Robust Data Clustering

Ana L.N. Fred
Institute of Telecommunications
Instituto Superior Técnico
afred@lx.it.pt

Anil K. Jain
Dept of Computer Science and Engineering
Michigan State University
jain@cse.msu.edu

Abstract

We address the problem of robust clustering by combining data partitions (forming a clustering ensemble) produced by multiple clusterings. We formulate robust clustering under an information-theoretical framework; mutual information is the underlying concept used in the definition of quantitative measures of agreement or consistency between data partitions. Robustness is assessed by variance of the cluster membership, based on bootstrapping. We propose and analyze a voting mechanism on pairwise associations of patterns for combining data partitions. We show that the proposed technique attempts to optimize the mutual information based criteria, although the optimality is not ensured in all situations. This evidence accumulation method is demonstrated by combining the well-known K-means algorithm to produce clustering ensembles. Experimental results show the ability of the technique to identify clusters with arbitrary shapes and sizes.

1. Introduction

Let $X = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}\}$ be a set of n objects or patterns. The problem of clustering consists of producing a partition of X into k "natural" groups or clusters, $P = \{C_1, C_2, \ldots, C_k\}$, k being in general unknown. Hundreds of clustering algorithms exist [16, 6, 19, 9, 11, 14, 3, 18, 1, 12], yet it is difficult to find a single clustering algorithm that can handle all types of cluster shapes and sizes. Instead of choosing a particular clustering algorithm for a given data set, the idea of combining the results of multiple clusterings in order to obtain robust data partitions has recently been proposed [13, 21]. Given N different partitions of the data X, which we define as a *clustering ensemble* $\mathbb{P} = \{P^1, P^2, \ldots, P^N\}$, where $P^i = \{C_1^i, C_2^i, \ldots, C_{k_i}^i\}$ has k_i clusters, the problem consists of producing a partition P^*, which is the result of a combination of the N partitions in $\mathbb{P}$. Ideally, P^* should satisfy the following properties:

(a) *Consistency with the clustering ensemble* $\mathbb{P}$. This means that the combined data partition P^* should somehow agree with the individual partitions, $P^i, i = 1, \ldots, N$.

(b) *Robustness to small variations in* $\mathbb{P}$. The number of clusters and the cluster membership in P^*, should not change significantly with small perturbation of the partitions in $\mathbb{P}$.

(c) *Goodness of fit with the ground truth information,* if available. P^* should be consistent with external cluster labels, or with perceptual evaluation of the data.

Fred and Jain [13] introduce the concept of evidence accumulation clustering, that maps the individual data partitions in a clustering ensemble into a new similarity measure between patterns, summarizing inter-pattern structure perceived from these clusterings; a final data partition is obtained by applying the single-link method to the new similarity matrix. Strehl and Ghosh [21] explore the concept of consensus between data partitions, using graph-theoretical approaches for consensus decisions, based on a cluster matching paradigm.

In this paper we propose an information-theoretic approach, based on the concept of mutual information and on variance analysis using bootstrapping to (i) measure the consistency between data partitions; (ii) define objective functions for criteria mentioned in (a) and (b) above; and (iii) define figures of merit concerning the agreement with ground truth information, as stated in (c). Optimality of the evidence accumulation strategy is analyzed in light of these objective functions. Experimental results are based on applying a combination of K-means clusterings to analyze both synthetic data and real data sets from the UCI repository.

2. Consistency of Data Partitions Using Mutual Information

A partition P^a describes a labelling of the n patterns in the data set X, into k_a clusters. Taking frequency counts as approximations for probabilities, the entropy

[4] of the data partition P^a is expressed by $H(P^a) = -\sum_{i=1}^{k_a} \frac{n_i^a}{n} \log\left(\frac{n_i^a}{n}\right)$, where n_i^a represents the number of patterns in cluster $C_i^a \in P^a$. The agreement between two partitions P^a and P^b is measured by the mutual information $I(P^a, P^b)$, as proposed by Strehl and Ghosh [21] $I(P^a, P^b) = \sum_{i=1}^{k_a}\sum_{j=1}^{k_b} \frac{n_{ij}^{ab}}{n} \log\left(\frac{\frac{n_{ij}^{ab}}{n}}{\frac{n_i^a}{n}\cdot\frac{n_j^b}{n}}\right)$, with n_{ij}^{ab} denoting the number of shared patterns between clusters C_i^a and C_j^b, $C_i^a \in P^a$ and $C_j^b \in P^b$. From the definition of mutual information [4], it is easy to demonstrate that $I(P^a, P^b) \leq \left(H(P^a) + H(P^b)\right)/2$. We define *normalized mutual information* (NMI) between two partitions P^a and P^b as $NMI(P^a, P^b) = \frac{2 \cdot I(P^a, P^b)}{H(P^a)+H(P^b)}$, which, after simplification, leads to the equation

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

Note that $0 \leq NMI(.,.) \leq 1$.

The agreement between a given partition, P, and the clustering ensemble, $\mathbb{P}$, designated by the *average normalized mutual information* [21], is defined by

$$NMI(P, \mathbb{P}) = \frac{1}{N}\sum_{i=1}^{N} NMI(P, P^i). \quad (2)$$

We further define the *average agreement between partitions* in a clustering ensemble $\mathbb{P}$ by $NMI(\mathbb{P}, \mathbb{P}) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} NMI(P^i, P^j)/\binom{N}{2}$.

3. Objective Functions and Optimality Criteria

Let $\check{\mathbb{P}}^k = \check{P}^{1^k}, \ldots, \check{P}^{m^k}$, $m = \frac{1}{k!}\sum_{l=1}^{k}\binom{k}{l}(-1)^{k-l} l^n$, represent the set of all possible partitions of the n patterns in X into k clusters. We define *k-cluster consensus partition*, P^{*^k}, as the k-cluster partition that best fits the clustering ensemble $\mathbb{P}$, maximizing the objective function $NMI(\check{P}^k, \mathbb{P})$, that is, satisfying the optimality criterion

$$P^{*^k} = \arg\max_i \left\{ NMI(\check{P}^{i^k}, \mathbb{P}) \right\}. \quad (3)$$

For each value of k, the criterion in equation (3) ensures the satisfaction of the property (a) in section 1.

In order to address the robustness property (b) in section 1, we perturb the clustering ensemble $\mathbb{P}$, using a bootstrap technique, and compute the variance of the resulting NMI values. Let $\mathbb{P}^B = \left\{\mathbb{P}^{b_1}, \ldots, \mathbb{P}^{b_B}\right\}$ denote the B bootstrap clustering ensembles produced by sampling with replacement from $\mathbb{P}$, and let $\mathbb{P}^{*B} = \left\{P^{*b_1}, \ldots, P^{*b_B}\right\}$ be the corresponding set of combined data partitions. The mean value of the average normalized mutual information between k-cluster combined partitions and the bootstrap clustering ensembles is given by

$$\overline{NMI(P^{*^k_b}, \mathbb{P}^b)} = \frac{1}{B}\sum_{i=1}^{B} NMI(P^{*^k_{b_i}}, \mathbb{P}^{b_i}), \quad (4)$$

and the corresponding variance is defined as follows

$$var\{NMI(P^{*^k_b}, \mathbb{P}^b)\} = \frac{1}{B-1}\sum_{i=1}^{B}\left(NMI(P^{*^k_{b_i}}, \mathbb{P}^{b_i}) - \overline{NMI(P^{*^k_b}, \mathbb{P}^b)}\right)^2. \quad (5)$$

It is expected that a robust data partition combination technique will be stable with respect to minor clustering ensemble variations; we model this robustness property through the minimum variance criterion

$$P^* : \min_k \left\{ var\{NMI(P^{*^k_b}, \mathbb{P}^b)\right\} \text{ is achieved.} \quad (6)$$

Let us define the *variance of NMI between bootstrap clustering ensembles* as

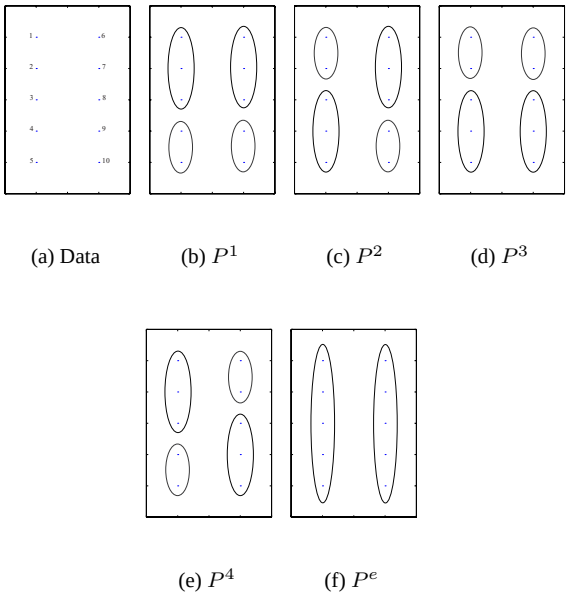
$$var\{NMI(\mathbb{P}^b, \mathbb{P}^b)\} = \frac{1}{B-1}\sum_{i=1}^{B}\left(NMI(\mathbb{P}^{b_i}, \mathbb{P}^{b_i}) - \overline{NMI(\mathbb{P}^b, \mathbb{P}^b)}\right)^2, \quad (7)$$

with $\overline{NMI(\mathbb{P}^b, \mathbb{P}^b)} = \frac{1}{B}\sum_{i=1}^{B} NMI(\mathbb{P}^{b_i}, \mathbb{P}^{b_i})$. Minimization of the variance criterion in equation (6) implies the following inequality:

$$var\{NMI(P^{*^k_b}, \mathbb{P}^b)\} \leq var\{NMI(\mathbb{P}^b, \mathbb{P}^b)\}. \quad (8)$$

The variability of the partition configurations is measured by $var\{NMI(P^{*_b}, P^{*_b})\}$; stable solutions have smaller variance, ideally equal to 0. In the following, standard deviation (std) will be used instead of variance.

The objective function in equation (3) is essential to guarantee that a partition combination technique provides the k-cluster partition that is consistent with the underlying clustering ensemble. It does not, however, serve as a criterion for deciding the correct number of clusters, k, in the final partition. The minimum variance criterion in equation (6), on the other hand, is able to decide the "optimal" number of clusters among various combination strategies. This is illustrated through a simple example in figure 1, consisting of 10 2D-patterns distributed along 2 straight lines (fig. 1(a)); figures 1(b) to 1(e) present 4 different partitions, $P^1, \ldots P^4$, of this data set into 4 clusters, forming the clustering ensemble $\mathbb{P}$. It is easy to see that any of these 4 partitions can be chosen as a 4-cluster consensus partition. In fact, $NMI(P^i, \mathbb{P}) = 0.8602$, $i = 1, \ldots, 4$, and



(a) Data (b) P^1 (c) P^2 (d) P^3

(e) P^4 (f) P^e

Figure 1. Illustration of clustering ensemble and combined data partition.

any other partition has a lower NMI value. The 2-cluster consensus partition, represented in figure 1(f) as P^e, however, is unique, with $NMI(P^e, \mathbb{P}) = 0.6732$. Although this has a lower NMI value than the 4-cluster solutions, it represents better the true structure of the patterns. Analyzing the partitions in $\mathbb{P}$, it is obvious that any partition that mixes patterns from the two sets containing patterns (1 to 5) and (6 to 10) is unacceptable. On the other hand, the evidence on pattern associations accumulated over the clusters in $\mathbb{P}$ reveals that the pattern pairings (1, 2), (4, 5), (6, 7), (9, 10) should always be maintained (they correspond to unanimous pattern associations), and that associations (2, 3), (3, 4), (7, 8) and (8,9) are present 50% of the time; therefore, either these associations are broken, leading to a 6-cluster partition, or are not broken, resulting in the 2-cluster partition P^e. Variance analysis corroborates the later decision: by bootstrapping on the clustering ensemble $\mathbb{P}$ ($B = 100$), different 4-cluster consensus partitions are obtained, with $std\{NMI(P^{*^4_b}, \mathbb{P}^b)\} = 0.04$, and $std\{NMI(P^{*^4_b}, P^{*^4_b}\} = 0.07$ (higher than $std\{NMI(\mathbb{P}^b, \mathbb{P}^b)\} = 0.05$); partition P^e shown in fig 1(f), however, continues to be the only stable 2-cluster consensus partition, with $std\{NMI(P^{*^2_b}, \mathbb{P}^b)\} = 0.00$.

4. Combining Data Partitions

4.1. Evidence Accumulation using a Voting Mechanism

The idea of evidence accumulation clustering is to combine the results of multiple clusterings into a single data partition, by viewing each clustering result as an independent evidence of data organization. A clustering algorithm, l, by organizing the n patterns into clusters according to the partition P^l, expresses relationships between objects in the same cluster; these are mapped into a binary $n \times n$ co-association matrix, $\mathcal{C}^l(i,j)$, where non-null pairwise relations, $\mathcal{C}^l(i,j) = 1$, express co-existence of patterns i and j in the same cluster of P^l. Assuming that patterns belonging to a "natural"cluster are very likely to be co-located in the same cluster in different clusterings, we take the co-occurrences of pairs of patterns in the same cluster as votes for their association; the clustering ensemble $\mathbb{P}$ is mapped into a $n \times n$ co-association matrix, as follows:

$$\mathcal{C}(i,j) = \frac{n_{ij}}{N} = \frac{\sum_{l=1}^{N} \mathcal{C}^l(i,j)}{N}, \qquad (9)$$

where n_{ij} is the number of times the pattern pair (i,j) is assigned to the same cluster among the N clusterings. Evidence accumulated over the N clusterings, according to equation (9), induces a new similarity measure between patterns, which is then used to recluster the patterns, yielding the combined clustering P^*. We use the single-link (SL) method to extract the final partition from the co-association matrix $\mathcal{C}$. We define the *lifetime of a k-cluster partition* as the absolute difference between its birth and merge thresholds in the dendrogram produced by the SL method; the final data partition is chosen as the one with the highest lifetime.

Figure 2 gives a schematic description of the proposed method. In order to reduce the computational complexity, the algorithm focuses on computing the associations between neighboring patterns. This results in a $n \times p$ co-association matrix, $\mathcal{C}$; $\mathcal{C}(i,j)$ represents the percentage of times pattern i and its jth nearest neighbor are assigned to the same cluster, among the N clusterings, $j = 1, \ldots, p$. This requires the pre-computation of a $n \times p$ matrix, which stores the indices of the p nearest neighbors for each of the n patterns [17]. The SL algorithm is applied to the corresponding $n \times p$ similarity matrix [7].

4.2. On the Optimality of the Proposed Technique

According to the information theoretical objective function in equation (3), the mutual information between partitions, as given by (1) and (2), is maximized based on the number of patterns shared between clusters in these partitions. The proposed voting mechanism maps the set of individual partitions into a new similarity measure, where the strength of the links between patterns is proportional to the percentage of times these patterns are shared by clusters in these partitions. By cutting weak links in the associated minimum spanning tree (MST), which is formally equivalent to cutting the dendrogram produced by the SL method

Input: n - number of patterns;
$n \times p$ nearest neighbor matrix
N - number of clusterings.
$\mathbb{P} = \{P^1, \ldots P^N\}$ - clustering ensemble

Output: P^* - Combined data partition.

Initialization: Set the $n \times p$ co-association matrix, $\mathcal{C}(.,.)$, to a null matrix.

1. For each partition $P^l \in \mathbb{P}$ do:

1.1. Update the co-association matrix: for each pattern pair (i, j) in the pth neighbor list, that belongs to the same cluster in P^l, set $\mathcal{C}(i, j) = \mathcal{C}(i, j) + \frac{1}{N}$.

2. Detect consistent clusters in the co-association matrix using the SL technique: compute the SL dendrogram; the final partition, P^*, is chosen as the one with the highest lifetime.

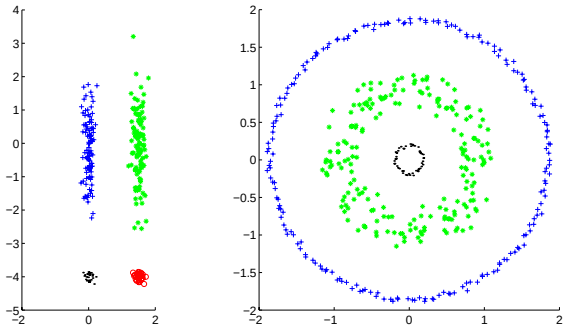
Figure 2. Data clustering using Evidence Accumulation.

[15], we are trying to maximize the number of shared patterns, based on a chain of high frequency pairwise associations, and therefore to maximize (3). The global optimum is, however, not ensured in all situations.

Having satisfied the consistency property with the clustering ensemble, we now address the robustness issue. By bootstrapping on the clustering ensemble $\mathbb{P}$, the corresponding dendrograms, produced by the SL method over the co-association matrix $\mathcal{C}$, will change. When cutting these dendrograms at the highest lifetime partition level, we are minimizing the effect of these changes on the final data partition, and therefore we are minimizing the variance of the average normalized mutual information, as given by equation (5); while the optimal global solution according to criterion (6) is not ensured, the companion necessary condition in (8) should be satisfied.

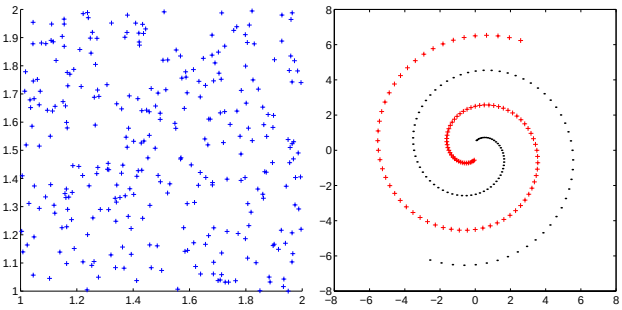
5. Experimental Results

We have tested the evidence accumulation combination method described above by combining K-means clusterings. The algorithm follows a split and merge technique: first the data is decomposed into a large number of small spherical clusters using the K-means algorithm; using N random initializations of the K-means, a clustering ensemble with N partitions is obtained; initial clusters are merged through the partition combination technique described earlier, leading to the combined data partition P^*. The value of k, in the K-means algorithm, can be either fixed to a constant value, or randomly selected in the range $[k_{min}, k_{max}]$.



(a) "Cigar" data set.

(b) 3-rings data set.



(c) Projection of 5-D random data set into a 2-D space.

(d) Spiral data set.

Figure 3. Robust Clustering: results on four artificial data sets.

Figures 3(a)-3(d) show the results of the evidence accumulation algorithm (each cluster has a distinct color), with $N = 50$, on four synthetic data sets: (a) 4-cluster data set (referred to as "cigar" data), $k = 15$; (b) rings data set, $k = 50$; (c) random data set (300 patterns uniformly distributed in a 5-D hypercube), $k = 15$; and (d) spiral data set, $k = 30$. Similar results are obtained when the value of k is randomly selected in the interval $[k_{\min}, k_{\max}]$. For all the four data sets, the evidence accumulation approach identifies the true clustering structure. Note that for the random data set of figure 3(c), our algorithm identifies a single cluster. The typical evolution of $\overline{NMI(P^{*k}_b, \mathbb{P}^b)}$ and of $std\{NMI(P^{*k}_b, \mathbb{P}^b)\}$ is illustrated in figure 4 (curve and error bars in black - thin line, referred to as $NMI(P^*, P)$) for the cigar data set; statistics were computed over $B = 100$ bootstrap experiments, and P^{*k}_b partitions were obtained by forcing $k-$cluster solutions using the SL method on the co-association matrices. While

the average normalized mutual information grows with increasing k (with a maximum at the number of clusters in the clustering ensemble, $k = 15$), the variance is a good indicator of the "natural" number of clusters, having a minimum value at $k = 4$; the partition lifetime criterion for extracting the combined partition from the dendrogram produced by the SL method, leads precisely to this number of clusters, as shown in figure 3(a). This also corresponds to the perceptual organization of the data, which we represent as P^o. The thick curve and corresponding error bars represent $\overline{NMI(P^{*^k_b}, P^o)}$ and $std\{NMI(P^{*^k_b}, P^o)\}$, respectively. Now, the zero variance is achieved for the 2-cluster and the 4-cluster solutions, meaning that a unique partition is produced as the corresponding $k-$cluster consensus partition; the maximum agreement with perceptual evaluation of the data is obtained for $k = 4$, which coincides with the minimum variance of $NMI(P^{*^k_b}, \mathbb{P}^b)$. Figure 5 shows plots of $std\{NMI(P^{*^k_b}, \mathbb{P}^b)\}$ (solid line curves) and of $std\{NMI(\mathbb{P}^b, \mathbb{P}^b)\}$ (dashed lines) for several data sets. It is interesting to note that, in the absence of a clustering structure, the $std\{NMI(P^{*^k_b}, \mathbb{P}^b)\}$ curve for the random data set (upper curve) has high values, for $k \geq 2$, compared to $std\{NMI(P^{*^k_b}, \mathbb{P}^b)\}$, and does not obey the inequality in equation (8); the evidence accumulation algorithm identifies a single cluster in this situation (figure 3(c)). With the remaining data sets, the evidence accumulation clustering decision corresponds to the minimum of $std\{NMI(P^{*^k_b}, \mathbb{P}^b)\}$, which falls below $std\{NMI(P^{*^k_b}, \mathbb{P}^b)\}$, thus obeying the inequality (8).

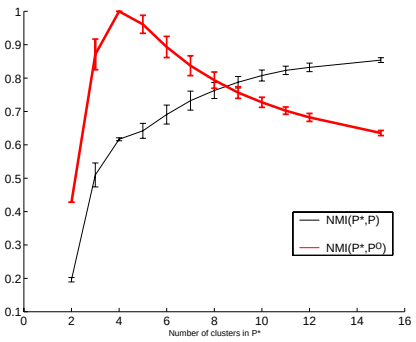


Figure 4. Variance analysis on the "cigar" data set.

The evidence accumulation clustering technique was applied to the Iris data set, with class labels (Setosa, Versicolor, and Virginica) being removed from the data. With $k = 15$ and $N = 50$, two clusters were identified (see fig. 5), corresponding to a merging of the types Virginica and Versicolor into a single cluster. These results are comparable with other techniques, such as the single link method, or the results in [20]. The difficulty in separating the Virginica

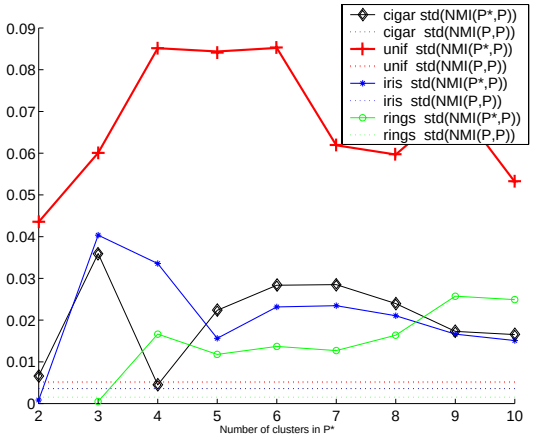


Figure 5. Standard deviations of NMI.

and Versicolor classes using clustering techniques is justified by the fact that these are touching clusters. Interpreting touching clusters as a particular case of noisy patterns, we removed the low density patterns, estimated using the shared nearest neighbor method in [8]. Eliminating about 20% of these atypical patterns, and applying the evidence accumulation technique, with fixed k or variable k (for instance $k \in [2, 20]$), a 3-cluster partition was obtained, with a classification error rate of 10.67% by comparing the clustering labels with the true class labels.

The original Wisconsin Breast Cancer data set (683 patterns represented by 9 integer-valued attributes, with class labels - benign and malignant - removed), available at the UCI Machine Learning Repository, is another example of touching clusters. In this case, we started by representing the original data by 250 centroids obtained by K-means clustering; applying the noise removal technique mentioned above to these centroids, 199 prototypes remained. The evidence accumulation clustering technique was applied to these prototypes, with $N = 100$, and $k \in [2, 10]$, leading to a 2-cluster partition with a correct classification rate of 96.93%. This result compares favorably to the cluster center based methods reported in [20] and [2], with accuracies of 94.28% and 95.5%, respectively. The evidence accumulation technique was also applied to 285 prototypes extracted from the new diagnostic Breast Cancer database (569 patterns, 30 numeric attributes), by using the same noise removal technique, $N = 100$, and $k \in [2, 10]$; the combined data partition contains 2 clusters, with a correct classification rate of 88.93%. The spectral kernel method described in [5] achieves (it is not clear for which of the two data sets) a 79.65% recognition rate when using a Gaussian kernel, and 97.29% recognition with a linear kernel.

The evidence accumulation clustering technique was also applied to the texture data set, that consists of 4000 patterns in a 19-dimensional feature space, representing an

image with 4 distinct textures [10]. This is a difficult data set due to the overlap between clusters. Using the combined prototype/sampling technique (600 prototypes) with random selection of k ($k \in [2, 20]$, $N = 200$, a 2-cluster partition was obtained, corresponding to the merging of natural clusters (defined based on a priori knowledge of the classification of the data into 4 texture classes) in groups of two; matching the 2-cluster partition with the corresponding merged classes gives an overall recognition rate of 95.5%. When trying to identify the most stable 4-cluster partition, we applied the K-means based evidence accumulation clustering algorithm with fixed $k = 4$ on the same prototypes. The most stable solution consisted of 3 clusters, corresponding to an overall recognition rate of 72.45% (two of the classes were still merged; matching the partition with the ideal classes, with these two merged, gives a 96.9% recognition rate); the next most stable solution corresponds to a 4-cluster partition, with a 91.95% recognition rate.

6. Conclusions

This paper has addressed the problem of robust clustering based on the combination of data partitions. Adopting an information theoretic-based approach, and with the goal of obtaining consistent and robust combination techniques, we defined objective functions and optimality criteria, based on the concept of mutual information, and on variance analysis using bootstrapping. The evidence accumulation technique was described, leading to a mapping of the clustering ensemble into a new similarity measure between patterns, by a voting mechanism on pairwise pattern associations. Optimality of this technique was discussed in light of the proposed criteria.

The proposed approach was tested on the combination of K-means clusterings; results obtained on both synthetic and real data sets illustrate the ability of the evidence accumulation technique to identify clusters with arbitrary shapes and arbitrary sizes, without using *a priori* information about the number of clusters, or ad-hoc specification of parameters. Results produced by our technique, by a simple combination of K-means clusterings, and without the need of parameter tuning, outperformed some of the results reported in the literature with more sophisticated unsupervised techniques. It is expected that the application of the evidence accumulation technique using more powerful clustering methods, than the K-means, can lead to even better clustering results.

Acknowledgments

This work was partially supported by the Portuguese Foundation for Science and Technology (FCT), Portuguese Ministry of Science and Technology, and FEDER, under grant POSI/33143/SRI/2000, and ONR grant no. N00014-01-1-0266.

References

[1] P. Bajcsy and N. Ahuja. Location- and density-based hierarchical clustering using similarity analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(9):1011–1015, 1998.

[2] S. V. Chakravarthy and J. Ghosh. Scale-based clustering using the radial basis function network. *IEEE Trans. Neural Networks*, 7:1250–1261, 1996.

[3] D. Comaniciu and P. Meer. Distribution free decomposition of multivariate data. *Pattern Analysis and Applications*, 2:22–30, 1999.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[5] N. Cristianini, J. Shawe-Taylor, and J. Kandola. Spectral kernel methods for clustering. In S. Becker T. G. Dietterich and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. MIT Press, Cambridge, MA, 2002.

[6] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, second edition, 2001.

[7] Y. El-Sonbaty and M. A. Ismail. On-line hierarchical clustering. *Pattern Recognition Letters*, pages 1285–1291, 1998.

[8] L. Ertoz, M. Steinbach, and V. Kumar. A new shared nearest neighbor clustering algorithm and its applications. In *Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining, http://www-users.cs.umn.edu/ kumar/papers/papers.html*, 2002.

[9] B. Everitt. *Cluster Analysis*. John Wiley and Sons, 1993.

[10] M. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(3):381–396, 2002.

[11] B. Fischer, T. Zoller, and J. Buhmann. Path based pairwise data clustering with application to texture segmentation. In M. Figueiredo, J. Zerubia, and A. K. Jain, editors, *Energy Minimization Methods in Computer Vision and Pattern Recognition*, volume 2134 of *LNCS*, pages 235–266. Springer Verlag, 2001.

[12] Chris Fraley and Adrian E. Raftery. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[13] A. Fred and A. K. Jain. Data clustering using evidence accumulation. In *Proc. of the 16th Int'l Conference on Pattern Recognition*, pages 276–280, 2002.

[14] E. Gokcay and J. C. Principe. Information theoretic clustering. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(2):158–171, 2002.

[15] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

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

[17] B. Kamgar-Parsi and L. N. Kanal. An improved branch and bound algorithm for computing k-nearest neighbors. *Pattern Recognition Letters*, I:195–205, 1985.

[18] G. Karypis, E-H Han, and V. Kumar. CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, 32(8):68–75, 1999.

[19] L. Kaufman and P. J. Rosseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, Inc., 1990.

[20] R. Kothari and D. Pitts. On finding the number of clusters. *Pattern Recognition Letters*, 20:405–416, 1999.

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