matrices represent the wrongly joined clusters (named 'D'). Their intracluster similarity is very high, but when analysing the minimum value of intracluster similarity it is possible to see that this cluster is not so stable as it first looks (0.77 similarity - compared with an average similarity of 0.90). If we consider the intercluster similarity, the corresponding stability value will significantly decrease. It can be noticed that there are vertical and horizontal lines on the matrix representing higher similarity values. The inclusion of the other sources of information avoids the inclusion of this bad cluster (merging two "natural" clusters) in the final combination process.

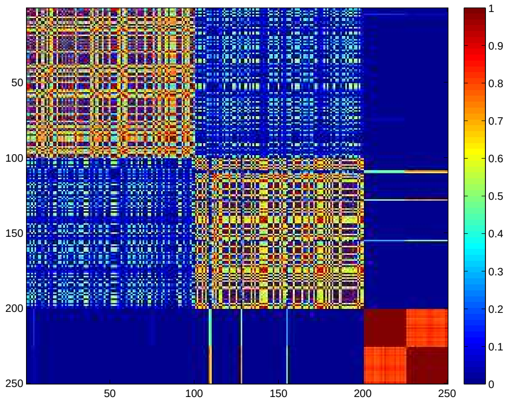


Figure 5: K-Means ensemble ($K = 4$) - Example of associated co-association matrix.

The new proposed criteria seem more restrictive than the **intrasum** (used before), selecting fewer clusters from the co-association matrixes obtained from each algorithm instantiation. Nevertheless the integration of all the algorithm instantiations, for almost all of the benchmark data-sets, resulted in the inclusion of clusters that covered all the data-set.

To summarize the results for the benchmark datasets, we present in table 3 the CI index using the SL and the AL hierarchical methods as extraction methods, marking for each data-set the maximal CI value.

The right half of the table presents more marked cells, showing that the AL extraction method conducted to better results. Comparison of the results shows that the Silhouete (*silh*), the *intermax*, and the *Dunn* index systematically leads to better results than the original **intrasum** selection criterium. The remaining do not show an evident superiority, and **intramin** is the index presenting the worst performance.

The *intermax* and **Dunn** criteria were the best in almost every data-set. The cases in which they did not obtain the best results correspond to situations where only a subset of samples was selected, since these criteria selected only a small part of the evaluated clusters. This fact caused that some objects were not part of any of the selected clusters, penalizing the overall result, since some natural clusters didn't have any match.

The criterion Silhouete gave also results comparable with Dunn and intermax criteria in almost all data-sets, allowing the coverage of all objects in every data-set.

7 CONCLUSIONS

Adopting the Multiple-Criteria Evidence Accumulation Clustering method (Multi-EAC) as baseline cluster combination method, we addressed the issue of selection of meaningful clusters from the multiple data partitions. In previous work, the authors proposed a cluster validity criterion based on cluster stability, assessed from intermediate co-association matrices, obtained from clustering ensembles produced by a single clustering algorithm by perturbing the data set using sub-sampling. In this paper we proposed new cluster validity criteria for the selection of clusters from the same intermediate co-associations matrices but using it on a different perspective. Instead of considering only the intra-cluster similarity, we propose indexes based on inter-cluster similarity and combination of intra-cluster and inter-cluster similarities. Comparison of the several criteria was based on the performance of the combined data partitions, obtained by accounting only on clusters that are selected according to the corresponding criteria.

Experimental results have shown that four out of the the five proposed criteria lead in general to better combination results than by using the cluster stability criterion. In particular, the criterion Silhouete and Dunn focusing both the intra and the inter-cluster separability, and the *intermax* focusing on intra-cluster separability, gave the overall best results.

Furthermore, the new methods can also be applied to clustering ensembles that do not make use of data sub-sampling, being of more general applicability. Additional experiments on larger data sets and on more real data sets are underway.

REFERENCES

Asuncion, A. and Newman, D. (2007). UCI ML repository.

Ayad, H. G. and Kamel, M. S. (2008). Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):160–173.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. In *Pacific Symposium on Biocomputing*.

Bezdek, J. C. and Pal, N. R. (1995). Cluster validation with generalized dunn's indices. In *ANNES '95: Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems*, page 190, Washington, DC, USA. IEEE Computer Society.

Bolshakova, N. and Azuaje, F. (2003). Cluster validation techniques for genome expression data. *Signal Process.*, 83(4):825–833.

Dubes, R. and Jain, A. (1979). Validity studies in clustering methodologies. *Pattern Recognition*, 11:235–254.

Dunn, J. C. (1974). A fuzzy relative of the isodata process and its use in detecting compact, well separated clusters. *Cybernetics and Systems*, 3(3):32–57.

Fern, X. Z. and Brodley, C. E. (2004). Solving cluster ensemble problems by bipartite graph partitioning. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 36, New York, NY, USA. ACM.

Fred, A. (2001). Finding consistent clusters in data partitions. In Kittler, J. and Roli, F., editors, *Multiple Classifier Systems*, volume 2096, pages 309–318. Springer.

Fred, A. and Jain, A. (2005). Combining multiple clustering using evidence accumulation. *IEEE Trans Pattern Analysis and Machine Intelligence*, 27(6):835–850.

Fred, A. and Jain, A. (2006). Learning pairwise similarity for data clustering. In *Proc. of the 18th Int'l Conference on Pattern Recognition (ICPR)*, volume 1, pages 925–928, Hong Kong.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Intelligent Information Systems Journal*, 17(2-3):107–145.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002a). Cluster validity methods: Part i. *SIGMOD Record*.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2002b). Cluster validity methods: Part ii. *SIGMOD Record*.

Jain, A. and Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall.

Jain, A. K. and Moreau, J. V. (1987). Bootstrap technique in cluster analysis. *Pattern Recognition*, 20:547 – 568.

Lange, T., Braun, M., Roth, V., and Buhmann, J. (2002). Stability-based model selection. In *NIPS*, pages 617–624.

Levine, E. and Domany, E. (2000). Resampling method for unsupervised estimation of cluster validity. *Aaa*.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. B. and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA.

Roth, V., Lange, T., Braun, M., and Buhmann, J. (2002). A resampling approach to cluster validation. In *Computational Statistics-COMPSTAT*.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sergios Theodoridis, K. K. (1999). *Pattern Recogniton*. Academic Press.

Strehl, A. and Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *J. of Machine Learning Research 3*.

Topchy, A., Jain, A. K., and Punch, W. (2005). Clustering ensembles: Models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1866–1881.

Verma, D. and Meila, M. (2003). A comparision of spectral clustering algorithms. Technical report, UW CSE Technical report.