Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 40 (2007) 807–824
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Model-based evaluation of clustering validation measures

Marcel Brun[a], Chao Sima[b], Jianping Hua[a], James Lowey[a], Brent Carroll[c], Edward Suh[a], Edward R. Dougherty[a,b,d,*]

[a]*Translational Genomics Research Institute, Phoenix, Arizona, USA*
[b]*Department of Electrical Engineering, Texas A&M University, College Station, TX, USA*
[c]*Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA*
[d]*Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, TX, USA*

Received 1 November 2005; received in revised form 27 May 2006; accepted 29 June 2006

Abstract

A cluster operator takes a set of data points and partitions the points into clusters (subsets). As with any scientific model, the scientific content of a cluster operator lies in its ability to predict results. This ability is measured by its error rate relative to cluster formation. To estimate the error of a cluster operator, a sample of point sets is generated, the algorithm is applied to each point set and the clusters evaluated relative to the known partition according to the distributions, and then the errors are averaged over the point sets composing the sample. Many validity measures have been proposed for evaluating clustering results based on a single realization of the random-point-set process. In this paper we consider a number of proposed validity measures and we examine how well they correlate with error rates across a number of clustering algorithms and random-point-set models. Validity measures fall broadly into three classes: internal validation is based on calculating properties of the resulting clusters; relative validation is based on comparisons of partitions generated by the same algorithm with different parameters or different subsets of the data; and external validation compares the partition generated by the clustering algorithm and a given partition of the data. To quantify the degree of similarity between the validation indices and the clustering errors, we use Kendall's rank correlation between their values. Our results indicate that, overall, the performance of validity indices is highly variable. For complex models or when a clustering algorithm yields complex clusters, both the internal and relative indices fail to predict the error of the algorithm. Some external indices appear to perform well, whereas others do not. We conclude that one should not put much faith in a validity score unless there is evidence, either in terms of sufficient data for model estimation or prior model knowledge, that a validity measure is well-correlated to the error rate of the clustering algorithm.

Keywords: Clustering algorithms; Clustering errors; Validation indices

1. Introduction

Data clustering has been used for decades in image processing and pattern recognition [1,2], and in recent years has become a popular technique in genomic studies using gene-expression microarrays [3–6]. Time-series clustering groups together genes whose expression levels exhibit similar behavior through time. Similarity is taken to indicate possible co-regulation. Another way to use expression data is to take expression profiles over various tissue samples, and then cluster these samples based on the expression levels for each sample. This approach offers the potential to discriminate pathologies based on their differential patterns of gene expression.

Despite the popularity of clustering, until very recently scant attention has been paid to what exactly is meant by the output of a clustering algorithm. A cluster operator takes a set of data points and partitions the points into clusters (subsets). But what is the meaning of the result? Is there more

* Corresponding author. Department of Electrical Engineering, 3128 Texas A&M University, College Station, TX 77840, USA. Tel.: +1 979 845 8896; fax: +1 979 845 7441.
E-mail address: e-dougherty@tamu.edu (E.R. Dougherty).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.06.026

than simply a picture? Is there any scientific content? Can it be argued that one clustering procedure is better than another? All of these questions point to the epistemological basis of clustering [7]. Unless clustering leads to predictions that can be tested with physical data, it lacks scientific content because, as Richard Feynman states, "It is whether or not the theory gives predictions that agree with experiment. It is not a question of whether a theory is philosophically delightful, or easy to understand, or perfectly reasonable from the point of view of common sense" [8]. Lacking inference in the context of a probability model, clustering is essentially a subjective visualization tool. Jain et al. have written, "Clustering is a subjective process; the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the process of clustering difficult" [1]. Subjective appreciations are certainly useful in the formulation of hypotheses, but these are constitutive of scientific knowledge only if they are set in a predictive framework.

The key to a predictive probabilistic theory of clustering is to recognize that, whereas the theory of classification is based on operators on random variables, the theory of clustering must be based on operators on random sets. The predictive capability of a clustering algorithm must be measured by the decisions it yields regarding the partitioning of random point sets. Once this is recognized, the path to the development of a predictive theory of clustering that can constitute scientific knowledge is clear and such a theory has been developed [9]. In particular, the error of a clustering algorithm is rigorously grounded within the random-set-based theory.

Historically, a host of "validity" measures have been proposed for evaluating clustering results based on a single realization of the random-point-set process [10–15]. No doubt one would like to measure the accuracy of a cluster operator based on a single application. But is this feasible? Clearly, it would be absurd to claim that one can assess the validity of a classifier based on the classification of a single point without knowledge of the true label of the point. Indeed, how would one hope to assess classifier validity given its actions on many points without access to their labels? Assessing the validity of a cluster operator on a single point set without knowledge of the true partition is analogous to assessing the validity of a classifier with a single unlabeled point. But there is a difference that provides hope. The output of a cluster operator consists of a partition of a point set. Therefore there is spatial structure to the output and one can define measures for different aspects of this structure, for instance, compactness. One can also consider the effects of a cluster operator on subsets of the data. It could be hoped that such measures can be used to assess the *scientific validity* of a clustering algorithm. For a validity measure to assess scientific validity, ipso facto, it must be closely related to the error rate of the cluster operator as that rate is defined within a probabilistic theory of clustering. In this paper we examine a number of proposed validity measures and see how well they correlate with error rates across a number of clustering algorithms and random-point-set models.

Validity measures proposed for clustering algorithms fall broadly into three classes. The first type is based on calculating properties of the resulting clusters, such as compactness, separation and roundness. This approach is called *internal validation* because it does not require additional information about the data [13,14,16]. A second approach is based on comparisons of partitions generated by the same algorithm with different parameters, or different subsets of the data. This is called *relative validation*, and also does not include additional information [13,4,17]. In the third way, called *external validation* and also based on comparison of partitions, the partitions to be compared consist of the one generated by the clustering algorithm and a given partition of the data (or a subset of the data) [14,18]. External validation corresponds to a kind of error measurement, either directly or indirectly. Therefore we should expect external methods to be better correlated to the true error; however, this is not always the case because it depends on the external validation procedure as well as the random labeled point process to which it is being applied and the specific clustering algorithm being tested. Fig. 1 shows a hierarchy of validation techniques.

On the issue of models, we have chosen several for this study. No doubt one could choose others. We have tried to choose models that would illustrate geometries that are both favorable and unfavorable to the various validity measures, thereby helping to provide conditions under which one might consider applying a particular validity index. If a validity index has been defined with the idea of measuring some property of the resulting clusters, then it might be expected to perform well when the random labeled point process generates sets possessing the property. But what happens when the process does not generate point sets possessing the property, or points sets having some degree of relation to the property? Does the validity measure still provide useful information or does it collapse completely and provide totally unreliable results? Obviously, every proposed validity index has a rationale behind it. But here we return to the epistemological question: Under what conditions is the rationale sound? This question can only be answered by experimentally examining the performance of a validity index under varied conditions: different clustering algorithms and different models.

The paper is organized in the following manner. Section 2 defines the error measure for cluster operators. Sections 3–5 define the internal, relative and external validation indices that we consider. Section 6 describes the clustering algorithms used in the study. Section 7 describes the model-based analysis employed. Section 8 describes the experiments. Section 9 analyzes the results relative to the different validation indices. Some concluding remarks are provided in Section 10. Owing to the size of the study, a substantial portion of the results are provided on a companion website at http://ee.tamu.edu/∼edward/validation/.

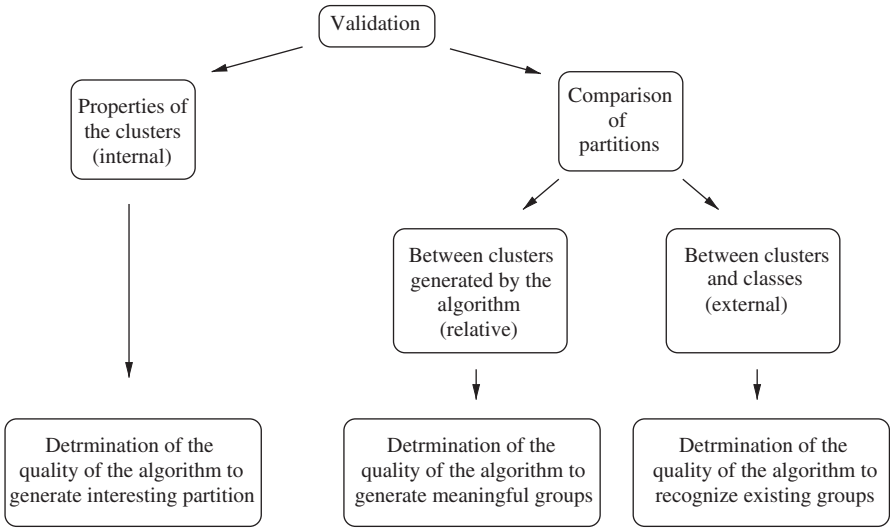

Fig. 1. A simplified classification of validation techniques.

2. Error measure

Although we will not cover the mathematical theory of Ref. [9], we believe it is necessary to summarize some points so that clustering error and error estimation are clear. As noted previously, in a probabilistic framework a clustering algorithm is an operator on random point sets. The points to be clustered are assumed to belong to a realization S of a random labeled point process Ξ and a clustering algorithm is a mapping Φ that assigns to S a label function, the latter being of the form $\varphi(\mathbf{x}) \in \{0, 1, 2, \ldots, K-1\}$ for all $\mathbf{x} \in S$, where K is the number of clusters forming a partition of S. This means that $\mathbf{x}_1$ and $\mathbf{x}_2$ are in the same cluster if and only if $\varphi(\mathbf{x}_1) = \varphi(\mathbf{x}_2)$. The error of a clustering algorithm is the expected difference between its labels and the labels generated by the labeled point process Ξ.

To quantify the matter, let S^{Φ} denote the labeling of S created by the clustering algorithm Φ, and let S^{Ξ} denote the labeling of the point process Ξ. Let $I_{\Phi}(S; \mathbf{x})$ and $I_{\Xi}(S; \mathbf{x})$ denote the label of $\mathbf{x}$ for S^{Φ} and S^{Ξ}, respectively. Then the *label error* between the two labelings is defined as the proportion of points that are differently labeled:

$$\varepsilon(S^{\Xi}, S^{\Phi}) = \frac{|\{\mathbf{x} : I_{\Xi}(S; \mathbf{x}) \neq I_{\Phi}(S; \mathbf{x})\}|}{|S|}, \tag{1}$$

where $|\bullet|$ indicates the number of elements of a set. Since the disagreement between two partitions should not depend on the indices used to label their clusters, the *partition error* is defined by

$$\varepsilon^*(S^{\Xi}, S^{\Phi}) = \min_{\pi} (S^{\Xi}, \pi S^{\Phi}), \tag{2}$$

where the minimum is taken over all of the possible permutations, πS^{Φ}, of the K sets in S^{Φ}. Since this error is for a specific realization S of the process Ξ, the error of the clustering algorithm Φ with respect to Ξ is given by the expected value

$$\varepsilon_{\Xi}(\Phi) = E[\varepsilon^*(S^{\Xi}, S^{\Phi})], \tag{3}$$

where the expectation is taken relative to the distribution of the random set Ξ (and here we defer to Ref. [9] for the theoretical details).

Error estimation is done in the usual manner: the expectation $E[\varepsilon^*(S^{\Xi}, S^{\Phi})]$ is estimated by generating realizations S of Ξ, computing $\varepsilon^*(S^{\Xi}, S^{\Phi})$ for each realization, and then averaging. In practice, we can generate independent synthetic data to test the performance of a cluster operator in the following manner: generate a sample of point sets $S_1, S_2, \ldots, S_m$ according to Ξ (so that $S_1^{\Xi}, S_2^{\Xi}, \ldots, S_m^{\Xi}$ are known), apply the clustering algorithm to $S_1, S_2, \ldots, S_m$ to obtain $S_1^{\Phi}, S_2^{\Phi}, \ldots, S_m^{\Phi}$, compute $\varepsilon^*(S_j^{\Xi}, S_j^{\Phi})$ for $j = 1, 2, \ldots, m$, and then average $\varepsilon^*(S_j^{\Xi}, S_j^{\Phi})$ for $j = 1, 2, \ldots, m$ to obtain an estimate of $\varepsilon_{\Xi}(\Phi)$ [18].

To illustrate error estimation we consider two simple two-dimensional labeled point processes. The first one consists of a mixture of two Gaussian distributions, so that points are labeled 0 or 1, depending on whether they are generated by the Gaussian with mean $(0, 3)$ and covariance matrix $2\mathbf{I}$, or by the Gaussian with mean $(3, 0)$ and covariance matrix $2\mathbf{I}$, with 50 points per class being generated. The second process consists of a mixture of a Gaussian with mean $(0, 0)$ and covariance matrix $0.2\mathbf{I}$, and a circular distribution with radius normally distributed according to $N(3, 0.2)$ and angle normally distributed in radians according to $N(0, 1)$, again 50 points being generated per class. Fig. 2 shows the results of single realization of the second process, part (a) showing the point set generated by the process and the remaining parts showing the results for five clustering algorithms. Table 1 shows the estimated error rates for the five clustering algorithms for the two random labeled point processes (based on 100 realizations).

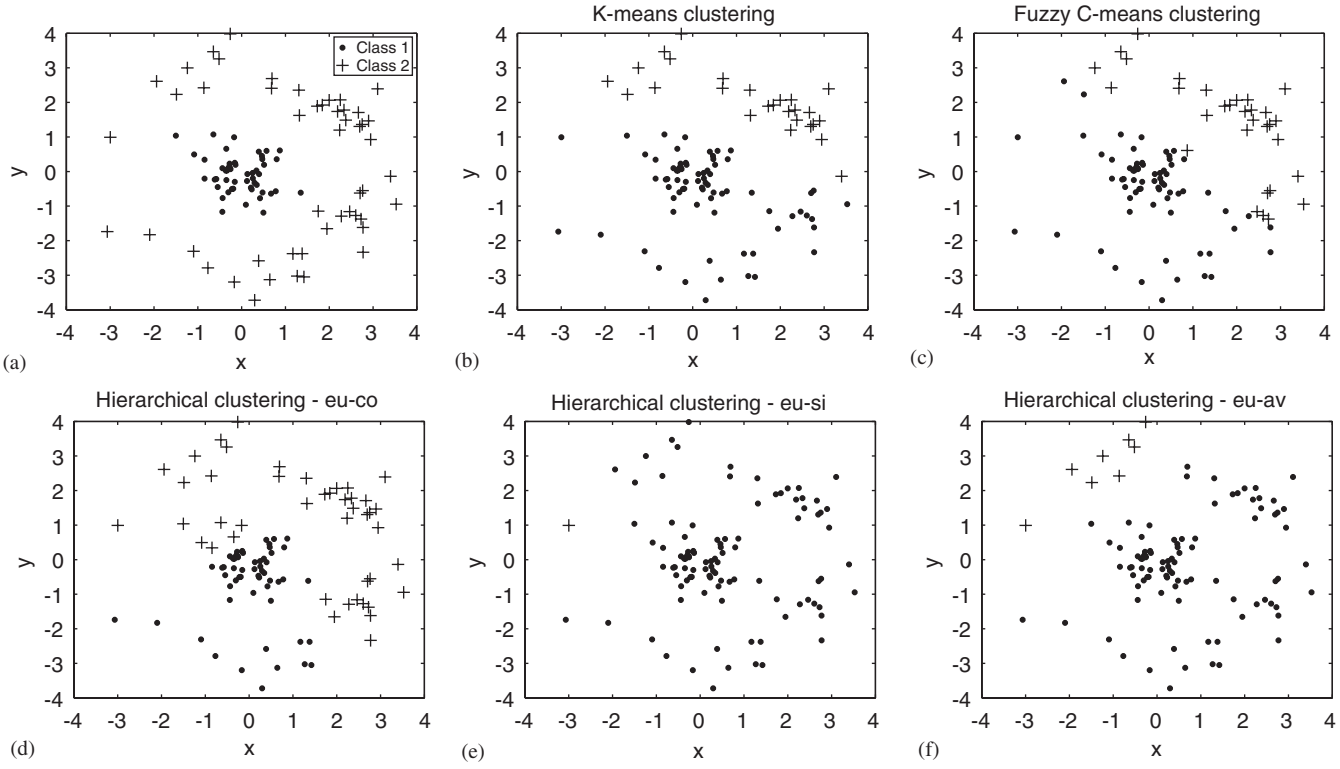

Fig. 2. (a) Labeled random set from second process; (b) K-means clustering (24 errors); (c) fuzzy C-means clustering (21 errors); (d) hierarchical (eu-co) clustering (18 errors); (e) hierarchical (eu-si) clustering (49 errors); (f) hierarchical (eu-av) clustering (42 errors). *Abbreviations in hierarchical clustering*: eu — Euclidean distance; si — Single linkage; co — Complete linkage; av — Average linkage.

Table 1
Estimated misclassification rate (%), over 100 realizations

	K-means	Fuzzy	Hierarchical		
		C-means	eu-co	eu-si	eu-av
Set 1	7.03	6.92	13.24	48.81	18.76
Set 2	18.9	13.97	26.59	46.14	40.75

3. Internal validation indices

For internal validation, the evaluation of the resulting clusters is based on the clusters themselves, without additional information or repeats of the clustering process. This family of techniques is based on the assumption that the algorithms should search for clusters whose members are close to each other and far from members of other clusters. We describe the internal validation indices used in the paper.

3.1. Dunn's indices

The *Dunn's validation index* is defined as the ratio between the minimum distance between two clusters and the size of the largest cluster [19–21]. If $\mathscr{C} = \{C_1, \ldots, C_K\}$ is a partition of the n points into K clusters, then the index is defined by

$$V(\mathscr{C}) = \frac{\min_{h,k=1,\ldots,K, h\neq k} d_C(C_k, C_h)}{\max_{k=1,\ldots,K} \Delta(C_k)}, \tag{4}$$

where $d_C(C_k, C_h)$ is the distance between the two clusters and $\Delta(C_k)$ is the size of the cluster C_k. The value of $V(\mathscr{C})$ depends on the selection of the distance measures. Several measures for the distances between clusters (or *linkage*) are proposed in Ref. [21]: *single*, *complete*, *average*, *average to centroid* and *Haussdorff metrics*. Table 2 shows the definition for each of these distance measures. The size of the cluster may be defined in many ways. Some of the measures defined in Ref. [21] are *complete*, *average* and *centroid*. Table 3 shows the definition for each of these measures. Each combination of distance measure and cluster-size measure defines a different Dunn's index.

3.2. Silhouette index

The *silhouette* is the average, over all clusters, of the *silhouette width* of their points [12,20,21]. If $\mathbf{x}$ is a point in the cluster C_k and n_k is the number of points in C_k, then the *silhouette width* of $\mathbf{x}$ is defined by the ratio

$$S(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{\max[b(\mathbf{x}), a(\mathbf{x})]}, \tag{5}$$

Table 2
Linkage methods for the distance between two clusters

Linkage	Equation	Alias
Single	$d_C(C_i, C_j) = \min_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} d(\mathbf{x}, \mathbf{y})$	min
Complete	$d_C(C_i, C_j) = \max_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} d(\mathbf{x}, \mathbf{y})$	max
Average[a]	$d_C(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\mathbf{x}\in C_i, \mathbf{y}\in C_j} d(\mathbf{x}, \mathbf{y})$	mean
Centroid[b]	$d_C(C_i, C_j) = d(\bar{\mathbf{x}}, \bar{\mathbf{y}})$	cen
Average to Centroid[b]	$d_C(C_i, C_j) = \frac{1}{n_i + n_j}\left[\sum_{\mathbf{x}\in C_i} d(\mathbf{x}, \bar{\mathbf{y}}) + \sum_{\mathbf{y}\in C_j} d(\mathbf{y}, \bar{\mathbf{x}})\right]$	cmean
Hausdorff metrics[c]	$d_C(C_i, C_j) = \max[d_H(C_i, C_j), d_H(C_j, C_i)]$	hausf

[a] n_i and n_j are the number of samples in clusters C_i and C_j, respectively.
[b] $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the centroid of clusters C_i and C_j, respectively.
[c] $d_H(A, B) = \max_{\mathbf{x}\in A} \min_{\mathbf{y}\in B} d(\mathbf{x}, \mathbf{y})$.

Table 3
Measures of cluster size

Measure	Equation	Alias
Complete	$\Delta(C) = \max_{\mathbf{x},\mathbf{y}\in C} d(\mathbf{x}, \mathbf{y})$	max
Average[a]	$\frac{1}{n*(n-1)} \sum_{\mathbf{x},\mathbf{y}\in C} d(\mathbf{x}, \mathbf{y})$	mean
Centroid[b]	$\Delta(C) = \frac{2}{\lvert C\rvert} \sum_{\mathbf{x}\in C} d(\mathbf{x}, \bar{\mathbf{x}})$	cen

[a] n is the number of samples in clusters C.
[b] $\bar{\mathbf{x}}$ is the centroid of clusters C.

where $a(\mathbf{x})$ is the average distance between $\mathbf{x}$ and all other points in C_k,

$$a(\mathbf{x}) = \frac{1}{n_k - 1} \sum_{\mathbf{y}\in C_k, \mathbf{y}\neq\mathbf{x}} d(\mathbf{x}, \mathbf{y}) \tag{6}$$

and $b(\mathbf{x})$ is the minimum of the average distances between $\mathbf{x}$ and the points in the other clusters,

$$b(\mathbf{x}) = \min_{h=1,\ldots,K, h\neq k} \left[\frac{1}{n_h} \sum_{\mathbf{y}\in C_h} d(\mathbf{x}, \mathbf{y})\right]. \tag{7}$$

Finally, the *global silhouette* index is defined by

$$S = \frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{n_k} \sum_{\mathbf{x}\in C_k} S(\mathbf{x})\right]. \tag{8}$$

For a given point $\mathbf{x}$, its silhouette width ranges from -1 to 1. If the value is close to -1, then it means that the point is closer, on average, to another cluster than the one to which it belongs. If the value is close to 1, then it means that its average distance to its own cluster is significantly smaller than to any other cluster. The higher the silhouette, the more compact and separated are the clusters.

3.3. Hubert's correlation with distance matrix

Let $\mathscr{C} = \{C_1, \ldots, C_K\}$ be a partition of the set of n objects into K groups, and let P be a similarity matrix between the n objects such that $P(i, j)$ is a measure of similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$. The relationship between two vectors, whether they belong to the same cluster or not, can be represented by a similarity matrix D defined by $D(i, j) = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, and $D(i, j) = 0$ if they belong to different clusters. The correlation Γ_D between both matrices gives a measure of similarity between them:

$$\Gamma_D = \frac{1}{M} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} D(i, j) P(i, j), \tag{9}$$

with $M = n(n-1)/2$, the number of pairs of different points.

The index Γ_D is classified as an internal index because it is based only on the partition $\mathscr{C}$ defined by the clustering algorithm and the similarity between the points to be grouped.

4. Relative validation indices

Relative validation is based on the measurement of the consistency of the algorithms, comparing the clusters obtained by the same algorithm under different conditions.

4.1. Figure of merit

The *figure of merit* (*FOM*) [17] is based on the assumption that, when used on microarray data, the clusters represent different biological groups, and therefore, points (genes) in the same cluster will possess similar pattern vectors (expression profiles) for additional features (arrays). Let m be the number of features, n the number of points and K the number of clusters. Let $\mathscr{C}^j = \{C_1^j, \ldots, C_K^j\}$ be the partition

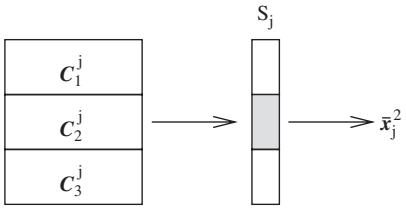

Fig. 3. Example of computation of $\bar{\mathbf{x}}_j^k$.

obtained by the algorithm when removing the feature S_j. The figure of merit for the feature S_j is computed as

$$FOM(K, j) = \sqrt{\frac{1}{n}\sum_{k=1}^{K}\sum_{i\in C_k^j}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j^k)^2}, \tag{10}$$

where $\bar{\mathbf{x}}_j^k$ is the jth element of the average of the vectors in C_k^j (Fig. 3). The figure of merit for a clustering algorithm, specifying K clusters, is computed as the following sum:

$$FOM(K) = \sum_{j=1}^{m} FOM(K, j). \tag{11}$$

If the partition defines compact sets in the removed feature, then their average distances to their centroids should be small. The FOM is the average measure of the compactness of these sets. The heuristic behind the figure of merit is that the lower the FOM, the better the clusters are to predict the removed feature and, therefore, the more consistent the result of the clustering algorithm.

A drawback of the FOM as defined is that its decrease as a function of the number of clusters may be artificial, due to the fact that more clusters means a smaller average size for the clusters. A solution to this problem is to adjust the values using a model-based correction factor, $\sqrt{(n-K)/n}$. The result is called *adjusted figure of merit*,

$$FOM^c(K) = \frac{1}{\sqrt{(n-K)/n}} \cdot FOM(K). \tag{12}$$

4.2. Stability

The stability measure has been introduced to assess the validity of the partitioning found by clustering algorithms and to select the number of clusters [22,23]. The stability measures the ability of a clustered data set to predict the clustering of another data set sampled from the same source. Let us assume that there exists a partition of a set S of n objects into K groups, $\mathscr{C} = \{C_1, \ldots, C_K\}$, and a partition of another set S' of n' objects into K' groups, $\mathscr{C}' = \{C'_1, \ldots, C'_{K'}\}$. Let the labelings α and α' be defined by $\alpha(\mathbf{x}) = i$ if $\mathbf{x} \in C_i$, for $\mathbf{x} \in S$, and $\alpha'(\mathbf{x}) = i$ if $\mathbf{x} \in C'_i$, for $\mathbf{x} \in S'$, respectively. The labeled set (S, α) can be used to train a classifier $f : \Re^n \to L$, which induces a labeling $\bar{\alpha}$ on S' by $\bar{\alpha}(\mathbf{x}) = f(\mathbf{x})$. The consistency of the pairs (S, α) and (S', α') is measured by the similarity between the original labeling α' and the induced labeling $\bar{\alpha}$ in S':

$$d_S(\mathscr{C}, \mathscr{C}') = \min_{\pi} d_\alpha(\alpha', \pi(\bar{\alpha})) \tag{13}$$

over all possible permutations π of the K' labels for $\mathscr{C}'$, with

$$d_\alpha(\alpha^1, \alpha^2) = \frac{1}{n'}\sum_{\mathbf{x}\in S'} \delta(\alpha^1(\mathbf{x}), \alpha^2(\mathbf{x})) \tag{14}$$

with $\delta(u, v) = 0$ if $u = v$ and $\delta(u, v) = 1$ if $u \neq v$.

The stability for a clustering algorithm is defined by the expectation E of the stability for pairs of sets drawn from the same source:

$$\xi = E_{(S,\mathscr{C})(S',\mathscr{C}')}[d(\mathscr{C}, \mathscr{C}')]. \tag{15}$$

In practice, there is only one set S of points with which to estimate the stability of a clustering algorithm. Estimation of the stability is obtained via a resampling schema [22]: the set S is partitioned into two disjoint subsets S_1 and S_2, the clustering algorithm is applied to obtain two partitions, $\mathscr{C}_1$ and $\mathscr{C}_2$, $d(\mathscr{C}_1, \mathscr{C}_2)$ is computed, and the process is repeated and the values averaged to obtain an estimate of ξ.

The stability index is dependent on the number of clusters, and therefore needs to be normalized when used for model selection [22,23]. The normalization is obtained by dividing it by the stability obtained when using a random estimator as classifier. The selection of the classification rule can influence the ability of this index to evaluate the quality of the clustering algorithm, since if the rule is too simple as to partition the space in the same fashion that the clustering algorithm does, then it may introduce false instability and downgrade the algorithm [23].

5. External validation indices

In *external validation*, the quality of the algorithm is evaluated by comparing the resulting clusters with pre-specified information.

5.1. Hubert's correlation

Assume that there exist two partitions of the same set of n objects into K groups: $\mathscr{C}^A = \{C_1^A, \ldots, C_K^A\}$, defined by additional information about the problem (called the *true* partition), and $\mathscr{C}^B = \{C_1^B, \ldots, C_K^B\}$, obtained by application of a clustering algorithm (called the *clustering* partition). The sets C_k^A are called *classes* and the sets C_k^B are called *clusters*. For each partition $\mathscr{C}$ the relationship between two vectors, whether they belong to the same cluster or not, can be represented by a similarity matrix defined by $\mathbf{d}(i, j) = 1$ if $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, and $\mathbf{d}(i, j) = 0$ if they belong to different clusters.

If $\mathbf{d}^A$ and $\mathbf{d}^B$ are the similarity matrices induced by two partitions, $\mathscr{C}^A$ and $\mathscr{C}^B$, then two similarity indices are com-

Table 4
Indices of agreement between partitions

Index	Equation
Rand statistic	$R = \frac{a+d}{M_a}$
Jaccard coefficient	$J = \frac{a}{a+b+c}$
Folkes and Mallow index	$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}$

puted as functions of the correlations and the covariances of these matrices, the Hubert Γ statistic:

$$\Gamma = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{d}^A(i,j)\mathbf{d}^B(i,j) \tag{16}$$

and the normalized Γ^* statistic:

$$\Gamma^* = \frac{1}{M\sigma^A\sigma^B}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (\mathbf{d}^A(i,j)-\mu_A) \times (\mathbf{d}^B(i,j)-\mu_B), \tag{17}$$

where $M = n(n-1)/2$ is the number of pairs of different points, and μ^A, μ^B, σ^A and σ^B are the respective sample means and standard deviations of the values in the matrices $\mathbf{d}^A$ and $\mathbf{d}^B$. The Hubert statistic is based on the fact that the more similar the partitions, the more similar the matrices would be, and this similarity can be measured by their correlation.

5.2. Rand statistics, Jaccard coefficient and Folkes and Mallows index

Given the true partition $\mathscr{C}^A = \{C_1^A, \ldots, C_K^A\}$ and the clustering partition $\mathscr{C}^B = \{C_1^B, \ldots, C_K^B\}$, for each pair of samples $\mathbf{x}$, $\mathbf{y}$ ($\mathbf{x} \neq \mathbf{y}$), there are four possible situations:

(a) $\mathbf{x}$ and $\mathbf{y}$ fall in the same cluster in both $\mathscr{C}^A$ and $\mathscr{C}^B$,
(b) $\mathbf{x}$ and $\mathbf{y}$ fall in the same cluster in $\mathscr{C}^A$ but in different clusters in $\mathscr{C}^B$,
(c) $\mathbf{x}$ and $\mathbf{y}$ fall in the different clusters in $\mathscr{C}^A$ but in the same cluster in $\mathscr{C}^B$,
(d) $\mathbf{x}$ and $\mathbf{y}$ fall in different clusters in both $\mathscr{C}^A$ and $\mathscr{C}^B$.

The measure of disagreement between $\mathscr{C}^A$ and $\mathscr{C}^B$ is quantified by the number of pairs of vectors that fall in situations (b) and (c). Let a, b, c, and d be the numbers of pairs of different vectors that belong to situations (a), (b), (c) and (d), respectively, and let $M = n(n-1)/2$ be the number of pairs of different vectors. The indices in Table 4 measure the agreement between the two partitions [13]: the *Rand statistic*, *Jaccard coefficient* and *Folkes and Mallow index*. The Rand statistic measures the proportion of pairs of vectors that agree by belonging either to the same cluster (a) or to different clusters (d) in both partitions. The Jaccard coefficient measures the proportion of pairs that belong to the same cluster (a) in both partitions, relative to all pairs that belong to the same cluster in at least one of the two partitions ($a+b+c$). The Folkes and Mallow (*FM*) index measures the geometric mean of the proportion of pairs that belong to the same cluster in both partitions (a), relative to the pairs that belong to the same cluster for each partition ($a+b$ for $\mathscr{C}^A$ and $a+c$ for $\mathscr{C}^B$).

6. Clustering algorithms

To simulate realistic conditions for the performance of the validation indices, they are applied to the outcomes of several clustering algorithms. We have selected five different algorithms. Variations in parameters raise the amount to a total of 12 different methods. The clustering algorithms used are:

- K-*means*: One of the most common iterative algorithms is the K-means algorithm [1,16], broadly used because of its simplicity of implementation, its convergence speed and the good quality of the clusters (for a limited family of problems).
- *Fuzzy* C-*means*: In the K-means algorithm, each vector is classified as belonging to a unique cluster (hard cluster), and the centroids are updated based on the classified samples. In a variation of this approach, known as fuzzy C-means [1,16], all vectors have a degree of membership of belonging to each cluster, and the respective centroids are calculated based on these membership degrees.
- *SOM*: By applying a self-organizing map to the data, clusters can be defined by the points of a grid that is adjusted to the data [24–27]. Usually the algorithm uses a two-dimensional grid in the higher-dimensional space, but for clustering it is usual to use a one-dimensional grid. For this paper we implement the SOM algorithm with Euclidean distance and two types of neighbors: bubble and Gaussian.
- *Hierarchical clustering*: Hierarchical clustering [1] creates a hierarchical tree of similarities between the vectors, called a dendrogram. The most common implementation of this strategy is agglomerative hierarchical clustering, which starts with a family of clusters with one vector each, and merges the clusters iteratively based on some distance measure until there is only one cluster left, containing all the vectors. For this paper we consider two distance metrics: Euclidean distance and correlation, and three linkage methods:
  - *Single linkage*. When two clusters are joined into a new cluster C_i, the distance between C_i and an existing cluster C_j is the minimum distance between the elements of C_i and C_j.
  - *Complete linkage*. When two clusters are joined into a new cluster C_i, the distance between C_i and an existing cluster C_j is the maximum distance between the elements of C_i and C_j.

Table 5
Clustering algorithms

Code	Algorithm	Parameters
km	K-means	
fcm	Fuzzy C-means	$b = 2$[a,b]
so[eu,b]	SOM	Distance = Euclidean, Neighborhood = bubble[b,c]
hi[eu,co]	Hierarchical	Distance = Euclidean, Linkage = Complete
hi[c,co]	Hierarchical	Distance = 1-abs(Pearson Corr), Linkage = Complete
hi[eu,si]	Hierarchical	Distance = Euclidean, Linkage = Single
hi[c,si]	Hierarchical	Distance = 1-abs(Pearson Corr), Linkage = Single
em[diag]	EM	Mixing Model = Diagonal[a,b]

[a]Tolerance = 0.001.
[b]Maximum number of iterations = 10000.
[c]Starting $\alpha = 0.9$, Stopping $\alpha = 0.01$.

- *Average linkage*. When two clusters are joined into a new group C_i, the distance between C_i and an existing cluster C_j is the average distance between the elements of C_i and C_j.

- *Expectation maximization*: Expectation maximization (EM) clustering [28–30] is based on the estimation of the density for the classes using the EM algorithm. The estimation is done in a two-step process similar to K-means clustering. In the first step the probabilities are estimated conditioned to the actual parameters, assigning each vector to one cluster (model), while in the second step the parameters of the models are estimated within the new clusters. The process is iterated until there is no more significant change in the parameters. The result is an estimated set of K multivariate distributions, each one defining a cluster, and each vector assigned to the cluster with maximum conditional probability. Different assumptions on the model result in different constraints on the covariance matrices. For this paper we use two constraints for the covariance matrix Σ_k of the class k:
  - *Pooled diagonal*. $\Sigma_k = \lambda I_d$ (where I_d is the identity matrix). The covariance matrices are all identical, diagonal, with the same value in the diagonal. The Gaussians are spherical.
  - *Diagonal*. $\Sigma_k = \lambda_k I_d$. The covariance matrices are all diagonal with the same value in the diagonal, but they can be different. The Gaussians are spherical, but they may have different volumes.

Table 5 presents a list of the clustering algorithms used in the paper. A more complete list is used for the companion web page. Hierarchical clustering is used four times, combining the two distance metrics and two linkage methods, complete and single. SOM is used once, for Euclidean distance and bubble-type neighbor. Finally, EM clustering is used also once, for diagonal covariance matrices. The purpose of using several algorithms is to have a broad spectrum of partitions of the data, all of them reflecting some structure of the data, and to evaluate the validation indices over the full spectrum.

Table 6
Example of computation of error rate and validation indices for 10 realizations of the random process

Error	Dunn[mean,cen]	Silhouette	FOM	Rand
17.20	0.430	0.420	1.076	0.714
13.40	0.444	0.465	1.037	0.767
14.00	0.430	0.437	1.073	0.759
14.60	0.437	0.450	1.092	0.750
12.20	0.459	0.472	1.000	0.785
14.40	0.433	0.449	1.037	0.753
13.00	0.459	0.445	1.058	0.773
11.60	0.413	0.449	1.046	0.795
13.60	0.460	0.442	1.036	0.765
12.80	0.414	0.458	1.014	0.776

7. Model-based analysis

Our method is a simulation-based study presenting several clustering algorithms against different labeled point processes to study how the validation measures correlate with the error of the algorithm as a label operator on random labeled point processes. The simulation is based on models of labeled point processes, with different separations between the different classes (label values) that make the clustering problem more or less complicated, and can easily be controlled by a variance parameter.

The misclassification error is an estimator of the true error of the cluster operators [9]. Each clustering algorithm can be considered as a heuristically defined cluster operator (not learned). The purpose of the paper is to study the relationship between validation indices and the cluster-operator errors. To visualize this relationship we plot the indices against the errors. To quantify the degree of similarity between the validation indices and the misclassification errors, we use Kendall's rank correlation between their values, based on the recognition that usually the indices are used to compare the performance of algorithms.

As an example, Table 6 shows the misclassification error and some validation indices computed over 10 realizations of the second random process introduced in Section 2,

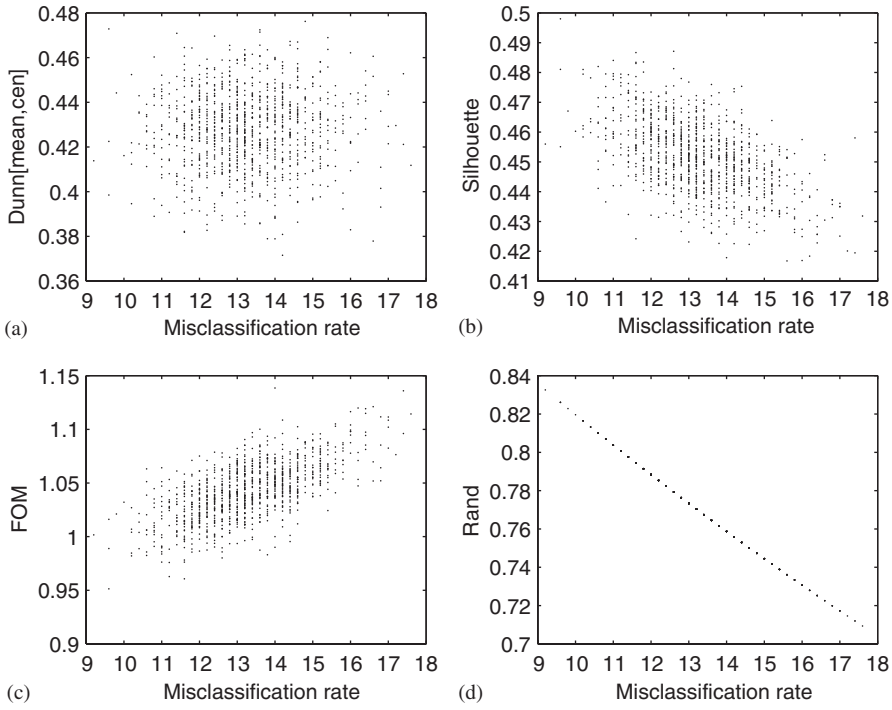

Fig. 4. Scatter plots against misclassification rate for (a) Dunn[mean,cen] index; (b) silhouette index; (c) FOM index; (d) Rand index.

with 250 samples for set, when the clusters are computed with the fuzzy C-means clustering algorithm. Based on 1000 pairs of values (error, validation), the computed rank correlations for this example are corr(Dunn[mean, cen]) = 0.03, corr(Silhouette) = 0.36, corr(FOM) = 0.45, corr(Rand) = 1.00. Fig. 4 shows the scatter plot over the 1000 realizations.

The overall procedure consists in simulating data, applying clustering, computing the indices, and comparing them to the error. The procedure can be characterized in six steps:

(1) *Specification of labeled point processes*: This stage requires determining some labeled point process with sufficient variability to obtain a broad range of error values, and also avoiding overly simple models that may be beneficial for some specific measures. We have approached this goal by allowing the processes to have a *variance multiplier*, ranging from very low variability in the data (allowing good performance of the clustering algorithms) to high variability, increasing the error by confusing the algorithms.
(2) *Generation of samples from the processes*: This step involves generating 100 sample sets (sets with their labels) for each process.
(3) *Application of clustering algorithms to the data*: This step involves computing the cluster labels for each data set using the clustering algorithms.
(4) *Estimation of the error of several algorithms from these samples*: The error is computed between the class labels, defined in step 2, against the cluster labels, defined in step 3, via Eq. (2).
(5) *Computation of the several validation measures for these algorithms on the same samples*: This step is done in a different way for relative indices than for internal and external ones.
   (a) Internal indices are computed based on the data points (spatial distribution of the points) and the cluster labels obtained in step 3.
   (b) External indices are computed based on the class labels, defined in step 2, and the cluster labels obtained in step 3.
   (c) Relative indices are computed based solely on the data points, applying repeatedly the clustering algorithms on subsets of the data, and computing the respective measures on the hold-out data. This is computationally the heaviest part of the process, because of the need to run the algorithm many times to compute a unique index (for example, for FOM and a 10-dimensional problem, the clustering algorithm needs to be run 10 times).
(6) *Quantification of the quality of the indices*: The measure of the ability of the validation indices to indicate the best clustering is determined by its rank correlation with the misclassification error, computed on all the samples based on the same labeled point process.

The analysis of the relationship between validation measures and misclassification, across several models, for different algorithms (label operators) and validation measures, displays the strengths and weaknesses of these measures.

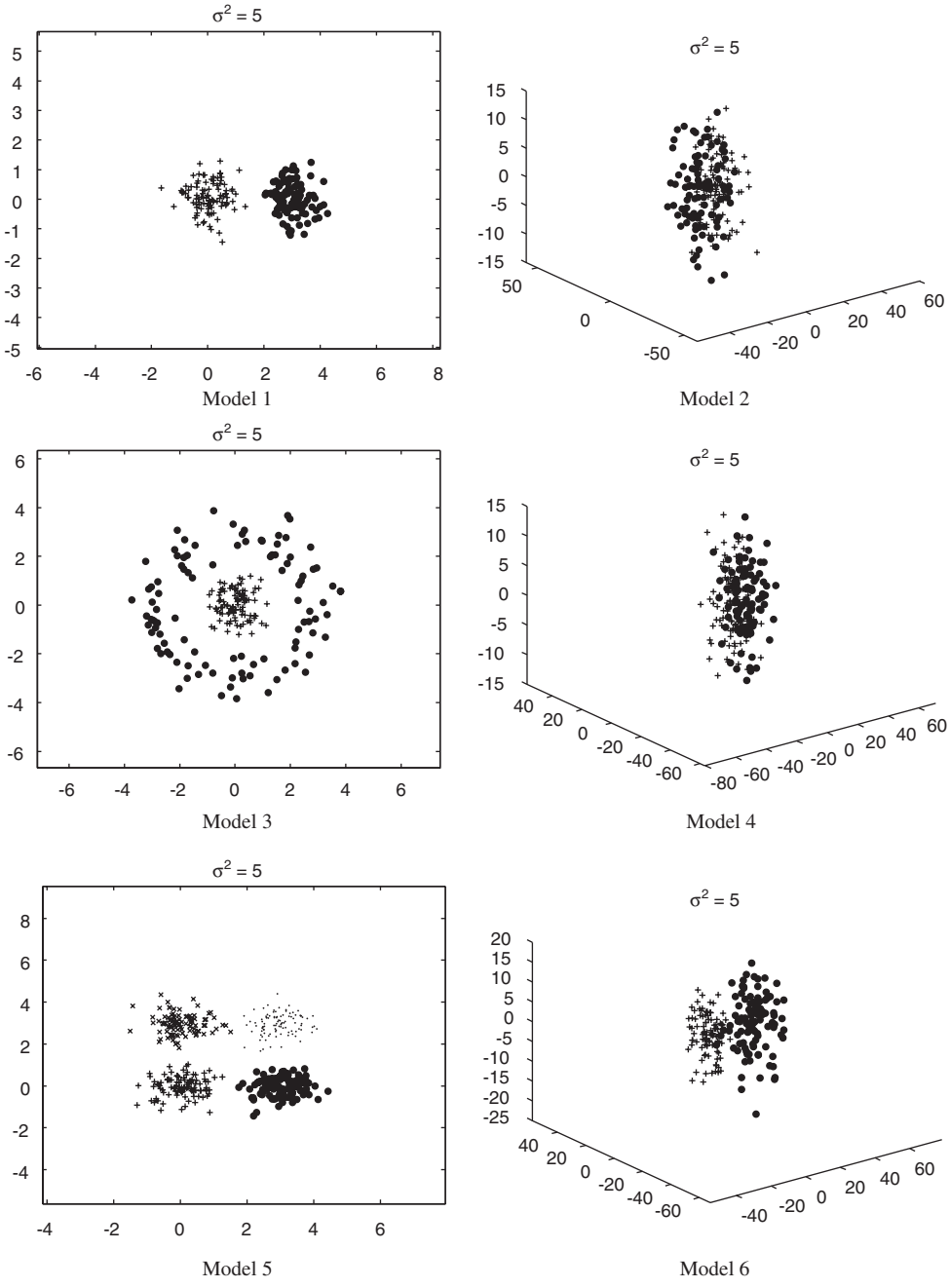

Fig. 5. Some examples of labeled sets generated for each model. The figures for models 4 and 6 show three-dimensional PCA plots.

8. Experiments

In this study, we generate sample point sets for three different models for the labeled point processes:

(1) *Model* 1—Two-dimensional mixture of two Gaussian distributions (Fig. 5).
(2) *Model* 2—Ten-dimensional mixture of two Gaussian distributions.
(3) *Model* 3—Two-dimensional mixture of two distributions where one distribution is Gaussian with covariance matrix $\sigma^2 I_d$ and the other is circular with normal distributions for both the radius and the angle, with variances σ^2 and 1 (Fig. 5).
(4) *Model* 4—Ten-dimensional mixture of a Gaussian and a distribution that is circular in its first two dimensions and Gaussian in its other eight dimensions.

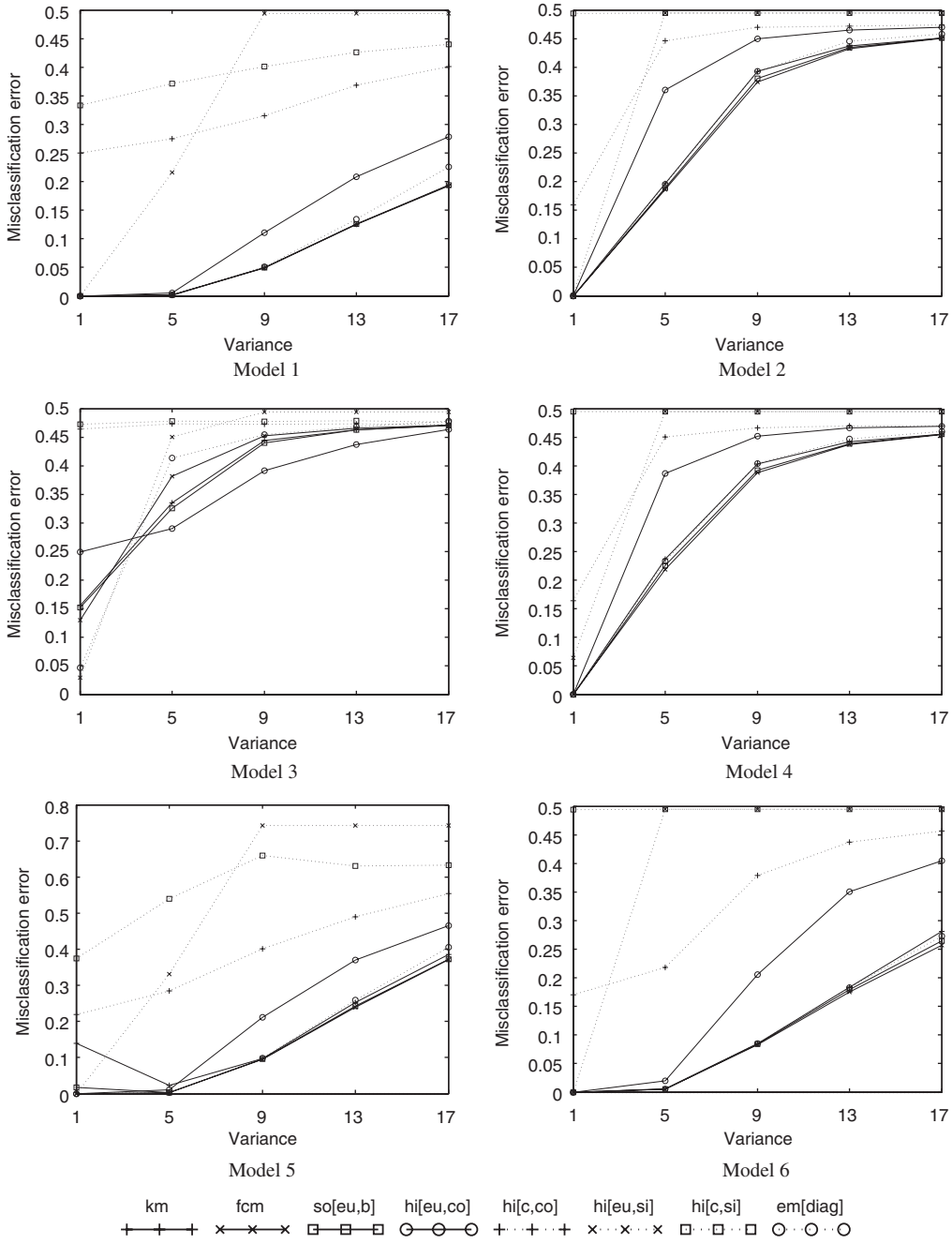

Fig. 6. Plot of misclassification as function of the variance of the model for several clustering algorithms.

(5) *Model* 5—Two-dimensional mixture of four Gaussian distributions.
(6) *Model* 6—Ten-dimensional mixture of a Gaussian and a distribution that is circular in its first two dimensions and Gaussian in its other eight dimensions. Class centers more separated than for model 4.

To obtain different error values, we use $\sigma^2 = 1$, 5, 9, 13 and 17. Fig. 5 shows examples of the six models, for $\sigma^2 = 5$, using three-dimensional PCA plot for models 4 and 6 [16]. Fig. 6 shows the misclassification error (in %) as a function of σ^2 for all the clustering algorithms analyzed.

Tables 5 and 7 present a list of the clustering algorithms and validation indices, respectively, used in the paper, along with unique IDs and short descriptions for using the figures and tables. Tables 8–13 show the rank correlations between the validation indices and the errors that have resulted from the experiments.

Table 7
Validation indices

Code	Algorithm	Parameters
$j1$	Trace criterion (J_e)	
$j2$	Determinant criterion (J_d)	
$j3$	Invariant criterion (J_f)	
dunn[cmean,max]	Dunn's validity index	Inter:meantocen–Intra:max
dunn[cmean,mean]	Dunn's validity index	Inter:meantocen–Intra:mean
dunn[cmean,cen]	Dunn's validity index	Inter:meantocen–Intra:cen
hubbd[eu]	Γ_D correlation	Distance = Euclidean
silh[eu]	Silhouette index	Distance = Euclidean
yfom	Figure of merit	
cfom	Corrected figure of merit	
stab[10,nn,LDA]	Stability	Rule = LDA[a]
stab[10,nn,PERC]	Stability	Rule = perceptron[a]
stab[10,nn,CEN]	Stability	Rule = centroid[a]
stab[10,nn,KNN,3]	Stability	Rule = 3NN[a]
hubert	Γ correlation	
nhubert	Normalized Γ^* correlation	
rand	Rand statistic	
jacc	Jaccard coefficient	
fm	Folkes and Mallows index	

[a]Repetitions = 10, normalization = No.

Table 8
Kendall's correlation for model 1

Index	km	fcm	so[eu,b]	hi[eu,co]	hi[c,co]	hi[eu,si]	hi[c,si]	em[diag]	Av.
$j1$	0.81	0.81	0.81	0.8	0.62	0.51	0.27	0.79	0.68
$j2$	0.8	0.8	0.79	0.8	0.57	0.5	0.15	0.78	0.65
$j3$	0.79	0.8	0.79	0.72	0.49	0.59	0.16	0.71	0.63
dunn[cmean,max]	0.77	0.77	0.77	0.76	0.57	0.69	0.35	0.77	0.68
dunn[cmean,mean]	0.8	0.8	0.8	0.81	0.7	0.7	0.47	0.8	0.74
dunn[cmean,cen]	0.8	0.8	0.8	0.81	0.79	0.71	0.54	0.8	0.76
hubbd[eu]	0.74	0.75	0.74	0.28	0.18	0.69	0.61	0.61	0.57
silh[eu]	0.81	0.81	0.81	0.83	0.75	0.58	0.65	0.81	0.76
yfom	0.77	0.77	0.77	0.71	0.45	0.56	0.27	0.77	0.63
cfom	0.77	0.77	0.77	0.71	0.45	0.56	0.27	0.77	0.63
stab[10,nn,lda]	0.81	0.81	0.8	0.76	0.42	0.48	0.09	0.83	0.63
stab[10,nn,perc]	0.73	0.72	0.72	0.72	0.4	0.36	0.2	0.75	0.57
stab[10,nn,cen]	0.81	0.81	0.8	0.76	0.42	0.51	0.09	0.83	0.63
stab[10,nn,knn,3]	0.8	0.8	0.8	0.76	0.37	0.37	0.13	0.82	0.61
hubert	0.99	0.99	0.99	0.84	0.42	0.59	0.68	0.93	0.8
nhubert	1	1	1	1	0.98	1	0.97	1	0.99
rand	1	1	1	1	1	1	1	1	1
jacc	1	1	1	0.96	0.76	0.63	0.18	0.99	0.82
fm	1	1	1	0.95	0.73	0.63	0.34	0.98	0.83

9. Analysis

9.1. Internal validation

Six internal validation indices have been analyzed: trace criterion J_e [16], determinant criterion J_d [16], invariant criterion J_f [16], Dunn's index, Γ_D correlation with Euclidean distance matrix and silhouette index. The 18 variants of Dunn's index correspond to all possible combinations of linkages and cluster size measures, and are presented in the companion web page.

9.1.1. Trace criterion, determinant criterion and invariant criterion

For the two-dimensional mixture of Gaussians, these criteria ($j1$, $j2$ and $j3$, respectively) can exhibit good behavior, their rank correlation with the error reaching values around 0.8 when used on clustering algorithms that tend to generate compact clusters. For other models or clustering algorithms, the rank correlation values are notably lower, specially for the situation where there is a circular distribution (model 3). The average correlation lies below 0.7 for all models, and below 0.5 for models 2–4, indicating very low information

Table 9
Kendall's correlation for model 2

Index	km	fcm	so[eu,b]	hi[eu,co]	hi[c,co]	hi[eu,si]	hi[c,si]	em[diag]	Av.
$j1$	0.65	0.66	0.65	0.47	0.44	0.55	0.14	0.45	0.5
$j2$	0.35	0.44	0.36	0.28	0.37	0.57	0.14	0.2	0.34
$j3$	0.33	0.43	0.34	0.36	0.39	0.57	0.14	0.14	0.34
dunn[cmean,max]	0.51	0.5	0.5	0.35	0.36	0.57	0.12	0.4	0.41
dunn[cmean,mean]	0.63	0.63	0.63	0.42	0.46	0.57	0.07	0.42	0.48
dunn[cmean,cen]	0.63	0.63	0.63	0.4	0.46	0.57	0.13	0.42	0.48
hubbd[eu]	0.65	0.68	0.67	0.3	0.38	0.56	0.04	0.41	0.46
silh[eu]	0.65	0.65	0.65	0.33	0.46	0.55	0.01	0.56	0.48
yfom	0.66	0.69	0.68	0.56	0.44	0.55	0.1	0.67	0.54
cfom	0.66	0.69	0.68	0.56	0.44	0.55	0.1	0.67	0.54
stab[10,nn,lda]	0.61	0.65	0.65	0.36	0.37	0.55	0	0.66	0.48
stab[10,nn,perc]	0.58	0.61	0.61	0.49	0.42	0.32	0.08	0.62	0.47
stab[10,nn,cen]	0.62	0.65	0.65	0.39	0.39	0.55	0.07	0.66	0.5
stab[10,nn,knn,3]	0.61	0.65	0.64	0.29	0.34	0.55	0.02	0.65	0.47
hubert	0.86	0.98	0.91	0.27	0.38	0.95	0.38	0.5	0.65
nhubert	0.99	1	0.99	0.91	0.94	1	0.97	0.96	0.97
rand	1	1	1	1	1	1	1	1	1
jacc	0.91	0.99	0.94	0.36	0.44	0.95	0.38	0.66	0.7
fm	0.91	0.99	0.94	0.34	0.44	0.95	0.38	0.65	0.7

Table 10
Kendall's correlation for model 3

Index	km	fcm	so[eu,b]	hi[eu,co]	hi[c,co]	hi[eu,si]	hi[c,si]	em[diag]	Av.
$j1$	0.09	0.12	0.1	0.14	0.07	0.66	0.04	0.29	0.19
$j2$	0.2	0.18	0.2	0.07	0.04	0.65	0.09	0.17	0.2
$j3$	0.09	0.12	0.08	0.13	0.08	0.65	0.06	0.31	0.19
dunn[cmean,max]	0.52	0.47	0.53	0.32	0.05	0.64	0	0.25	0.35
dunn[cmean,mean]	0.3	0.17	0.31	0.06	0.01	0.52	0.01	0.01	0.17
dunn[cmean,cen]	0.17	0.04	0.18	0.01	0.09	0.56	0.1	0.05	0.15
hubbd[eu]	0.49	0.51	0.5	0.12	0.03	0.72	0.21	0.51	0.39
silh[eu]	0.56	0.39	0.58	0.49	0.1	0.2	0.11	0.29	0.34
yfom	0.62	0.57	0.63	0.55	0.06	0.52	0.02	0.6	0.45
cfom	0.62	0.57	0.63	0.55	0.06	0.52	0.02	0.6	0.45
stab[10,nn,lda]	0.21	0.3	0.24	0.24	0.05	0.35	0.01	0.5	0.24
stab[10,nn,perc]	0.19	0.29	0.21	0.25	0.06	0.5	0.02	0.45	0.25
stab[10,nn,cen]	0.19	0.29	0.22	0.26	0.05	0.35	0.01	0.51	0.24
stab[10,nn,knn,3]	0.21	0.29	0.23	0.31	0.01	0.25	0.04	0.5	0.23
hubert	0.79	0.84	0.8	0.34	0.05	0.15	0.22	0.09	0.41
nhubert	0.99	0.99	0.99	0.96	0.91	1	0.63	0.9	0.92
rand	1	1	1	1	1	1	1	1	1
jacc	0.84	0.89	0.85	0.63	0.12	0.19	0.2	0.05	0.47
fm	0.84	0.89	0.85	0.58	0.12	0.17	0.2	0.02	0.46

represented by the indices. Departure from Gaussian models or clustering algorithms that do not generate compact clusters negatively affects the quality of these indices.

9.1.2. Dunn index

On average, for a low-dimensional mixture of Gaussians (models 2 and 5), the Dunn index (*dunn*) attains better rank correlation when used with a linkage based on the centroids, both *centroid* (*dunn*[*cmean*,*cen*]) and *average to centroids* (*dunn*[*cmean*,*mean*]), reaching average values above 0.8, and the cluster size measure does not considerably affect the results. For the other models, this index has a consistent low correlation to the error.

9.1.3. Γ_D correlation with Euclidean distance matrix

This index (*hubbd*) has an average correlation between 0.4 and 0.5, except for model 5, with average correlation of 0.66. Its behavior is highly variable, reaching its maximum value for hierarchical clustering on model 3 (correlation of 0.72) (Table 10).

9.1.4. Silhouette

The silhouette index (*silh*) is affected by lack of normality and higher-dimensional space. The correlation is not low for models 1, 5 and 6, but it drops below 50% for models 2–4. The reasons for this may reside in the distance-based nature of the index and the fact that for some models the index

Table 11
Kendall's correlation for model 4

Index	km	fcm	so[eu,b]	hi[eu,co]	hi[c,co]	hi[eu,si]	hi[c,si]	em[diag]	Av.
$j1$	0.61	0.63	0.63	0.43	0.42	0.52	0.12	0.44	0.48
$j2$	0.34	0.4	0.35	0.24	0.35	0.53	0.12	0.22	0.32
$j3$	0.33	0.4	0.35	0.33	0.37	0.53	0.12	0.16	0.32
dunn[cmean,max]	0.48	0.48	0.47	0.33	0.38	0.53	0.1	0.39	0.4
dunn[cmean,mean]	0.59	0.58	0.59	0.39	0.43	0.53	0.07	0.4	0.45
dunn[cmean,cen]	0.59	0.58	0.59	0.37	0.43	0.53	0.1	0.4	0.45
hubbd[eu]	0.64	0.68	0.67	0.27	0.35	0.53	0.03	0.44	0.45
silh[eu]	0.62	0.63	0.62	0.3	0.43	0.53	0.02	0.53	0.46
yfom	0.66	0.69	0.68	0.53	0.41	0.53	0.07	0.67	0.53
cfom	0.66	0.69	0.68	0.53	0.41	0.53	0.07	0.67	0.53
stab[10,nn,lda]	0.57	0.62	0.63	0.33	0.37	0.22	0.01	0.64	0.42
stab[10,nn,perc]	0.55	0.59	0.58	0.45	0.4	0.29	0.05	0.6	0.44
stab[10,nn,cen]	0.59	0.64	0.62	0.35	0.37	0.22	0.06	0.64	0.44
stab[10,nn,knn,3]	0.57	0.62	0.61	0.27	0.34	0.23	0.01	0.63	0.41
hubert	0.84	0.97	0.9	0.24	0.38	0.95	0.3	0.51	0.64
nhubert	0.99	1	0.99	0.89	0.94	1	0.97	0.96	0.97
rand	1	1	1	1	1	1	1	1	1
jacc	0.89	0.98	0.94	0.31	0.44	0.95	0.3	0.65	0.68
fm	0.89	0.98	0.94	0.3	0.43	0.95	0.3	0.62	0.68

Table 12
Kendall's correlation for model 5

Index	km	fcm	so[eu,b]	hi[eu,co]	hi[c,co]	hi[eu,si]	hi[c,si]	em[diag]	Av.
$j1$	0.69	0.84	0.82	0.85	0.71	0.37	0.27	0.82	0.67
$j2$	0.63	0.84	0.81	0.86	0.72	0.5	0.2	0.83	0.67
$j3$	0.69	0.84	0.82	0.85	0.71	0.37	0.27	0.82	0.67
dunn[cmean,max]	0.76	0.78	0.76	0.78	0.6	0.41	0.17	0.78	0.63
dunn[cmean,mean]	0.78	0.8	0.8	0.78	0.55	0.36	0.2	0.8	0.63
dunn[cmean,cen]	0.78	0.8	0.79	0.77	0.5	0.4	0.19	0.8	0.63
hubbd[eu]	0.51	0.78	0.74	0.59	0.58	0.75	0.62	0.72	0.66
silh[eu]	0.7	0.83	0.81	0.84	0.8	0.44	0.73	0.84	0.75
yfom	0.55	0.79	0.76	0.77	0.67	0.59	0.27	0.79	0.65
cfom	0.55	0.79	0.76	0.77	0.67	0.59	0.27	0.79	0.65
stab[10,nn,cen]	0.37	0.83	0.78	0.81	0.62	0.38	0.25	0.85	0.61
hubert	0.81	0.99	0.96	0.87	0.79	0.33	0.24	0.93	0.74
nhubert	0.89	0.99	0.98	0.94	0.87	0.67	0.87	0.99	0.9
rand	0.89	0.99	0.98	0.93	0.86	0.76	0.92	0.98	0.91
jacc	0.88	0.99	0.98	0.94	0.87	0.39	0.46	0.98	0.81
fm	0.88	0.99	0.98	0.93	0.86	0.39	0.28	0.98	0.79

flattens fast as a function of the variance of the model, as is shown in Fig. 7 for model 4. In this figure we use a three-dimensional surface view of the scatter plot to appreciate where the majority of the points lie.

9.2. Relative validation

9.2.1. Figure of merit

The figure of merit (*yfom*) shows consistent high correlation (above 0.6) for most of the clustering algorithms that tend to form compact clusters, and for most of the models, but falls below 0.5 correlation when used for algorithms based in correlation instead of Euclidean distance (algorithms hi [C,Co] and hi [C,Si]). A key drawback, shared by other internal and relative indices, is that it relies on the disposition of the points to be clustered. Therefore it is affected by changes in the variance of the model, even when the clustering algorithms may yield consistent results. This is exemplified in Fig. 8(a), where each strip is generated from a different variance multiplier. The FOM lies in different ranges for different variances, while the clustering error covers a broad range for all of them. In this case the FOM is unable to accurately predict the quality of the clusters. The reason is clear when comparing the average values for FOM against the values for the error rate, as a function of the variance (Figs. 8(b) and (c), respectively). For FOM, the index is essentially a function of the variability of the data, independent of the classifier used or individual samples of the data.

Table 13
Kendall's correlation for model 6

Index	km	fcm	so[eu,b]	hi[eu,co]	hi[c,co]	hi[eu,si]	hi[c,si]	em[diag]	Av.
$j1$	0.84	0.83	0.83	0.79	0.68	0.55	0.27	0.79	0.7
$j2$	0.75	0.78	0.75	0.64	0.53	0.57	0.27	0.69	0.62
$j3$	0.72	0.77	0.72	0.65	0.52	0.57	0.27	0.59	0.6
dunn[cmean,max]	0.75	0.75	0.75	0.71	0.59	0.57	0.15	0.73	0.63
dunn[cmean,mean]	0.83	0.82	0.82	0.76	0.72	0.57	0.13	0.78	0.68
dunn[cmean,cen]	0.82	0.82	0.82	0.76	0.74	0.57	0.21	0.78	0.69
hubbd[eu]	0.77	0.78	0.77	0.38	0.42	0.57	0.12	0.7	0.56
silh[eu]	0.84	0.83	0.83	0.74	0.74	0.56	0.02	0.81	0.67
yfom	0.81	0.81	0.8	0.75	0.58	0.56	0.09	0.8	0.65
cfom	0.81	0.81	0.8	0.75	0.58	0.56	0.09	0.8	0.65
stab[10,nn,lda]	0.84	0.85	0.85	0.74	0.55	0.57	0.04	0.85	0.66
stab[10,nn,perc]	0.75	0.73	0.72	0.69	0.55	0.46	0.09	0.72	0.59
stab[10,nn,cen]	0.84	0.85	0.85	0.75	0.57	0.57	0.08	0.85	0.67
stab[10,nn,knn,3]	0.84	0.84	0.85	0.69	0.53	0.57	0.05	0.84	0.65
hubert	0.99	1	0.99	0.67	0.6	0.96	0.68	0.93	0.85
nhubert	1	1	1	0.99	0.98	1	0.99	1	1
rand	1	1	1	1	1	1	1	1	1
jacc	1	1	1	0.82	0.74	0.96	0.68	0.98	0.9
fm	0.99	1	1	0.8	0.73	0.96	0.68	0.98	0.89

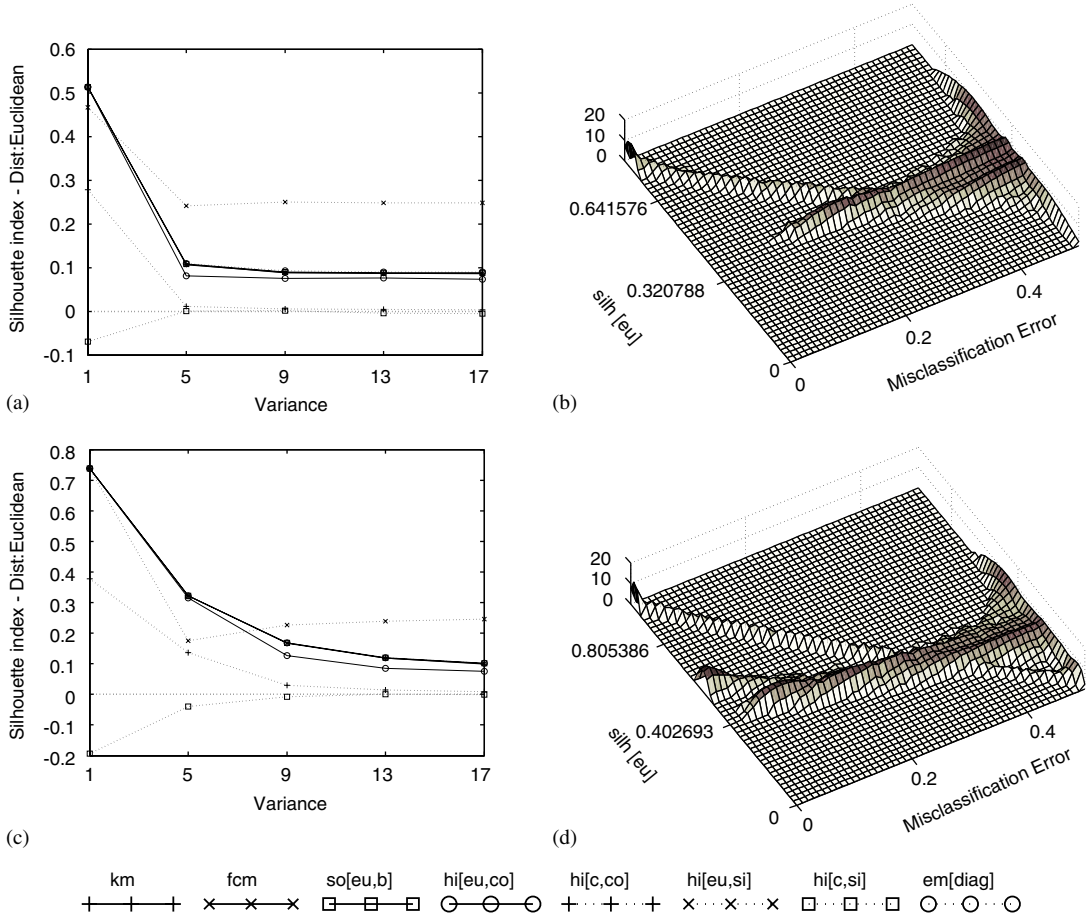

Fig. 7. Silhouette index as function of the variance and its scatter plot (a,b) for model 4 and (c,d) for model 6.

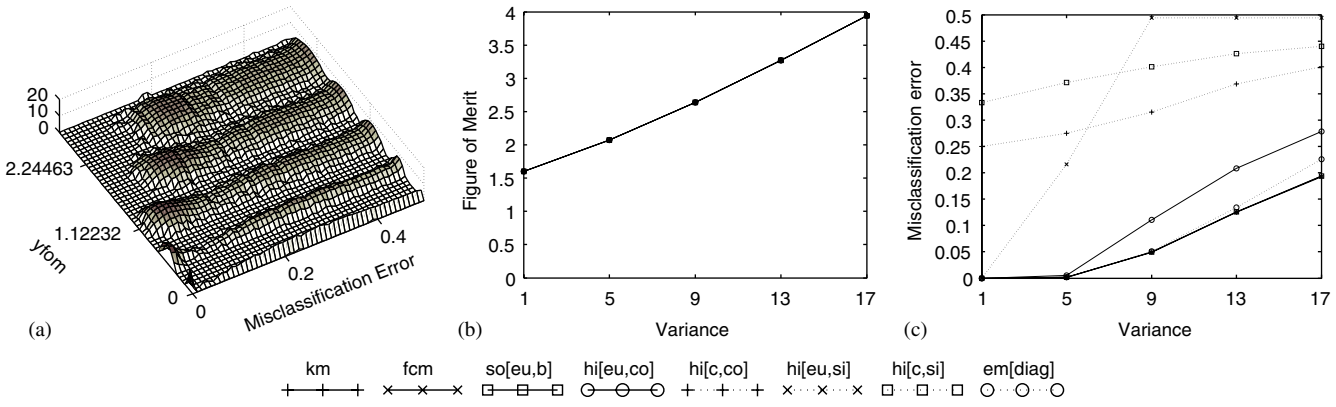

Fig. 8. Figure of merit: (a) scatter plots, (b) FOM as function of the variance, (c) misclassification rate as a function of the variance.

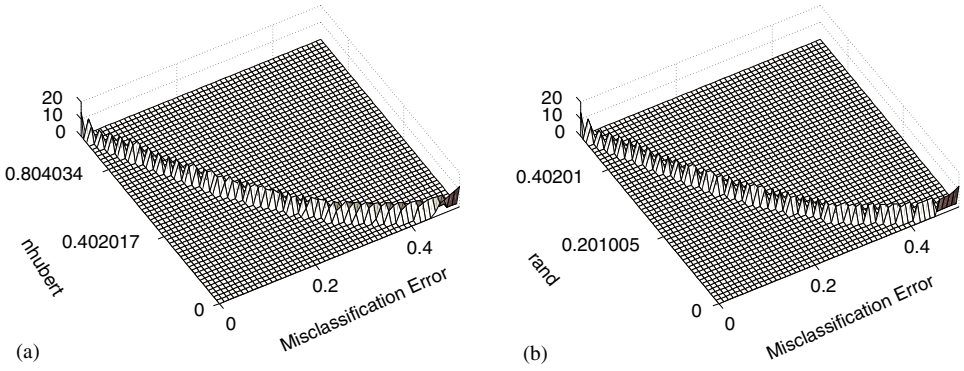

Fig. 9. Scatter plots for (a) normalized Γ^* correlation and (b) Rand index.

9.2.2. Stability

The stability indices (*stab*) show some of the highest correlation values (above 0.8 for models 1 and 6) but very low ones also (below 0.5 for model 3). The overall behavior is not significantly better than some internal validation indices, while the computational cost is extremely larger, involving a Monte Carlo approach (to partition the data) plus clustering and training a classifier in each step.

9.3. External validation

The only indices to have an average and combined correlation close to 1 are the Rand index (*rand*) and the normalized Γ^* correlation (*nhubert*). Figs. 9(a) and (b) show that even if the relationships are not linear (the average linear Pearson correlation between them and the indices being 0.98, not shown here), there is a one-to-one relationship between them and the error.

The same observation does not repeat for the other external measures: Γ and Γ^* correlation (*hubert*, *nhubert*), the Jaccard coefficient (*jacc*) and the Folkes and Mallow index (*fm*). Fig. 10 shows that for high values of the error the relationship between the indices and the error is no longer one-to-one.

10. Conclusion

For simulations or when additional information is known about the true classes, the choice of validity index is clearly in favor of external indices; however, not all of them are good predictors of the clustering error. For external indices, the Rand statistic is the best replacement for the error rate: it can be computed quickly, it does not deviate from the error for the 2-class case, and the deviation is small for models with more than two classes (an average correlation of 0.9 for model 5). In some cases the other external indices give information associated with the Rand index, like the correlation between the similarity matrices (hubert and nhubert), but in other cases they measure different properties of the relationship between clusters and classes, like the Jaccard coefficient and Folkes and Mallows indices, and may not correlate well with clustering error.

In the absence of information to apply external validation, intuitively it might seem that the relative indices should be more desirable than the internal indices since they try to exploit data redundancy; however, most of the results show that even for simple models the relative indices do not give substantial improvement over the simpler internal indices, while at the same time increasing the computational costs beyond the limits of a desktop PC. In general, internal indices have a

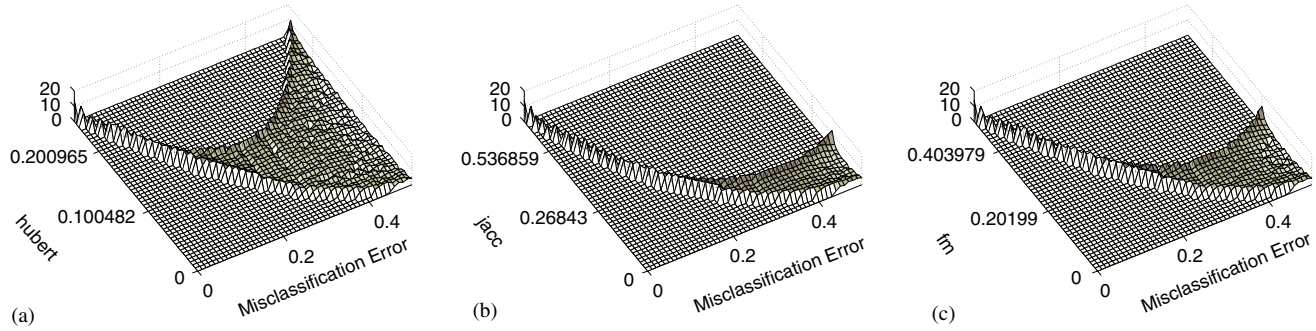

Fig. 10. Scatter plots for (a) Γ^* correlation, (b) Jaccard coefficient and (c) Folkes and Mallow index.

satisfactory behavior when the conditions are appropriately constrained, such as using Gaussian models with compact clustering algorithms; however, when the models get more complex or the algorithms give more complex clusters, the internal indices fail to correlate well with the error of the algorithm. In that case, the relative indices also fail to almost the same degree. If a choice is to be made, based on our extensive simulations among varied models, it appears that the silhouette index should be the choice, since it almost always outperforms the other internal indices, and its performance is close to that of the best relative indices.

What we believe, as has been demonstrated by our extensive analysis, is that, when investigating the performance of a proposed clustering algorithm, it is best to consider varied models and use the true clustering error. In applications where one wishes to get an idea of the accuracy of the clustering when there is only a single sample, unless there is some evidence, either in terms of sufficient data for model estimation or prior model knowledge, that a validity measure is well-correlated to the error rate for the algorithm, one should not refer to a validity score to justify a claim of clustering accuracy. Indeed, relative to clustering being scientifically constitutive, the historical evolution of validity indices might be seen as being premature. Without a predictive theory of clustering, there is no hope of checking the meaningfulness of a validity index. What is now needed is a rigorous accounting of the distributional conditions that warrant the use of an already proposed validity index and the development of new validity indices that highly correlate to the performance of clustering algorithms under well-documented circumstances.

References

[1] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surv. 31 (3) (1999) 264–323.

[2] A.K. Jain, A. Topchy, M. Law, J.M. Buhmann, Landscape of clustering algorithms, in: Pattern Recognition, 2004, ICPR 2004, Proceedings of the 17th International Conference on, vol. 1, iss., 23–26 August 2004, 2004, pp. 260–263.

[3] M.B. Eisen, P. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[4] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, J. Comput. Biol. 6 (3/4) (1999) 281–297.

[5] H. Chipman, T. Hastie, R. Tibshirani, Clustering Microarray Data, Statistical Analysis of Gene Expression Microarray Data, Chapman & Hall, CRC Press, London, Boca Raton, 2003.

[6] M. Brun, C.D. Johnson, K.S. Ramos, Clustering: revealing intrinsic dependencies in microarray data, in: Genomic Signal Processing and Statistics, EURASIP Book Series on Signal Processing and Communications, Hindawi Publishing Corporation, 2005, pp. 129–162.

[7] E.R. Dougherty, U. Braga-Neto, Epistemology of computational biology: mathematical models and experimental prediction as the basis of their validity, Biol. Syst. 14(1) (2006) 65–90.

[8] P. Richard Feynman, The Strange Theory of Light and Matter, Princeton University Press, Princeton, 1985.

[9] E.R. Dougherty, M. Brun, A probabilistic theory of clustering, Pattern Recognition 37 (2004) 917–925.

[10] L. Fisher, J.W. Van-Ness, Admissible clustering procedures, Biometrika 58 (1) (1971) 91–104.

[11] J.W. Van-Ness, Admissible clustering procedures, Biometrika 60 (2) (1973) 422–424.

[12] S. Guenter, H. Bunke, Validation indices for graph clustering, in: J.-M. Jolion, W. Kropatsch, M. Vento (Eds.), Proceedings of the 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, 2001, pp. 229–238.

[13] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Intell. Inf. Syst. J. 17 (2–3) (2001) 107–145.

[14] Z. Lubovac, B. Olsson, P. Jonsson, K. Laurio, M.L. Anderson, Biological and statistical evaluation of clusterings of gene expression profiles, in: C.E. D'Attellis, V.V. Kluev, N.E. Mastorakis (Eds.), Proceedings of Mathematics and Computers in Biology and Chemistry, WSES Press, 2001, pp. 149–155.

[15] V. Roth, T. Lange, M. Braun, J.M. Buhmann, A resampling approach to cluster validation, in: B.R. Wolfgang Hrdle (Ed.), Proceedings in Computational Statistics: 15th Symposium Held in Berlin, Germany 2002 (COMPSTAT2002), Physica-Verlag, Heidelberg, 2002, pp. 123–128.

[16] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2002.

[17] K.Y. Yeung, D.R. Haynor, W.L. Ruzzo, Validating clustering for gene expression data, Bioinformatics 17 (2001) 309–318.

[18] E.R. Dougherty, J. Barrera, M. Brun, S. Kim, R.M. Cesar, Y. Chen, M.L. Bittner, J.M. Trent, Inference from clustering with application to gene-expression microarray, J. Comput. Biol. 9 (1) (2002) 105–126.

[19] F. Azuaje, A cluster validity framework for genome expression data, Bioinformatics 18 (2002) 319–320.

[20] F. Azuaje, N. Bolshakova, Clustering genomic expression data, in: D. Berrar, W. Dubitzky, M. Granzow (Eds.), Design and Evaluation Principles, A Practical Approach to Microarray Data Analysis,

Copyright 2002, Kluwer Academic Publishers, Boston, Dordrecht, London.
[21] N. Bolshakova, F. Azuaje, Cluster validation techniques for genome expression data, Technical Report TCD-CS-2002-33, Computer Science Department, The University of Dublin.
[22] T. Lange, M. Braun, V. Roth, J.M. Buhmann, Stability-based model selection, Advances in Neural Information Processing Systems.
[23] V. Roth, M. Braun, T. Lange, J.M. Buhmann, Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data, Springer, Berlin, 2002.
[24] T. Kohonen, Self-Organizing Maps, second ed., Springer, New York, 1997.
[25] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation, Proc. Natl. Acad. Sci. USA, vol. 96, 1999, pp. 2907–2912.
[26] P. Toronen, M. Kolehmainen, G. Wong, E. Castren, Analysis of gene expression data using self-organizing maps, FEBS Lett. 451 (1999) 142–146.
[27] J. Wang, J. Delabie, H.C. Aasheim, E. Smeland, O. Myklebost, Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study, BMC Bioinformatics 3 (1) (2002) 36.
[28] C. Fraley, A.E. Raftery, Mclust: software for model-based clustering and discriminant analysis, J. Classification 16 (1999) 297–306.
[29] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, W.L. Ruzzo, Model-based clustering and data transformation for gene expression data, Bioinformatics 17 (10) (2001) 977–987.
[30] C. Fraley, A.E. Raftery, Model-based clustering, discriminant analysis, and density estimation, J. Am. Statistical Association 97 (2002) 611–631, Technical Report No. 380, Department of Statistics, University of Washington, October 2000.

About the Author—MARCEL BRUN received his Ph.D. in Computer Sciences from the University of So Paulo, Brazil. He was involved in research in genomics signal processing at the Electrical Engineering department, at Texas A&M University, and the Department of Biochemistry and Molecular Biology, at the University of Louisville, from 2000 to 2004. Currently he is an Associated Investigator at TGen, Arizona, with research focusing on computational biology, centered in design and simulation of genetic networks and analysis of large-scale biological data.

About the Author—BRENT CARROLL is working on the B.S. degree in the Department of Electrical Engineering at Rice University. He worked on clustering validation during an internship in the Genomic Signal Processing Laboratory at Texas A&M University.

About the Author—EDWARD R. DOUGHERTY is a Professor in the Department of Electrical Engineering at Texas A&M University in College Station, TX, Director of the Genomic Signal Processing Laboratory at Texas A&M University, and Director of the Computational Biology Division of the Translational Genomics Research Institute in Phoenix, AZ. He holds a Ph.D. in mathematics from Rutgers University and an M.S. in Computer Science from Stevens Institute of Technology. He is author of 12 books, editor of five others, and author of more than 190 journal papers. He is an SPIE fellow, is a recipient of the SPIE President's Award, and has served as editor of the Journal of Electronic Imaging for six years. Prof. Dougherty has contributed extensively to the statistical design of nonlinear operators for image processing and the consequent application of pattern recognition theory to nonlinear image processing. His current research is focused in genomic signal processing, with the central goal being to model genomic regulatory mechanisms for the purposes of diagnosis and therapy.

About the Author—JIANPING HUA received the B.S. and M.S. degrees in Electrical Engineering from the Tsinghua University, Beijing, China, in 1998 and 2000, respectively. He received the Ph.D. degree in Electrical Engineering from Texas A&M University in 2004. Currently, he is a senior post-doc fellow in Translational Genomics Research Institute (TGen) at Phoenix, AZ. His main research interest lies in bioinformatics, genomic signal processing, signal and image processing, image and video coding and statistic pattern recognition.

About the Author—JAMES LOWEY is the Assistant Director of the High Performance Biocomputing Center at the Translational Genomics Research Institute (TGen). Mr. Lowey is responsible for the architecture, management and daily operation of TGen's high performance computer systems that include a 512 node parallel cluster computer and various large SMP machines. He works closely with TGen scientists to implement and provide computational tools and data management systems to facilitate and accelerate translational genomics research. Prior to joining TGen, Mr. Lowey worked as a consultant at various Fortune 500 companies, implementing and managing large-scale computational systems.

About the Author—CHAO SIMA received his Ph.D. degree in 2006 in the Department of Electrical and Computer Engineering at Texas A&M University in College Station, under the supervision of Dr. E.R. Dougherty. He received his B.E. degree in 1995 at Xi'an Jiaotong University, PR China. He is now working as a Postdoc researcher in the Department of Statistics at Texas A&M University in College Station, and his current research interest includes feature selection and classification in genomic signal processing, Bayesian analysis and developing statistical models for gene-expression microarray and aCGH data, and other sources of biological data.

About the Author—Dr. EDWARD SUH is the Chief Information Officer of the Translational Genomics Research Institute (TGen), where he leads and manages Biomedical Informatics, Information Technology and High Performance Biocomputing programs. Dr. Suh and his team develop and provide data mining and data management systems, computational algorithms and application software, and high-performance biocomputing and secure information technology infrastructure for rapid collection, integration, analysis and dissemination of biomedical data for the discovery of novel biomarkers, diagnostics and prognostics, leading to the treatment of diseases. Dr. Suh has served multiple NIH grants in the capacity of an IT director and an investigator. Dr. Suh joined TGen after 15 years at NIH, where he held increasingly important positions in the Division of Computational Bioscience (DCB) of the Center for Information Technology, finally serving as its Associate Director. Dr. Suh began his career in electrical engineering. After earning a Sc.D. in computer science from George Washington University, he married the two career fields and now specializes in the application of computational science and engineering methodologies to biomedical data mining, systems biology and high performance biocomputing. Dr. Suh authored and co-authored numerous articles in journals such as Science, Journal of Computational Biology, Bioinformatics and Cancer Research.