A New Cluster Isolation Criterion Based on Dissimilarity Increments

Ana L.N. Fred, *Member*, *IEEE*, and José M.N. Leitão, *Member*, *IEEE*

Abstract—This paper addresses the problem of cluster defining criteria by proposing a model-based characterization of interpattern relationships. Taking a dissimilarity matrix between patterns as the basic measure for extracting group structure, dissimilarity increments between neighboring patterns within a cluster are analyzed. Empirical evidence suggests modeling the statistical distribution of these increments by an exponential density; we propose to use this statistical model, which characterizes context, to derive a new cluster isolation criterion. The integration of this criterion in a hierarchical agglomerative clustering framework produces a partitioning of the data, while exhibiting data interrelationships in terms of a dendrogram-type graph. The analysis of the criterion is undertaken through a set of examples, showing the versatility of the method in identifying clusters with arbitrary shape and size; the number of clusters is intrinsically found without requiring ad hoc specification of design parameters nor engaging in a computationally demanding optimization procedure.

Index Terms—Clustering, hierarchical methods, context-based clustering, cluster isolation criteria, dissimilarity increments, model-based clustering.

1 Introduction

In this section, we review existing clustering methodologies and algorithms, and outline the goals and the main ideas proposed in this paper.

1.1 Review of Clustering Approaches

Clustering has been applied in a variety of domains, whose main goals are exploratory pattern analysis and data mining, decision-making, and machine learning. Most of the existing work in clustering deals with developing new clustering algorithms. Two main strategies have been adopted: hierarchical methods and partitional methods [1], [2].

Partitional methods organize patterns into a small number of clusters. Model-based techniques assume that patterns belonging to a cluster can be given a simple and compact description in terms of a parametrical distribution (such as a Gaussian), a representative element (the centroid or the median, for instance), or some geometrical primitive (lines, planes, circles, ellipses, curves, surfaces, etc.). Such approaches assume particular cluster shapes, partitions being obtained, in general, as a result of an optimization process using a global criterion. Parametric density approaches, such as mixture decomposition techniques [3], [4], [5], [6], and prototype-based methods, such as central clustering [7], square-error clustering [8], K-means [2], [1], or K-medoids clustering [9], emphasize compactness, imposing hyperspherical clusters in the data. Model order selection is sometimes left as a design parameter or it is incorporated in the clustering procedure [10], [11], [5]. The K-means is probably the best known and most widely used algorithm in this category. Assuming a priori knowledge about the number of classes, and based on the square-error criterion, it is a computationally efficient clustering technique that identifies hyperspherical clusters. Extensions of the basic method include: use of Mahalanobis distance to deal with hyperellipsoidal clusters [2]; fuzzy algorithms [12]; adaptations to straight line fitting [13]. Optimization-based clustering algorithms adopting shape fitting approaches include [14], [15], [16]. Cost-functional clustering methods based on a minimum variance criterion favor spherical clusters. Other optimization-based clustering algorithms do not assume particular cluster shapes, such as the work in [17], proposing a pairwise clustering cost function emphasizing cluster connectedness. Nonparametric density-based clustering methods attempt to identify high-density clusters separated by low-density regions by either exploiting regions of high sample density [18] or regions with less data, such as in valley seeking clustering algorithms [19], [20].

Hierarchical methods, mostly inspired by graph theory [21], consist of a sequence of nested data partitions in a hierarchical structure, that can be represented graphically as a dendrogram [2]. Both agglomerative [2], [22] and divisive approaches [23] (such as those based on the minimum spanning tree—MST [2]) have been attempted. Variations of the algorithms are obtained depending on the definition of similarity measures between patterns and between clusters [24], the later ultimately determining the structure of the clusters identified. The single-link (SL) and the complete-link (CL) methods [2] are the best known techniques in this class, emphasizing, respectively, connectedness and compactness. Prototype-based hierarchical methods define similarity between clusters based on cluster representatives, such as the centroid or the median; like the prototype-based partitional algorithms, these techniques fail to identify clusters of arbitrary shapes and sizes, imposing spherical structure in the data. Variations of the prototype-based hierarchical clustering include the use of multiple prototypes per cluster, as in the CURE algorithm [25]. Other algorithms compute similarity between clusters by the aggregate of the similarities

• *The authors are with the Instituto de Telecomunicações, Instituto Superior Técnico, Av. Rovisco Pais 1049-001, Lisbon, Portugal.*
E-mail: {afred, jleitao}@lx.it.pt.

Manuscript received 14 Feb. 2001; revised 26 Mar. 2002; accepted 27 Dec. 2002. Recommended for acceptance by R. Kumar.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 113621.

(emphasizing interconnectivity, such as the group-average method [2]) among pairs of patterns belonging to distinct clusters, or selecting a particular pair. Other hierarchical agglomerative clustering algorithms follow a split and merge technique; the data being initially split into a high number of small clusters and merging being based on intercluster similarity. A final partition is selected among the clustering hierarchy by thresholding techniques or based on measures of cluster validity. Density-based techniques usually define initial clusters by seeking high-density points (by simple use of K-means clustering [28], applying kernel-based density estimation [18] or using density gradient estimation, the modes being detected with the hill climbing *mean shift* procedure [29], [30]), density similarity guiding the merging process; simple thresholding [28] or cluster validity indices weighting intercluster connectivity and cluster isolation (low-density regions separating clusters) [18] are used to select a clustering. In the work in [30], an initial random space tessellation is produced to which a *mean shift* procedure is applied to detect cluster centers. A two phase clustering algorithm is presented in [31], according to which initial subclusters are obtained using a graph partitioning technique to the K-nearest neighbor graph of the data set, followed by a dynamic merging of subclusters under a hierarchical agglomerative framework. The density-based clustering algorithm presented in [32] explores the idea of intracluster homogeneity and uniformity, working on links from a complete graph.

1.2 Goals and Outline of the Paper

In this paper, we address the problem of cluster defining criteria under a model-based framework. A new cluster isolation criterion, briefly outlined in [33], underlying a hypothesis of smooth dissimilarity increments between neighboring patterns, is presented and discussed. It is shown that dissimilarity increments between neighboring patterns within a cluster have a smooth evolution, whose statistical distribution can be modeled by an exponential density function. Dissimilarity increments, by means of their statistical model, characterize context. The proposed isolation criterion is supported on a pair-wise context analysis. This isolation criterion is merged in a hierarchical agglomerative clustering algorithm, producing a data partitioning and simultaneous accessibility to the intrinsic data interrelationships in terms of a dendrogram-type graph. The structure of the obtained dendrogram, unlike conventional hierarchical clustering methods, is constrained by the isolation criterion, expanding the range of pattern structures handled by these methods, namely, situations containing both sparse and dense clusters. Additionally, the problem of deciding the number of clusters is subsumed and intrinsically dictated by the criterion.

Section 6 studies the distribution of dissimilarity increments, supporting the smooth evolution hypothesis, and outlines the new cluster isolation criterion (Section 2.2). Critical evaluation and mathematical manipulation of the parametric context model—exponential distribution—leads to the definition of an intrinsic isolation parameter (Section 2.3). A hierarchical agglomerative algorithm adopting this criterion is described in Section 3. The novelty of the proposed method and its relation to work in the literature is outlined in Section 4. The characteristics of the new method are analyzed and illustrated through a set of examples (Section 5), covering synthetic data (random data, Gaussian mixtures, concentric patterns, and clusters of arbitrary shape and size) and examples from the UCI Machine Learning Repository [34] (Iris data and the Wisconsin Breast Cancer Data Set). Results are compared with the single-link method and the k-means algorithm. A discussion of the proposed method with the SL and the K-means algorithm is presented in Section 6. Conclusions are drawn in Section 7.

2 Smoothness Hypothesis and Cluster Isolation Criterion

Let X be a set of patterns, and x_i represent an element in this set. Assume that interpattern relationships are measured by some dissimilarity function, $d(.,.)$. The definition of $d(.,.)$ is problem and data representation dependent; it may be, for instance, the Euclidean distance for patterns in multidimensional feature spaces; string edit distances [35], [36], [37], [38] are commonly used for quantifying resemblance between string patterns.

The proposed cluster isolation criterion is derived from the following intuitive concepts and assumptions:

- A cluster is a set of patterns sharing important characteristics, defining a context.
- Dissimilarity between neighboring patterns within a cluster should not occur with abrupt changes.
- The merging of well separated clusters results in abrupt changes in dissimilarity values.

The first concept states that a cluster gathers interrelated patterns, the pattern dependence profile being a characteristic of the cluster, thus defining a context; this enables its distinction from other clusters. The last two items state a hypothesis of smooth evolution of dissimilarity changes, or increments, between neighboring patterns within a cluster, nonsatisfaction of this condition being associated with cluster isolation. This smoothness hypothesis is the genesis of the proposed cluster isolation criterion, the dissimilarity increments measuring continuity within a cluster.

2.1 Distribution of Dissimilarity Increments

Consider a set of patterns X. Given x_i, an arbitrary element of X and some dissimilarity measure, $d(.,.)$, between patterns, let (x_i, x_j, x_k) be the triplet of nearest neighbors, obtained as follows:

$$\begin{aligned}(x_i, x_j, x_k) - \text{nearest neighbors}\\ x_j : j = arg\ \min_l \{d(x_l, x_i) \quad , l \neq i\}\\ x_k : k = arg\ \min_l \{d(x_l, x_j) \quad , l \neq i, l \neq j\}.\end{aligned}$$

The dissimilarity increment between the neighboring patterns is defined as

$$d_{inc}(x_i, x_j, x_k) = |d(x_i, x_j) - d(x_j, x_k)|,$$

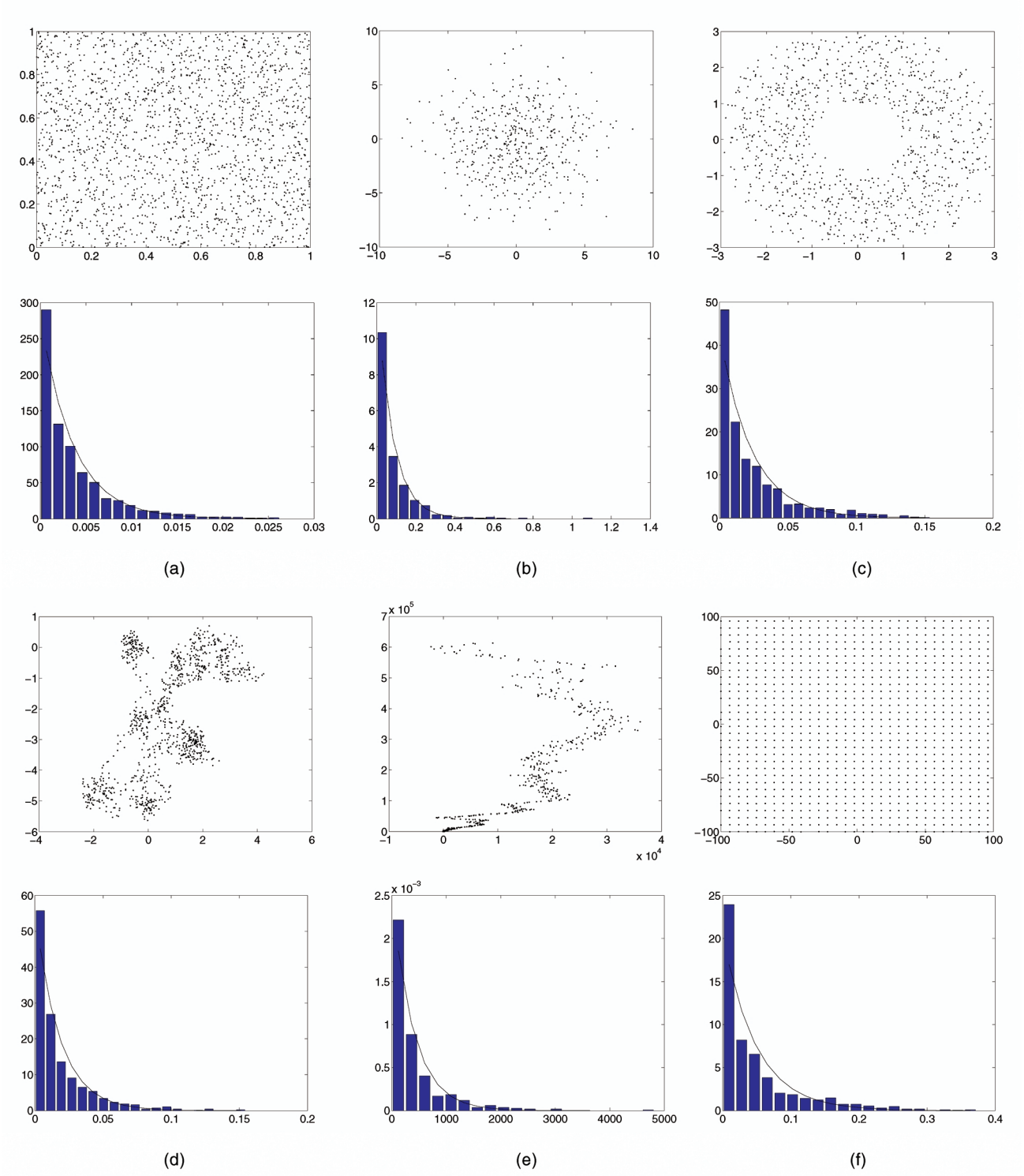


Fig. 1. Histograms (bar graphs) and fitted exponential distributions (solid line curves) of the dissimilarity increments computed over neighboring patterns in the data. The Euclidean distance was used as the dissimilarity measure. (a) There are 2,000 uniformly distributed patterns within a square, (b) 500 patterns generated from a Gaussian distribution ($N([0,0],[10\ 0;0\ 10])$), (c) ring-shaped data (1,000 random patterns), (d) 1,000 patterns generated according to the stochastic model: $y(k+1) = y(k) + n_1(k)$, $x(k+1) = x(k) + n_2(k)$, with $n_1(k)$, $n_2(k)$ being noise uniformly distributed in the interval $[-.25; .25]$, (e) directional expanding data generated by the model: $x(k+1) = x(k) + n_s(k)k$, $y(k+1) = y(k) + n(k)$, where n_s and $n(k)$ represent uniform noise in the range $[-10; 10]$ and $[0; 10]$, respectively, and (f) grid corrupted by zero mean Gaussian noise, with standard deviation 0.1.

which can be seen as the first derivative of the dissimilarity function at the first point of the ordered list of neighboring samples.

There is experimental evidence that the increments of the dissimilarity measure between neighboring patterns, as defined above, typically exhibit an exponential distribution, $p(x) = \beta \exp^{-\beta x}, x > 0$, as illustrated in Fig. 1. This figure plots histograms and corresponding fitted distributions of dissimilarity increments for a variety of data sets. Two-dimensional examples were chosen for simplicity of representation:

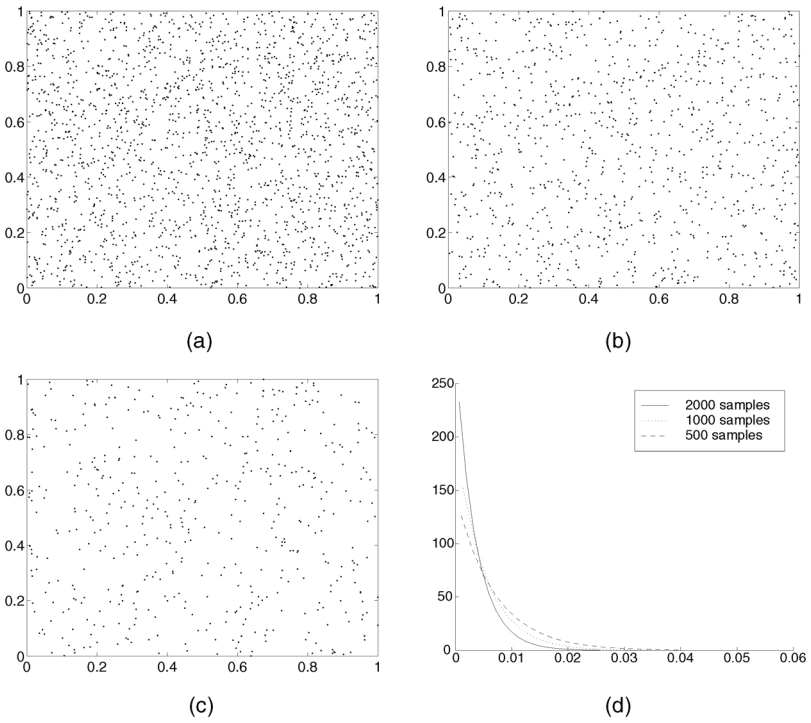


Fig. 2. Fitted exponential distributions for the dissimilarity increments of 2D data, randomly generated from a uniform distribution: (a) 2,000 samples, (b) 1,000 samples, and (c) 500 samples. (d) Steep exponentials (higher β parameter) correspond to high-density patterns.

- random samples (uniform distribution),
- 2D Gaussian process,
- noisy ring shaped pattern,
- 2D stochastic process,
- directional expanding pattern, and
- grid corrupted by Gaussian noise.

The Euclidean distance is used as the dissimilarity measure in these examples.

As shown in Fig. 2d, the statistical distribution of the dissimilarity increments within the same context or data formation model (cluster) has a smooth evolution, where the parameter β of the fitted exponential probability density function characterizes data sparseness. It can be observed that distinct data generation models lead to very similar curves (for instance, patterns in Figs. 1c and 1d), while an increasing number of observations from the same process (corresponding to decreasing data dispersion levels) results in increasing values for the parameter β of the exponential distribution (see Fig. 2).

Thus, by adopting the dissimilarity derivatives as features for context characterization, a single parametric model (exponential distribution) is obtained for distinct cluster shapes or data generation paradigms. When considering well-separated clusters, it is clear that dissimilarity increments between patterns in different clusters are positioned far on the tail of the distribution associated with the other cluster. We explore this property in defining a cluster isolation criterion in the next section.

2.2 Isolation Criterion

We extend the previous concept of dissimilarity increments between neighboring patterns to define the concept of *gap* between clusters.

Let C_i, C_j be two clusters candidate for merging, as the ones shown in Fig. 3, and consider the nearest pattern pair, (x_i, x_j), linking these clusters, such that $x_i \in C_i$ and $x_j \in C_j$ ($x_i \equiv x_{12}$ and $x_j \equiv x_{18}$ Fig. 3). We shall represent the dissimilarity between these patterns, $d(x_i, x_j)$, as $d(C_i, C_j)$ (corresponding to the distance between the two clusters, according to the nearest-neighbor rule). Let x_k be the nearest neighbor of x_i within C_i (pattern x_3 in Fig. 3), and let $d_t(C_i) = d(x_i, x_k)$. The triplet (x_k, x_i, x_j), therefore, corresponds to neighboring patterns. We define *dissimilarity increment* or *gap* between clusters i and j as the asymmetric increase in the dissimilarity value, needed in order to allow the data association into a single cluster:

$$gap_i = |d(C_i, C_j) - d_t(C_i)|. \quad (1)$$

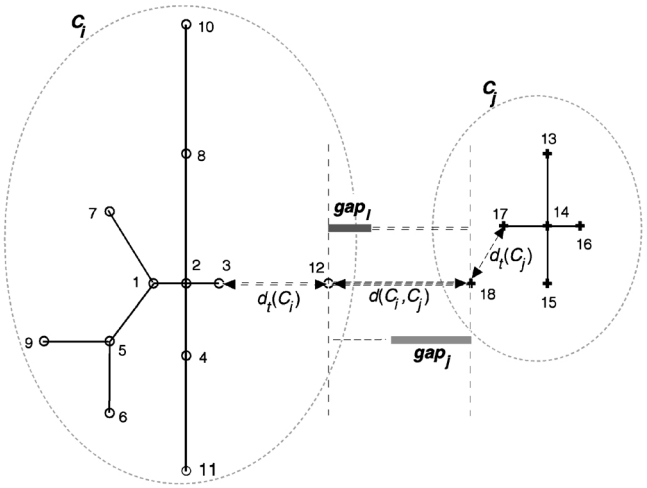


Fig. 3. Definition of gap. The figure shows 18 two-dimensional patterns grouped in two clusters. The patterns are linked by the minimum spanning tree, adopting the Euclidean distance as the edge weight.

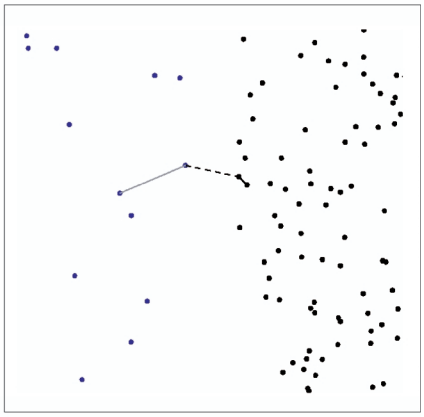


Fig. 4. Touching classes with distinct densities.

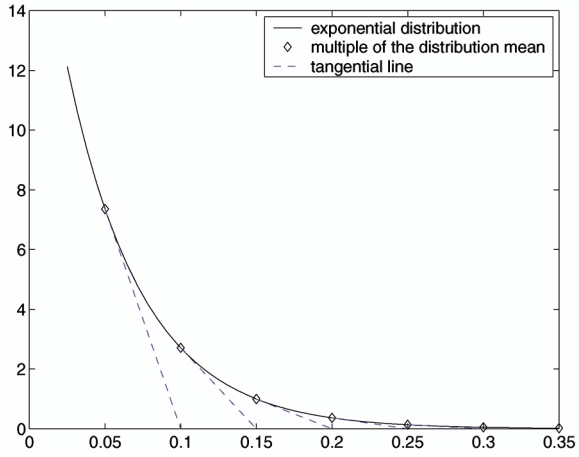


Fig. 5. Defining a threshold on the gap values (x axis). Dots are located on points which are multiple of the distribution mean, $\frac{1}{\beta} = 0.05$ and dashed lines are tangents at those points. The crossing at the x axis occur at points $\frac{i}{\beta}$, i being a positive integer. Values for i in the range $[3, 5]$ cover the most significant part of the distribution.

In a similar way, we find x_l (x_{17} in Fig. 3), the nearest pattern to x_j belonging to C_j, and define *gap* between cluster j and i: $gap_j = |d(C_i, C_j) - d_t(C_j)| = |d(C_i, C_j) - d(x_j, x_l)|$.

Dissimilarity increments between neighboring patterns within a cluster is a measure of pattern continuity. The statistical distribution of dissimilarity increments is modeled by an exponential distribution. Let $\hat{\mu}_i = \frac{1}{\hat{\beta}_i}$, $\hat{\mu}_j$, be the average of dissimilarity increments in cluster C_i and C_j, respectively. Tails of these distributions correspond to patterns in frontier or borderline situations, where continuity is broken. The gaps, gap_i, gap_j, represent the increase in neighboring pattern distances needed in order to join the two clusters, measuring intercluster continuity, as seen from each cluster perspective. If the two clusters are well separated, these gaps will have high values (compared to intracluster statistics), being located on the tails of each cluster statistic, and corresponding to a discontinuity in both clusters structure. In situations of touching clusters with distinct densities, as in the example shown in Fig. 4, context analysis is needed in order to identify the clusters. The dashed line in Fig. 4 links the nearest-neighbor patterns connecting the two clusters; remaining lines link the intracluster nearest neighbors to each of these elements. From this figure, it is intuitive to see that the element from the cluster on the right could naturally be included in the left cluster since the increment ($gap_1 = 0.0150$) is small compared to the intracluster statistic ($\hat{\mu}_1 = 0.0268$). From the context of the cluster on the right, however, the dissimilarity increment ($gap_2 = 0.0542$) is large compared to the average dissimilarity increments within this cluster: $\hat{\mu}_2 = 0.0068$. Therefore, taking the one-sided perspective of cluster C_1, the two clusters could be merged; from the context of C_2, the clusters are isolated.

The cluster isolation criterion consists of setting a limit on the dissimilarity increments, such that most of the patterns exhibiting the same statistical structure or model (densely or sparsely connected) are included in the same cluster, while all others, not satisfying this smoothness hypothesis, are rejected:

- *Let C_i, C_j be two clusters which are candidates for merging, and let μ_i, μ_j be the respective mean values of the dissimilarity increments in each cluster. Compute the increments for each cluster, gap_i and gap_j, as defined in (1). If $gap_i \geq \alpha\mu_i$ ($gap_j \geq \alpha\mu_j$), isolate cluster C_i (C_j) and continue the clustering strategy with the remaining patterns. If neither cluster exceeds the gap limit, merge them.*

Notice that the above criterion can be regarded as a context-dependent cluster isolation rule where the context is modeled by the parametric distribution of dissimilarity increments. The isolation rule consists of comparing the value of the dissimilarity increment, seen from the context of each cluster, with a dynamic threshold, $\alpha\mu_i$, computed from this context; inconsistency of gap values in a given context (cluster) determines the isolation of that cluster.

The design parameter, α, constrains the degree of isolation; values in the range 3 to 5 provide reasonable choices, as justified in the next section.

2.3 Setting the Isolation Parameter

As seen previously, the structure of the dissimilarity increments within a cluster is summarized by an exponential distribution; the parameter β of this distribution thus characterizes each cluster. Well-separated clusters are clearly identified by the analysis of these distributions, as samples not belonging to a given cluster will be placed far in the tail of the cluster distribution. A reasonable choice for the isolation parameter, α, is to set it at a point on the tail that does not reject a significant amount of data nor does it allow grouping of patterns that are clearly atypical.

Theoretical analysis of the exponential distribution leads to the following interesting result (see Appendix A): The crossing of the tangential line, at points which are multiples of the distribution's mean value, $i \times \frac{1}{\beta}$, with the x axis, is given by $(i + 1) \times \frac{1}{\beta}$; this is shown in Fig. 5.

Therefore, setting the threshold, α, to some multiple of the distribution mean, i.e., α inside the interval 3 to 5 is a reasonable choice. In examples throughout the paper, the typical value used is $\alpha = 3$.

3 Hierarchical Clustering Algorithm

In this section, we incorporate the cluster isolation criterion described in Section 2.2 in a hierarchical agglomerative clustering algorithm. Each cluster, C_i, is characterized by: $\mu[i]$—the estimate of the mean value of the dissimilarity increments within the cluster; $jumps[i]$—the number of elements used in this estimate. The algorithm starts with

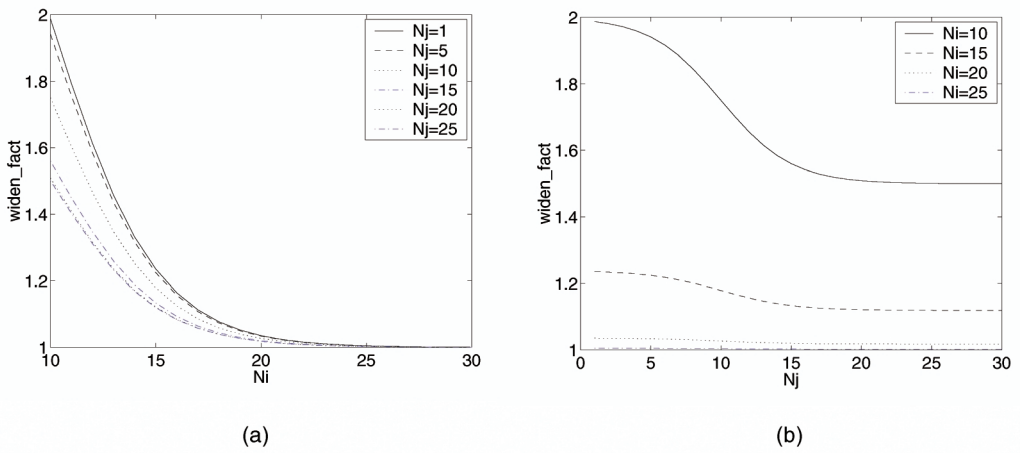


Fig. 6. Amplification term $widen_{fact}$ associated with the estimate $\hat{\mu}$ for cluster C_i. (a) Amplification factor as a function of the number of terms used in the computation of the gaps distribution mean for cluster C_i ($\beta = 1$). (b) Reinforcement of the amplifying term as a function of the number of elements in cluster C_j ($\beta = 1$).

each pattern in a cluster, the dissimilarity matrix between pattern pairs being computed. It evolves by selecting the most similar pair of clusters and applying the cluster isolation criterion from each cluster context; clusters are thus either isolated (one or both) and frozen on the dendrogram, or merged; frozen clusters are not available for further merging. Statistics $\mu[i]$ are updated along the merging process.

Estimates of the mean values $\mu[i]$ are not reliable for very small cluster sizes; this may lead to premature isolation of clusters. In order to overcome this situation, widening of the isolation parameter α for small cluster sizes may be adopted [39]; alternatively, inhibition of cluster isolation actions may be implemented when clusters are very small [33]. In this paper, we replace the term $\alpha\hat{\mu}_i$ by the dynamic threshold

$$t_{dyn\_C_i}(\alpha, \hat{\mu}_i, ni, nj) = \alpha\hat{\mu}_i \times widen_{fact}(ni, nj) + delta_{fact}(ni). \tag{2}$$

Expression (2) has two terms. The first term increases the value of the estimate $\hat{\mu}_i$ by multiplying it by a factor greater than or equal to 1, $widen_{fact}(ni, nj)$, where $ni \equiv jumps[i]$ and $nj \equiv jumps[j]$ are the number of elements available for the computation of the distribution means for cluster C_i and C_j, respectively. We define the amplifying factor $widen_{fact}(ni, nj)$ as a monotonous decreasing function of ni, nj:

$$widen_{fact}(ni, nj) = 1 + \beta \times \underbrace{\left(1 - \frac{1}{1 + e^{-.4(ni-10)}}\right)}_{f_1(ni)} \times \underbrace{\left(2 - \frac{1}{1 + e^{-.4(nj-10)}}\right)}_{f_2(nj)}. \tag{3}$$

The reasoning underlying (3) is the following (see Fig. 6). If cluster C_i has few samples, the estimate $\hat{\mu}(C_i)$ should be enlarged to compensate for possible underestimation of the true distribution mean; this widening effect smoothly vanishes as the number of terms ni used in the computation of the estimate $\hat{\mu}(C_i)$ increases (Fig. 6a), which is modeled by the term $f_1(ni)$, a sigmoid-like function. The term $f_2(nj)$ expresses the reinforcement of the widening effect when the number of elements in the competing cluster C_j is also low (Fig. 6b), taking values greater or equal to 1. When both clusters have low cardinality the combined action of f_1 and f_2 favors clusters merging. When cluster C_i has already a sufficiently large number of elements, the estimate of $\hat{\mu}(C_i)$ is considered to be reliable and term $f_1(ni)$ tends to zero, thus annihilating the influence of term f_2 (the size of cluster C_j becomes irrelevant—see Fig. 6a, $ni \geq 25$). In (3), β is a scaling parameter (default value: 3).

When the number of elements available for the estimation of the dissimilarity increments statistic, n_i, is extremely low (such as when the number of cluster's samples is less than 10), the estimate for the μ parameter is very poor. Applying a multiplicative factor to the threshold term may not solve the underestimation problem in this situation, in particular, when $\hat{\mu}$ is near zero. The second term in (2), with large values vanishing for $ni = 10$, boosts near zero estimates for extremely small sized clusters:

$$delta_{fact}(ni) = big_{val} \times \left(1 - \frac{1}{1 + e^{-10(ni-5)}}\right), \tag{4}$$

where big_{val} is a large positive number.

In order to compute the gap between clusters, one needs to know the distances between nearest neighbor patterns. Using the nearest-neighbor rule for updating intercluster dissimilarity, $d(C_i, C_j)$ gives the desired distance between nearest neighbors in each cluster. Considering that most similar patterns are joined first, dissimilarity values growing along the evolution of the clustering algorithm, we will approximate the exact value of the gap by $gap_i = d(C_i, C_j) - d_t[i]$, with $d_t[i]$ representing the dissimilarity in the last merging performed in cluster C_i (see Fig. 7). This approximation prevents further computation of nearest neighbors in each cluster, leading to a computationally more efficient algorithm.

The following gives a schematic description of the clustering algorithm.

Input: N samples; α (default value is 3).
Output: Data partitioning.
Steps:

1. Set: $Final\_clusters = \phi$; $n = N$;
   Put the ith sample in cluster $C_i, i = 1, \ldots, n$;
   $Clusters = \bigcup_i C_i, i = 1, \ldots, n$;
   $d_t[i] = \mu[i] = jumps[i] = 0, i = 1, \ldots, n$;
2. *If* ($Clusters == \phi$) or ($n == 1$)

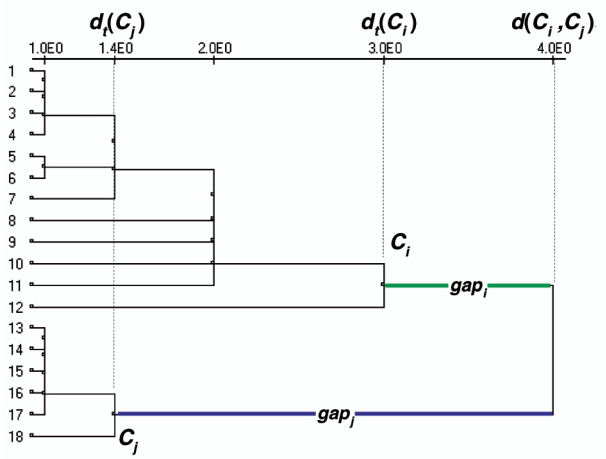


Fig. 7. Definition of gap on the dendrogram produced by the single-link method for the data in Fig. 3.

then stop, returning the clusters found in $Final\_clusters \bigcup Clusters$;
else continue.

3. Choose the most similar pair of clusters (C_i, C_j) from $Clusters$. Let
$gap_i = d(C_i, C_j) - d_t[i] \qquad ni = jumps[i]$
$gap_j = d(C_i, C_j) - d_t[j] \qquad nj = jumps[j]$
4. *If* $((gap_i < t_{dyn\_C_i}(\alpha, \mu[i], ni, nj))$ and $(gap_j < t_{dyn\_C_j}(\alpha, \mu[j], nj, ni)))$
then
join the clusters C_i, C_j into cluster $C_{i,j} : C_{i,j} = C_i \bigcup C_j$
Let I be the index for the merged cluster;
Replace C_i, C_j by $C_{i,j}$ in $Clusters$;
$d_t[I] = d(C_i, C_j)$;
$jumps[I] = jumps[i] + jumps[j] + 2$;
$\mu[I] = \mu[i]\frac{jumps[i]}{jumps[I]} + \mu[j]\frac{jumps[j]}{jumps[I]} + \frac{gap_i + gap_j}{jumps[I]}$;
Go to step 2.
else continue.
5. *If* $(gap_i \geq t_{dyn\_C_i}(\alpha, \mu[i], ni, nj))$
then set $Final\_clusters = Final\_clusters \bigcup C_i$;
Remove C_i from $Clusters$;
$n = n - 1$.
end if
If $(gap_j \geq t_{dyn\_C_j}(\alpha, \mu[j], nj, ni))$
then set $Final\_clusters = Final\_clusters \bigcup C_j$;
Remove C_j from $Clusters$;
$n = n - 1$.
end if
Go to step 2.

4 Related Work

The distinctive feature of the proposed scatter measure, which forms the basis of the cluster isolation criterion, consists of analyzing and modeling dissimilarity increments in neighboring patterns, instead of statistical or geometrical manipulations of the dissimilarity values between patterns. Dissimilarity increments measure continuity within a cluster. The work presented in [32] explores the concept of uniformity to detect clusters with similar interior distances. It works on links from a complete graph. Initial clusters are defined by gathering links differing in length by no more than a given threshold. The length difference within these clusters, which is similar to the dissimilarity increment proposed in this paper, has an a priori fixed upper value; in our method, increments are compared to an adaptive threshold, which depends on individual cluster statistics. The merging process proposed in [32] is based on the comparison of intracluster average distances; in our method, the distribution of increments within a cluster is modeled by a parametric model (exponential distribution), the parameter summarizing cluster structure being the average value of increments between neighboring patterns. Increment values computed from the nearest pair of patterns in distinct clusters are compared to each cluster statistic to decide for merging.

The proposed cluster isolation criterion has been evaluated in the context of hierarchical agglomerative clustering, adopting a nearest-neighbor rule for measuring the similarity between clusters. This new algorithm is therefore closely related to graph-theoretical methods, in particular, with the single-link method: both methods start with single element clusters, merging most similar clusters first, and updating the similarity matrix according to the nearest-neighbor rule. A major distinction between the two methods is that the standard SL method uses a fixed threshold on dissimilarity values for cutting the resulting dendrogram, while the herein proposed method uses an adaptive threshold on dissimilarity first derivatives, based on the computation of intracluster statistics of dissimilarity increments. These statistics are scatter measures, characterizing density of clusters. With the proposed cluster isolation criterion, the new algorithm is able to identify clusters with different densities, which requires special treatment when using graph-theoretical methods, such as detecting and removing denser clusters, and then clustering the remaining patterns. With the proposed approach, this situation is easily handled as, according to the asymmetric isolation criterion, denser clusters are identified and frozen on the dendrogram, the clustering process based on dissimilarity increments proceeding with the remaining data. Some authors have adopted postprocessing of the dendrogram produced by the SL method [26], [27] or, equivalently, processing of the minimum spanning tree (MST), in order to obtain a final data partition. Zhan [21] proposed a technique for the identification of clusters from a minimum spanning tree by removing inconsistent links based on the comparison of the link distance (dissimilarity between linked patterns) with the average of nearby link distances on both sides of the link. Inconsistent links removal is therefore based on local dissimilarity statistics; our method, however, evaluates overall clusters statistics (of dissimilarity increments instead of distances) along the clustering process, eventually conditioning the final form of the dendrogram. This dynamic construction of the dendrogram, the final topology being conditioned by intracluster statistics, opposes to the static behavior of the above methods, based on postprocessing of structures. A dynamic hierarchical agglomerative procedure is proposed in [31]. In that work, however, similarity between clusters combines interconnectivity and relative closeness measures based on the K-nearest neighbor graph of the data set, isolation criteria consisting of the comparison of the similarity value with a user specified parameter, controlling, simultaneously with the K parameter, the characteristics of the desired clusters.

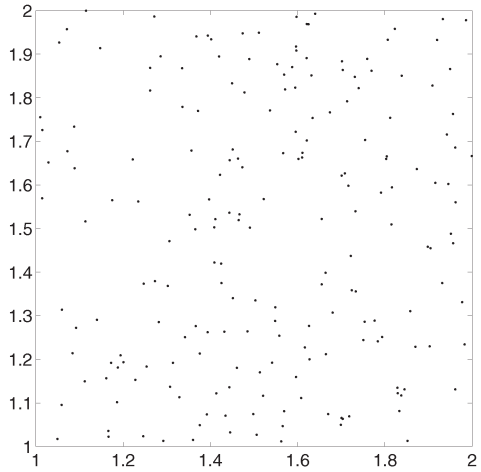


Fig. 8. Two-dimensional projection of 200 samples in a nine-dimensional space. A single cluster is identified for $\alpha \geq 1$.

5 Analysis and Illustrative Examples

The analysis of the proposed criterion will be carried out through a set of examples. Clustering results will be compared with two very popular and well-known strategies: the single-link method and the k-means algorithm.

5.1 Clustering of Random Data

The first question that should be asked whenever a clustering algorithm is to be applied concerns the clustering tendency of the data. Does the data entail some structure, ultimately exposed by some clustering algorithm or is it random data? A typical approach consists of applying some test for randomness before further analysis by clustering be performed. This is a wise strategy as most clustering algorithms can impose inappropriate clustering structure in the presence of unstructured or random data.

In this section, we analyze how the algorithm behaves in the presence of patterns randomly generated from uniform or Gaussian distributions. These examples are also used to illustrate the effect of the parameter α on the data partitioning.

There were 25 tests performed with random data uniformly distributed in d-dimensional hypercubes, with d in the interval $[2; 10]$ and the number of points being randomly selected from the interval $[100; 1,000]$. A typical example is shown in Fig. 8. Results on these trials show a single cluster, for $\alpha \geq 3$. When using $\alpha \leq 2$ (which is in disagreement with the analysis performed in Section 2.3), a large cluster and additional spurious, small sized clusters were obtained due to too narrow limits on the exponential distribution: With $\alpha = 2$, a single cluster was obtained in 16 data sets; in the remaining nine data sets, a spurious cluster was obtained with one or two patterns (no cluster size limiting rule was applied). The occurrence of spurious, small sized clusters increases for $\alpha = 1$ (12 cases).

Data sets drawn from Gaussian distributions produce nonhomogeneous scatter plots, with a small percentage of patterns dispersed around a high-density nucleus. Typical clustering results obtained with the proposed algorithm consists of two clusters, the atypical data being gathered in a cluster, as illustrated in Fig. 9. In order to merge all the data in a single cluster, higher values of α are usually required ($\alpha \geq 4$).

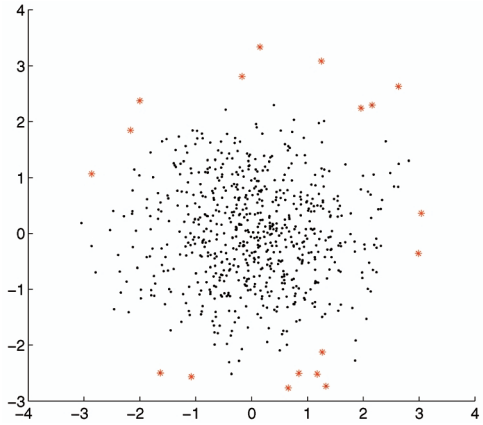


Fig. 9. Clustering 700 patterns from a bivariate Gaussian. The plot corresponds to $\alpha = 3$. A single cluster is identified for $\alpha \geq 4$.

5.2 Mixture of Gaussians

The separation of Gaussian mixtures with equal covariance matrix is illustrated in a study of the number of clusters found as a function of the Mahalanobis distance between two Gaussian distributions. To this end, data was generated with sample sizes of 100, 200, 300, 400, and 500 per class, each class obeying a two-dimensional Gaussian distribution; separation between the classes is measured by the Mahalanobis distance (MD). Each experiment consists of the generation of data from two classes with a given MD and cluster size. There were 15 realizations of the experiment performed for each situation (75 experiments for each MD, when considering variable cluster sizes).

The proposed algorithm consistently separated the data sets in two large clusters for Mahalanobis distances higher than 5, a third cluster being formed gathering spurious data (see Fig. 10a for a typical example). For MD = 5, the two natural clusters were merged most of the times (as in Fig. 10b); no cluster separation was obtained for lower MD values. The single link method produced similar results. The K-means algorithm always finds two clusters (with $k = 2$), outperforming the proposed method for MD < 6, although results are dependent on centroid initialization (see Fig. 10c).

With the examples provided above, where clusters exhibit identical covariance, it would appear that, among the three approaches evaluated, the k-means is the best performing method, allowing separation of clusters for lower values of the MD. The Mahalanobis distance, however, is not an adequate index to characterize the performance of the methods, as they present remarkably different behaviors, for instance, in situations of uneven data sparseness. This aspect is put in evidence in the example depicted in Fig. 11a, concerning a mixture of two Gaussians with identical mean (MD = 0) and disparate covariance matrices. In this case, the situation of coinciding cluster centroids is responsible for the failure of the k-means algorithm. The single link method is unable to handle the distinct data sparseness, joining patterns around the denser cluster, and breaking down the low-density cluster into a set of small sized clusters (Fig. 11b). The proposed method, on the other hand, identifies two concentric clusters, which is a partition consistent, for instance, with a Bayes classifier for the given data. The ability to separate the overlapping clusters results from the distinctive feature of unbalanced cluster densities, which is exploited by the method. Cluster separability, as addressed by the

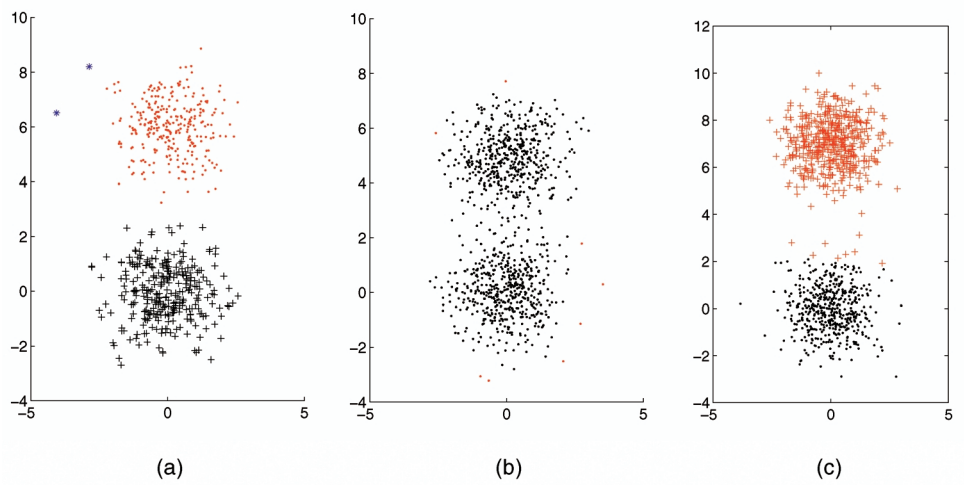


Fig. 10. (a) MD = 6, (b) MD = 5, (c) MD = 7, K-means results. Dependency of clustering of mixtures of two Gaussians with unit covariance on the Mahalanobis distance (MD) between the distributions means: $\sqrt{(\mu_1-\mu_2)^T(\mu_1-\mu_2)}$. $\alpha = 3$. (a) and (b) are obtained with the proposed method and (c) is the result of inadequate initialization of the K-means algorithm, leading to poorer data partitions.

isolation criterion, does not necessarily require well-separated clusters, as shown in this example.

5.3 Concentric Clusters

The situation of concentric patterns is examined in this section with several experiments with ring-shaped patterns. Fig. 12 shows a series of increasingly difficult clustering problems, by either tightening the gaps between the rings, or increasing the number of clusters. The k-means method is unable to handle this type of patterns, imposing globular shaped clusters on the data. The single link method can only cope with situations where the separation between clusters is higher than the maximum within cluster distance between neighboring patterns. The proposed method consistently outperforms both methods in all situations, in terms of correct identification of the number of clusters and data assignment into clusters (plots on the left of Fig. 12).

5.4 Arbitrary Shape Clusters

A complex composition of 739 patterns organized into eight irregularly shaped, variable sized clusters is proposed here for analysis. Data comprises (see Fig. 13) two concentric ring-shaped clusters (200 patterns each); two parallel bar-delimited groups of random patterns (uniform distribution), with 100 patterns per class; two neighboring sets of 50 points, drawn from Gaussian distributions with distinct covariances; a star-shaped cluster with 29 patterns; 10 equally spaced points forming an outer circle, intermingled with bar and ring shaped clusters.

Figs. 13 and 14 present partitions of this complex pattern composition using the three methods. As shown in Fig. 13a, although the reasonable choice for α is 3 or 4, there is a large range of values $(2 < \alpha < 9)$ for which adequate pattern associations are produced. Values above the upper limit of the interval for α leads to the gathering of clusters (the bar-shaped ones being the first to be merged), while values below the lower limit break down sparser clusters (the star-shaped is the first candidate for splitting) into spurious, lower size clusters.

Accounting for data proximity directly, and comparing this with a threshold, a design parameter for which no a priori selection criteria exists, the single link method is unfit to handle the variability of density of data, splitting sparse groups of data into, often, single point clusters (Fig. 13b). The k-means algorithm does not cope with irregularly shaped and/or concentric patterns, producing odd pattern associations (see Fig. 14).

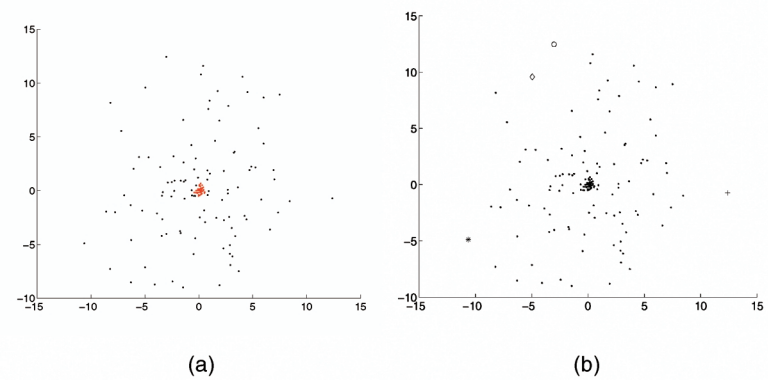


Fig. 11. Clustering of a mixture of two overlapping Gaussians with zero mean and covariance matrices $20I$ (100 patterns) and $0.1I$ (50 patterns), with I being the identity matrix. (a) Proposed method, $\alpha = 3$. (b) Single-link method, $th = 3$.

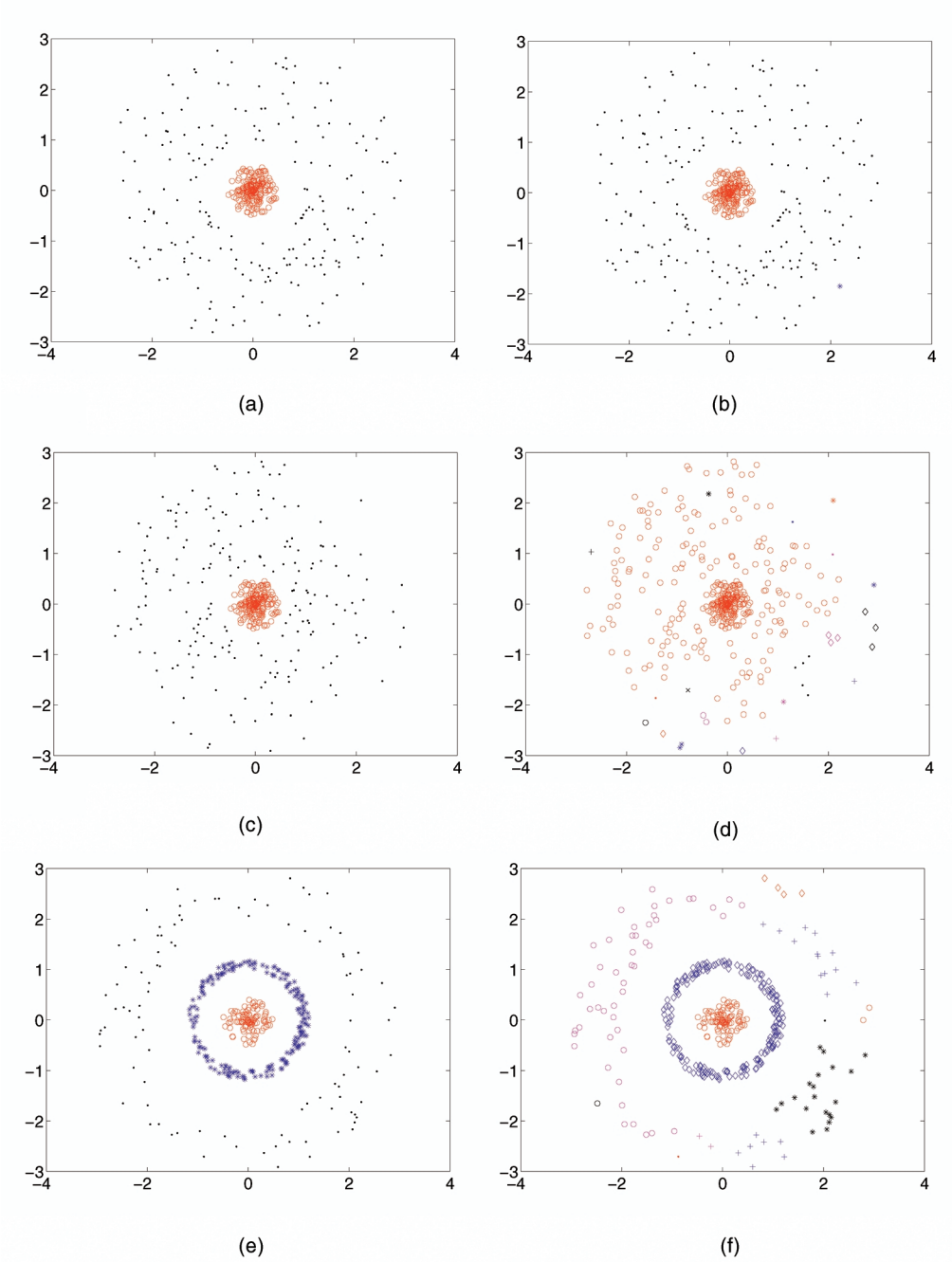


Fig. 12. Concentric patterns. Left column plots: results with the proposed method, $\alpha = 3$. Right column: single-link method. Threshold are as follows: (b) $th = .53$ (three clusters). (d) $th = .4$ (21 clusters). (f) $th = .49$ (12 clusters).

5.5 Iris Data

The Iris data set consists of three types of Iris plants (Setosa, Versicolor, and Virginica) with 50 instances per class, represented by four features. This data, extensively used in classifier comparisons, is known to have one class (Setosa) linearly separable from the remaining two classes, the latter not being linearly separable.

Two clusters are found with the proposed method ($1 < \alpha < 10$), corresponding to a merging of types Virginica and Versicolor, and a single cluster for the Setosa type. Results are therefore comparable with the ones obtained with the single link method, according to which the same data partition is obtained by adequate selection of a threshold on the dendrogram (see Fig. 15). A similar result is reported in [40] where the proposed criteria for selecting the number of cluster leads to two clusters. The K-means method gave the best clustering results, one cluster including the Setosa type, and the other types of plants being separated in two clusters, with an overall error rate of 11 percent.

5.6 Breast-Cancer Data

The final test data consists of the Wisconsin Breast Cancer Data set available at the UCI Machine Learning Repository [34]. Data of two types (benign and malignant, 444 and 239 samples, respectively) are represented by nine features (fully instantiated, class labels are ignored in clustering).

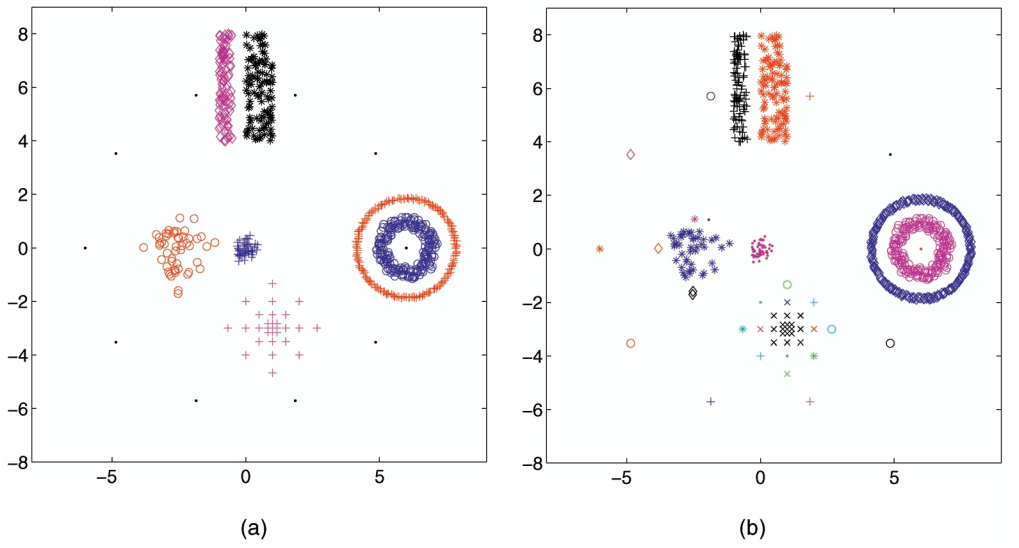


Fig. 13. Clustering of complex cluster structures. (a) Clustering using the proposed method: for $2 < \alpha < 9$, our algorithm correctly identifies eight clusters. (b) Single-link method with $th = .5$: 33 clusters are identified.

Fig. 16 represents the dendrogram produced by the single-link method by ordering data according to their class labels: benign patterns are on the right side of the graph. As shown, this method is not able to differentiate between the two types of data: Simple thresholding on this graph leads to a cluster with most of the samples and spurious single pattern clusters. It also obvious that the two classes are not well separated but exhibit different structures of interpattern distances. Therefore, with the proposed method, a single cluster is obtained for $\alpha = 3$ (which assumes good cluster separation), but by lowering this threshold to the value 1 two clusters are identified. By comparing the partitions, thus obtained with the patterns class labels, a recognition rate of 96.63 percent was achieved (23 samples were misclassified). This result compares favorably to the cluster center based methods reported in [40] and [41], where performances obtained on the same data were 94.28 percent and 95.5 percent, respectively. Using the k-means algorithm, with k = 2, results are dependent on the initial cluster centers. After several experiments, the best accuracy achieved was 96.49 percent (24 samples misclassified) which is comparable with the result obtained with the proposed method. The corresponding cluster centers are indicated in Table 1.

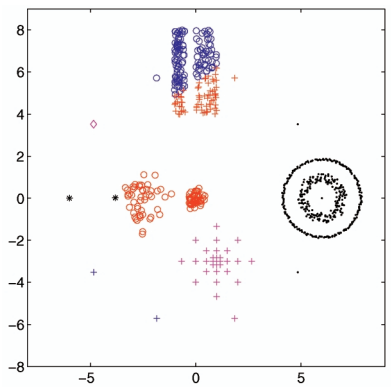


Fig. 14. K-means partition when imposing eight clusters.

6 Discussion

The k-means algorithm [2] is a square-error partitional method. Its major drawbacks are the necessity of a priori knowledge of the number of clusters, dependence of the partition found on the centroids initialization, and an inability to identify irregularly shaped clusters. While methods have been proposed to handle the selection of the number of clusters [11], the centroid-based error computation leads to globular shaped clusters. The inadequacy in identifying other shapes has been illustrated, for instance, in Fig. 14. More recently, k-means derived methods have been proposed that can identify specific shapes in patterns [13], [16], the models of which (line, circle, ...) being defined in advance, not data driven.

The single-link method, manipulating a dissimilarity matrix between patterns, imposes a hierarchical structure on data, graphically displayed as a dendrogram; it is able to identify irregularly shaped clusters whenever the minimum dissimilarity between clusters in higher than within cluster dissimilarity between neighboring patterns. As illustrated in the examples provided above, unbalanced density clusters are not adequately handled by this method. Also, an undesirable characteristic of the method consists of the "chaining effect," meaning the gathering of distinct clusters whenever there is a chain of data points bridging the gap.

The method outlined in Section 3 incorporates the proposed cluster isolation criterion into a hierarchical agglomerative type algorithm. Although it provides a dendrogram type graph describing the structure of data, the new algorithm is a partitional procedure that intrinsically identifies the number of clusters without necessity of ad hoc definition or a priori knowledge of design parameters. The distinctive features of the new method responsible for overcoming difficulties not solved by the previous methods are now discussed.

The first aspect that distinguishes the proposed method from other clustering methods is the exploration of the first derivative of the dissimilarity (called *dissimilarity increments* between neighboring patterns, or *gaps*, when considering

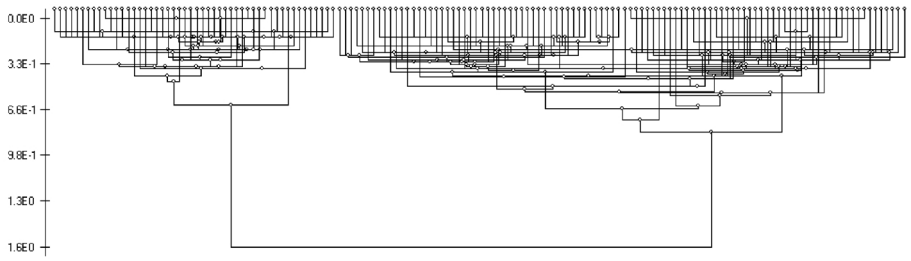


Fig. 15. Dendrogram produced by the single-link method for the iris data. From the graphs, the iris types Virginica and Versicolor are undistinguishable-rightmost 100 samples, while the Setosa forms a well-separated cluster.

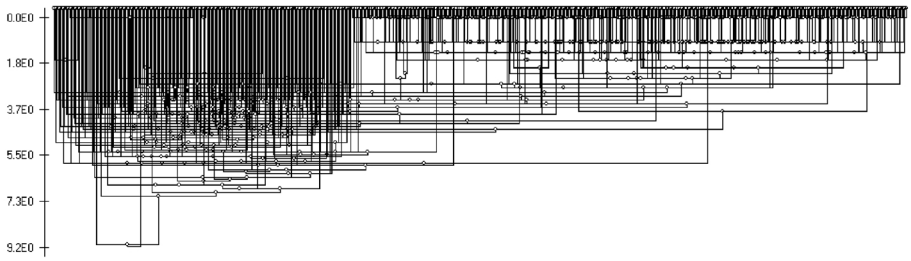


Fig. 16. Dendrogram produced by the single-link method for the breast cancer data.

clusters), instead of the dissimilarity directly, for cluster evaluation.

A parametrical distribution—exponential density—models the statistical properties of this feature within a cluster, forming the basis of the proposed isolation criterion. Continuously updated along the cluster formation process, the distribution mean, $1/\beta$, multiplied by the factor α (by default, 3), constitutes an adaptive, cluster dependent threshold to which dissimilarity increments are compared when two clusters are considered for merge. This is illustrated in Fig. 17 showing the dendrogram produced by the single-link method for the three concentric patterns in Fig. 12e. In this figure, $d1$ and $d2$ correspond to distances between clusters. While clusters present diverse structure and are well separated, cluster isolation based on a global threshold on the distances, as happens with the single-link method, poses difficulties: The inner clusters cannot be isolated without consequent fragmentation of the outer cluster. Group structure is assessed by the proposed approach by means of the statistical model for the dissimilarity increments within the cluster. According to the new isolation criterion, the proposed method, instead of looking at distances, analyzes the gaps $g1$ and $g2$ and compares each with the distribution of dissimilarity increments of the adjoint cluster. As a result, these two clusters are isolated and frozen in the dendrogram. Association steps continue with the remaining data which leads to the formation of a third cluster. Therefore, in this case, application of the proposed clustering method is equivalent to cutting the dendrogram produced by the single-link method at three points with the overall shape of the graph remaining the same.

The dynamic thresholding strategy applied during the cluster formation process may, however, conduct to drastic changes in pattern associations, reflected in distinct topology dendrograms. This is illustrated in the next example,

TABLE 1
Cluster Center Locations Obtained with the k-Means Algorithm

Cluster Center location
2.9843 1.2601 1.3789 1.3363 2.0762 1.2937 2.0717 1.2242 1.0717
7.1857 6.7089 6.6709 5.6414 5.4135 7.7806 6.0295 5.9662 2.6034

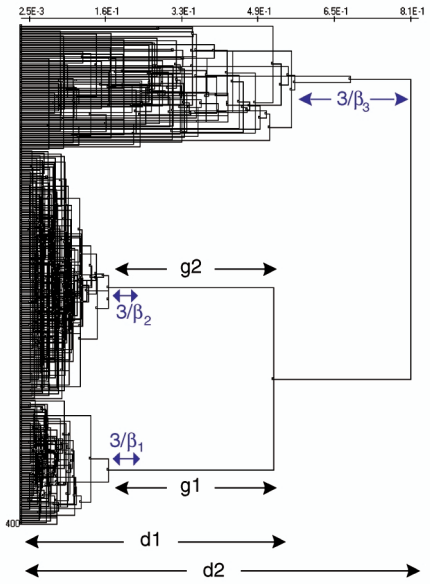


Fig. 17. Dendrogram produced by the single-link method for the data in Fig. 12e. The distances $d1$ and $d2$ are here plotted as levels at which clusters are joined. Single-link method: A vertical line on the dendrogram defines a data partitioning. Proposed method: Dissimilarity increments are compared with a dynamic, class dependent threshold. For instance, the gaps, $g1$ and $g2$, are compared with the corresponding class threshold ($3/\beta_1 = 0.053$ and $3/\beta_2 = 0.035$, respectively). As a result, two clusters are isolated and frozen in the dendrogram; merging steps continue with the remaining data, thus leading to a third cluster, as gaps are smaller than the cluster threshold, $3/\beta_3 = 0.21$.

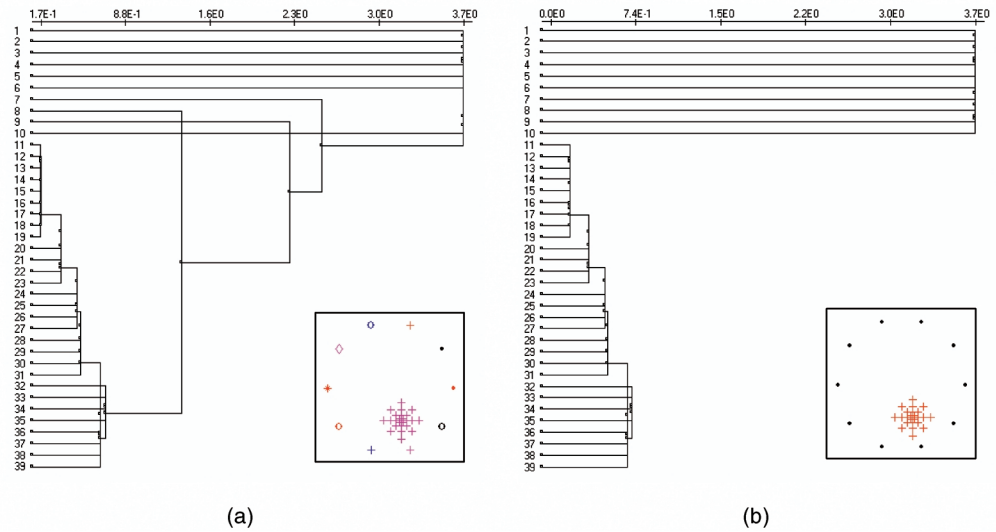


Fig. 18. Dendrograms produced by the single-link and the proposed method. A plot of the clustering obtained is overlayed on the graph. (a) Single-link method. Clustering: cut at level 2. (b) Proposed method: clustering obtained with $\alpha < 10$.

which clarifies how the outer circle in Fig. 13a is detected. For simplicity, only the circle and the star-shaped clusters are considered. Fig. 18a shows the dendrogram produced by the single-link method. Due to spatial proximity, a few points of the circle are associated with the star-shaped cluster. The proposed method (Fig. 18b) changes the way the dendrogram is produced by eliminating the association of the star with its nearest point on the circle (pink cross on the plot in Fig. 18a). Since this association is not possible according to the statistics of dissimilarities in the star pattern, the later is frozen in the dendrogram; further associations continue with the remaining data, making it possible to connect the circle (plot in Fig. 18b).

7 Conclusions

We have shown that clusters of distinct shapes or different data generation paradigms can be adequately modeled by an exponential distribution when analyzing the dissimilarity increments between neighboring patterns; the mean value of this parametric model is closely related to data sparseness, irrespective of orientation or shape features.

Adopting this parametric model for cluster representation, a new criterion for cluster isolation was proposed. A hierarchical agglomerative algorithm adopting the proposed isolation criterion was described and applied to several test data. A comparison of the new method with other approaches covered the following techniques: a hierarchical agglomerative clustering algorithm—the single-link method; a cluster center based technique—the k-means algorithm. The analysis of the method and of the results obtained revealed its ability to identify clusters that have arbitrary shape and size, greatly outperforming the single-link and k-means methods, taken as reference; the number of clusters is intrinsically found without requiring ad hoc specification of design parameters or engaging computationally demanding optimization processes. Furthermore, the algorithm does not impose clusters on the data, as corroborated by the results with random data.

Results with the Iris data set and with Gaussian data with equal covariance matrices and varying Mahalanobis distance, revealed sensitivity of the method to noise. While using a global statistic of dissimilarity increments to characterize cluster structure, the isolation criterion is applied locally to a single pair of patterns: the nearest neighbors linking the clusters candidate for merging. Therefore, the presence of noise may induce the merging of clusters with similar structure. In order to overcome this difficulty, one can apply some denoising technique over the data, eliminating atypical patterns. Otherwise, the isolation criterion can be applied to the average dissimilarity increments, computed over a set of pairs (instead of a single pair) of nearest-neighbor patterns linking the two clusters.

Examples provided in this paper used the Euclidean distance as dissimilarity measure between patterns described as real-valued vectors. The proposed method is not, however, conditioned to any specific dissimilarity measure or pattern representation form. An application example concerning clustering of contour images described in the string format and using a normalized string edit distance [38], [24] as dissimilarity measure has been presented in [33].

The proposed cluster isolation criterion based on the concept of continuity between neighboring patterns within a cluster (the overall structure being captured by dissimilarity increments statistics) has been evaluated in this paper in the context of hierarchical clustering. Ongoing work includes the application of this criterion and its extension to local neighborhoods, to other clustering frameworks, namely, in K-means-based clustering and formal justification for the exponential behavior of dissimilarity increments.

Appendix A

Let $p(x) = \beta \exp^{-\beta x}$ be an exponential distribution with mean value $\bar{x} = \frac{1}{\beta}$. The slope of the distribution at points which are multiples of the mean, $x = i\bar{x} = \frac{i}{\beta}$, is given by

$$\left.\frac{dp(x)}{x}\right|_{x=\frac{i}{\beta}} = -\beta^2 \exp^{-\beta\frac{i}{\beta}} = -\beta^2 \exp^{i}.$$

The equation of the tangential line at this point is of the form

$$y = -\beta^2 \exp^{i} x + y_0.$$

Since $y \equiv p(x = \frac{i}{\beta}) = \beta \exp^{-i}$, from the equality

$$\beta \exp^{-i} = -\beta^2 \exp^{-i} \frac{i}{\beta} + y_0$$

one obtains $y_0 = \beta \exp^{-i}(i+1)$.

The crossing, x_0, of this tangential line with the x-axis is therefore given by

$$\begin{aligned} y = 0 \Rightarrow \quad & 0 = -\beta^2 \exp^{-i} x_0 + \beta \exp^{-i}(i+1) \\ & \beta \exp^{-i}(i+1-\beta x_0) = 0 \\ & x_0 = \frac{i+1}{\beta} = (i+1)\bar{x}. \end{aligned}$$

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. The breast cancer database was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. The database is available at the UCI repository of Machine Learning Databases http://www.ics.ici.edu/~mlearn/MLRepository.html.

REFERENCES

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

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

[3] G. McLachlan and K. Basford, *Mixture Models: Inference and Application to Clustering.* New York: Marcel Dekker, 1988.

[4] S. Roberts, D. Husmeier, I. Rezek, and W. Penny, "Bayesian Approaches to Gaussian Mixture Modeling," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, Nov. 1998.

[5] M. Figueiredo, J. Leitão, and A.K. Jain, "On Fitting Mixture Models," *Energy Minimization Methods in Computer Vision and Pattern Recognition,* E. Hancock and M. Pellilo, eds., pp. 54-69, Springer-Verlag, 1999.

[6] J.D. Banfield and A.E. Raftery, "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics,* vol. 49, pp. 803-821, Sept. 1993.

[7] J. Buhmann and M. Held, "Unsupervised Learning without Overfitting: Empirical Risk Approximation as an Induction Principle for Reliable Clustering," *Proc. Int'l Conf. Advances in Pattern Recognition,* S. Singh, ed., pp. 167-176, 1999.

[8] B. Mirkin, "Concept Learning and Feature Selection Based on Square-Error Clustering," *Machine Learning,* vol. 35, pp. 25-39, 1999.

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

[10] H. Tenmoto, M. Kudo, and M. Shimbo, "MDL-Based Selection of the Number of Components in Mixture Models for Pattern Recognition," *Advances in Pattern Recognition,* A. Amin, D. Dori, P. Pudil, and H. Freeman, eds., vol. 1451, pp. 831-836, 1998.

[11] H. Bischof and A. Leonardis, "Vector Quantization and Minimum Description Length," *Proc. Int'l Conf. Advances on Pattern Recognition,* S. Singh, ed., pp. 355-364, 1999.

[12] N.R. Pal and J.C. Bezdek, "On Cluster Validity for the Fuzzy C-Means Model," *IEEE Trans. Fuzzy Systems,* vol. 3, pp. 370-379, 1995.

[13] P-Y. Yin, "Algorithms for Straight Line Fitting Using k-Means," *Pattern Recognition Letters,* vol. 19, pp. 31-41, 1998.

[14] D. Stanford and A.E. Raftery, "Principal Curve Clustering with Noise," technical report, Univ. of Washington, http://www.stat.washington.edu/raftery, 1997.

[15] H. Frigui and R. Krishnapuram, "A Robust Competitive Clustering Algorithm with Applications in Computer Vision," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 21, no. 5, pp. 450-466, May 1999.

[16] Y. Man and I. Gath, "Detection and Separation of Ring-Shaped Clusters Using Fuzzy Clusters," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 16, no. 8, pp. 855-861, Aug. 1994.

[17] B. Fischer, T. Zoller, and J. Buhmann, "Path Based Pairwise Data Clustering with Application to Texture Segmentation," *Energy Minimization Methods in Computer Vision and Pattern Recognition,* M. Figueiredo, J. Zerubia, and A.K. Jain, eds., vol. 2134, pp. 235-266, 2001.

[18] E.J. Pauwels and G. Frederix, "Finding Regions of Interest for Content-Extraction," *Proc. IS&T/SPIE Conf. Storage and Retrieval for Image and Video Databases VII,* vol. 3656, pp. 501-510, Jan. 1999.

[19] K. Fukunaga, *Introduction to Statistical Pattern Recognition.* New York: Academic Press, 1990.

[20] E. Gokcay and J.C. Principe, "Information Theoretic Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 2, pp. 158-171, Feb. 2002.

[21] C. Zahn, "Graph-Theoretical Methods for Detecting and Describing Gestalt Structures," *IEEE Trans. Computers,* vol. 20, no. 1, pp. 68-86, Jan. 1971.

[22] Y. El-Sonbaty and M.A. Ismail, "On-Line Hierarchical Clustering," *Pattern Recognition Letters,* pp. 1285-1291, 1998.

[23] M. Chavent, "A Monothetic Clustering Method," *Pattern Recognition Letters,* vol. 19, pp. 989-996, 1998.

[24] A.L. Fred and J. Leitão, "A Comparative Study of String Dissimilarity Measures in Structural clustering," *Proc. Int'l Conf. Advances in Pattern Recognition,* S. Singh, ed., pp. 385-394, 1998.

[25] S. Guha, R. Rastogi, and K. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases," *Proc. 1998 ACM-SIGMOID Int'l Conf. Management of Data,* 1998.

[26] R. Dubes and A.K. Jain, "Validity Studies in Clustering Methodologies," *Pattern Recognition,* vol. 11, pp. 235-254, 1979.

[27] T.A. Bailey and R. Dubes, "Cluster Validity Profiles," *Pattern Recognition,* vol. 15, no. 2, pp. 61-83, 1982.

[28] E.W. Tyree and J.A. Long, "The Use of Linke Line Segments for Cluster Representation and Data Reduction," *Pattern Recognition Letters,* vol. 20, pp. 21-29, 1999.

[29] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, pp. 790-799, 1995.

[30] D. Comaniciu and P. Meer, "Distribution Free Decomposition of Multivariate Data," *Pattern Analysis and Applications,* vol. 2, pp. 22-30, 1999.

[31] G. Karypis, E.-H. Han, and V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling," *IEEE Computer,* vol. 32, no. 8, pp. 68-75, Aug. 1999.

[32] P. Bajcsy and N. Ahuja, "Location- and Density-Based Hierarchical Clustering Using Similarity Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 9, pp. 1011-1015, Sept. 1998.

[33] A.L. Fred and J. Leitao, "Clustering under a Hypothesis of Smooth Dissimilarity Increments," *Proc. 15th Int'l Conf. Pattern Recognition,* vol. 2, pp. 190-194, 2000.

[34] C.J. Merz and P.M. Murphy, "UCI Repository of Machine Learning Databases," Dept. of Information and Computer Science, Univ. of California, Irvine, http://www.ics.uci.edu/mlearn/MLRepository.html, 1996.

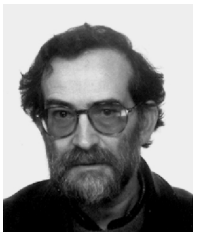
[35] E.S. Ristad and P.N. Yianilos, "Learning String-Edit Distance," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 20, no. 5, pp. 522-531, May 1998.

[36] D. Sankoff and J. Kruskal, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison,* reprint, with a forward by J. Nerbonne, Stanford, Calif.: CLSI Publications, (1983), 1999.

[37] B.J. Oomen and R.S.K. Loke, "Pattern Recognition of Strings Containing Traditional and Generalized Transposition Errors," *Proc. Int'l Conf. Systems, Man, and Cybernetics,* pp. 1154-1159, 1995.

[38] A. Marzal and E. Vidal, "Computation of Normalized Edit Distance and Applications," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 9, pp. 926-932, Sept. 1993.
[39] A. Fred, "Clustering Based on Dissimilarity First Derivatives," *Proc. Second Int'l Workshop Pattern Recognition in Information Systems,* J. Iñesta and L. Micó, eds., pp. 257-266, 2002.
[40] R. Kothari and D. Pitts, "On Finding the Number of Clusters," *Pattern Recognition Letters,* vol. 20, pp. 405-416, 1999.
[41] S.V. Chakravarthy and J. Ghosh, "Scale-Based Clustering Using the Radial Basis Function Network," *IEEE Trans. Neural Networks,* vol. 7, pp. 1250-1261, 1996.

Ana L.N. Fred received the MS and PhD degrees in electrical and computer engineering, in 1989 and 1994, respectively, both from the Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal. She has been a faculty member of IST since 1986, where she is currently a professor with the Department of Electrical and Computer Engineering. She is a researcher at the Communication Theory and Pattern Recognition Group of the Institute of Telecommunications. Her research interests include information theory, pattern recognition, signal processing, and artificial intelligence. She is a member of the IEEE.



José M.N. Leitão (M'95) received the EE and PhD degrees in electrical engineering, in 1970 and 1983, respectively, both from the Instituto Superior Técnico (IST), Technical University of Lisbon, Portugal. He received the "Agregado" degree in electrical and computer engineering, also from IST, in 1992. He was with the Laboratory of Physiology of the Instituto Gulbenkian de Ciência, in Oeiras, Portugal, from 1970 to 1972. After spending three years at the University of Tübingen, Germany, he joined the faculty of IST in 1976, where he is currently a full professor with the Department of Electrical and Computer Engineering. He is also the coordinator of the Communication Theory and Pattern Recognition Group of the Institute of Telecommunications. His main research interests are communication and information theory, pattern recognition, signal, and image processing. He is a member of the IEEE.