Pattern Recognition 43 (2010) 1292–1305

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A stability based validity method for fuzzy clustering

M. Falasconi [a,*], A. Gutierrez [b,c], M. Pardo [a,d], G. Sberveglieri [a], S. Marco [b,c]

[a] SENSOR Laboratory, Department of Chemistry and Physics for Engineering and Materials, University of Brescia and INFM-CNR, Via Valotti 9, I-25123 Brescia, Italy
[b] Departament d'Electrònica, Universitat de Barcelona, Martí i Franquès, 1, 08028 Barcelona, Spain
[c] Artificial Olfaction Group, Institute for Bioengineering of Catalonia (IBEC), Baldiri i Rexach 13, 08028 Barcelona, Spain
[d] Computational Molecular Biology Department, Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin, Germany

ARTICLE INFO

Article history:
Received 8 October 2008
Received in revised form
25 September 2009
Accepted 2 October 2009

Keywords:
Fuzzy *c*-means
Cluster validity
Number of clusters
Cluster stability

ABSTRACT

An important goal in cluster analysis is the internal validation of results using an objective criterion. Of particular relevance in this respect is the estimation of the optimum number of clusters capturing the intrinsic structure of your data. This paper proposes a method to determine this optimum number based on the evaluation of fuzzy partition stability under bootstrap resampling. The method is first characterized on synthetic data with respect to hyper-parameters, like the fuzzifier, and spatial clustering parameters, such as feature space dimensionality, clusters degree of overlap, and number of clusters. The method is then validated on experimental datasets. Furthermore, the performance of the proposed method is compared to that obtained using a number of traditional fuzzy validity rules based on the cluster compactness-to-separation criteria. The proposed method provides accurate and reliable results, and offers better generalization capabilities than the classical approaches.

1. Introduction

Cluster analysis (CA) is an unsupervised learning method frequently used in exploratory data analysis to make a preliminary assessment of the data structure, to discover hidden structures in the data sets, and to extract (or compress) the information by drawing data prototypes. Clustering essentially refers to the assignment of patterns into groups (clusters) so that the objects belonging to the same group are more similar to each other than those within different groups [1].

Most of the partitional clustering algorithms output a hard (or crisp) partition of the data. This means that a pattern can belong to one and only one cluster of the partition; hence, the clusters in a hard partition are disjoint. Fuzzy clustering [2] extends the notion of partition to associate each pattern with every cluster using a membership function whose values span from zero to one. Membership values closer to one indicate higher confidence in the assignment of the pattern to the given cluster. The output of fuzzy clustering algorithms is then called "fuzzy" partition. Nowadays, fuzzy clustering is a vibrant area of research that finds application in many different fields such as data mining, image analysis, and bioinformatics [3].

The most popular fuzzy clustering technique is arguably the Fuzzy *C*-Means (FCM) algorithm proposed by Bezdek 25 years ago [4], though many variants have been proposed since then [3]. The algorithm selects a random initial fuzzy partition, and then it tries to minimize a proper objective function—for example the weighted sum of square errors—by iteratively changing the patterns membership and the cluster prototype position. This procedure is repeated until convergence, which means until the variation of the cost function is less than a pre-specified (small) amount.

There are some recognized benefits of fuzzy clustering with respect to hard partitioning methods. First, a hard partition can be viewed as a special case of a fuzzy partition by simply assigning the pattern to the class where it gets the maximum membership. Second, a fuzzy partition contains more information than a hard one since the membership value can be interpreted as an indication of the similarity of a certain pattern to a certain cluster. In case of overlapping clusters, the membership values show which patterns are located in this overlapping region. Finally, FCM presents some computational advantages with respect to its crisp counterpart, known as hard *c*-means (HCM or k-means). FCM is faster and less prone to converge towards local minima [3].

A relevant problem which arises in CA is the interpretation of clustering results. Most of current clustering algorithms do not provide any estimate of the significance of the returned results. It is also well known that every clustering algorithm tends to produce clusters irrespectively of the data containing true clusters or not. The verification of clustering results is therefore a crucial task.

It is common practice to base the validation of results on a visual and lengthy exploration process of the data. Clearly, this procedure is highly subjective, and may be a dangerous endeavor. In particular, researchers may unwittingly overrate clusters that reinforce their own assumptions, and ignore surprising or contradictory results. For this reason, cluster validity (CV)

[*] Corresponding author. Tel.: +39 30 3715789; fax: +39 30 2091271.
E-mail address: matteo.falasconi@ing.unibs.it (M. Falasconi).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.10.001

methods have been developed to objectively and quantitatively assess the quality of the clustering results. For an overview on this topic the reader can refer to the Jain's book [1] or to more recent reviews [5,6].

CV methods have the potential to provide an analytical assessment of the amount and type of structure captured by a partitioning, and should therefore be a key tool in the interpretation of clustering results. CV techniques are named internal when they do not use additional knowledge in the form of class labels, but base their quality estimate on the information intrinsic to the data alone. Specifically, they attempt to measure how well a given partition corresponds to the natural cluster structure of the data.

A recurrent problem in cluster analysis is the selection of the model order, which means estimating the proper number of clusters. Most clustering algorithms need the number of cluster to be prespecified. Consequently, the user is confronted with the problem of selecting among different partitions resulting from running the algorithm with different number of clusters. Internal CV methods are designed to assess the partition recovery performance as a function of the number of clusters. If both the clustering algorithm employed and the internal measure are adequate for the data set under consideration, the best number of clusters can often be identified as a feature (a maximum/ minimum or a flex point) in the resulting plot of the validity index.

Traditionally, CV rules have been defined by following the intuitive notion of cluster: a group of patterns which is compact and isolated. For this reason, the corresponding objective function is often based on a suitable combination of the pooled within cluster sum of squares around cluster means with the total between cluster-centers pairwise distance [7,8].

Several such CV indices have been also proposed for fuzzy clustering [9–24]. The papers recently presented by Wang and Zhang [10] and by Bouguessa et al. [11] review a number of indices available in the literature by providing a comparison study. These studies show that traditional CV indices have intrinsic limitations when trying to estimate the number of clusters.

A fundamental problem shared by all classical CV rules is related to their bias towards the geometrical structure of the partition, especially the shape of clusters and their degree of overlap. This happens because the evaluation of compactness and separation is intrinsically based on some geometrical criterion. Thus, for example, the combination of the classical FCM method—which is known to recover spherical clusters—with any of previous CV rules can provide a wrong estimate of the number of clusters if the clusters are intrinsically not spherical. Moreover, the assumption that valid clusters should be compact and separated leads to classes with tight boundary regions, i.e. to intrinsically hard partitions, and this contrasts the benefit of using a fuzzy clustering strategy.

An attempt in this direction was undertaken by JianYu and Cui-Xia Li [25] who developed a cluster validity index based on the optimality test of the FCM objective function [26]. The method provides a criterion specifically designed for FCM that relies on the stability of the FCM outputs. In this approach the concept of "stability" of a certain partition is related to the probability with which this partition is obtained as the output of the clustering algorithm.

One of the major gaps to be filled by fuzzy validity methods is the need of better generalization capabilities without loosing of recovery performance. Ideally the validation scheme should be independent of the clustering algorithm, which also guarantees the possibility to adapt it to different data structures by choosing the most appropriate fuzzy clustering algorithm for the data and hence by ensuring an improvement of recovery ability. For this, it seems necessary a novel validity paradigm for fuzzy partitions that relaxes the constraints and uses more efficiently the information contained in the fuzzy membership function.

In this paper, we adopt an alternative approach to cluster validity based on the concept of partition stability under resampling and perturbation of data that was earlier proposed by Breckenridge [27]. Thereafter a number of stability methods have been proposed for validating crisp partitions [28–33] while theoretical basis have been provided by Ben-David et al. [34]. The goal of our paper is to propose a stability based CV index for validating fuzzy partitions, and in particular for estimating the best number of fuzzy clusters in a data set.

With respect to classical approaches, stability based CV methods employ less stringent definitions of partition quality, e.g. they do not make assumptions on the degree of overlap of the clusters, and therefore the bias effects due to the spatial clusters distribution or to the model inherent the clustering algorithm can be strongly reduced.

Stability methods rely on the hypothesis that the more stable the partition is under perturbations of the data the better, i.e. the more adherent to reality it is. These methods can be used to predict the best number of clusters in the data set. In fact, the stability of a partition varies with the number of clusters that are inferred. For instance, inferring too many clusters leads to arbitrary splits of the data, and the solution is influenced heavily by sample fluctuations. Inferring too few clusters might also provide unstable solutions, since the lack of degrees of freedom forces the algorithm to ambiguously mix structures that should be kept separate. Following these considerations, the "correct" number of clusters is that related to clustering solutions of maximum stability under perturbation of the data.

The data can be perturbed in several ways and various strategies have been proposed for evaluating partition stability. The main taxonomical subdivision is between supervised and unsupervised approaches.

In supervised schemes [28–30] the data are repeatedly split into training and a test sets (typically of equal size and with no overlap), and both sets are clustered. The partitioning on the training set is then employed to derive a classifier to predict all class labels for the test set. The disagreement between the prediction and the partitioning on the test set can then be computed using an external binary validation index (such as Folwkes–Mallows index [29]). Clearly, the classifier used for prediction has a significant impact on the performance of the method.

Unsupervised strategies [31–33] also repeatedly draw subsamples (with or without replacement) of the same data set. Such sampling procedure is repeated a number of times. Each subsample is clustered individually, and then the resulting partitions are cross compared—again by using a partitions similarity measure—to obtain an average consistency value.

In this paper we generalize to fuzzy partitions the approach proposed by Law and Jain [33]; in the following we will refer to this approach with the acronym BPSE (bootstrap partition stability estimation). BPSE belongs to unsupervised strategies and it is based on the evaluation of partition stability by repeated bootstrap sampling of the data. The method is well funded from the theoretical point of view, and provides a simple strategy for determining the optimal clustering solution. The BPSE numerical implementation is straightforward and intuitive. Finally, the method can be easily generalized for the validation of fuzzy partitions.

The validation of fuzzy clustering by means of stability methods has been recently explored by Borgelt [35]. Borgelt presented a general overview of the problem, and then studied the influence on the validity results of: (1) different ways of comparing fuzzy partitions matrices (i.e. through the direct comparison of patterns assignment or by pairwise assignments), and (2) different t-norms needed to combine the fuzzy membership degrees. For estimating the partition stability, Borgelt used a supervised resampling scheme (without replacement) based on repeated two-fold cross validation. Although the experimental study is limited to few simulated data sets and to only one experimental data set, the paper demonstrates that the stability approach is applicable to fuzzy clustering. The comparison with other fuzzy validity rules is missing.

Our work differs substantially from Borgelt's contribution and extends it in some aspects. The most important difference is the use of a different resampling scheme; there are good arguments in favor and against bootstrap sampling. The second important difference is the explicit evaluation of the variance of the clustering algorithm, if we consider this one as an estimator of the best partition over the space of all possible partitions. In this work, we just focus on the comparison of pairwise patterns assignment (*coincidence matrix*) because it is computationally less intensive than comparing patterns.[1] Finally, we carry out an accurate study for validating our proposal which includes the investigation on both artificial and experimental data sets, and its comparison with a number of traditional fuzzy validity methods.

In Section 2 we propose a possible extension of BPSE to fuzzy partitions (thereafter called fBPSE). A key step is to generalize the concept of similarity measure between two fuzzy partitions, namely between the corresponding two membership functions. For hard partitions several well-suited binary measures based on the contingency table of the pairwise assignment of data items are known (see Ref. [1, Chapter 4]); probably the best known is the Rand index [36]. Campello [37] recently proposed a fuzzy extension of the Rand index, as well as of others binary measures; thus we took advantage from this work to define and compute the fuzzy BPSE.

In Section 3 we illustrate the artificial and the experimental data sets tested in this work. In this work, we have chosen to use FCM for clustering. However, it is important to emphasize the proposed method is valid for any fuzzy clustering algorithm. Simulated data allow to study the effects of several parameters influencing FCM outcome, and hence CV performance: feature space dimensionality, true number of clusters, clusters' overlap, and fuzzifier m. Further, four well known experimental data sets available at the Machine Learning Repository [38] were analyzed; these are characterized by different degrees of complexity, i.e. a different number of features and number of true clusters.

In Section 4 we present and discuss the cluster validity results. We have compared the fBPSE with some classical CV indices selected from literature, based on concepts of cluster compactness and separation. In order to verify the benefits of using the entire information content of the fuzzy membership matrix, we also compared the generalized fuzzy version fBPSE with the corresponding hard version (hBPSE) obtained by first converting the fuzzy membership into a crisp partition.

2. Methods

2.1. Fuzzy c-means (FCM)

FCM definition is based on the minimization of the following objective function:

$$J_m(U,V)=\sum_{j=1}^{n}\sum_{i=1}^{C}u_{ij}^{m}\|\mathbf{x}_j-\mathbf{v}_i\|^2 \tag{1}$$

where n is the number of patterns, C is the number of clusters, and $\|\cdot\|$ stands for the Euclidean distance between the pattern $\mathbf{x}_j$ and the i-th cluster centre $\mathbf{v}_i$. The corresponding fuzzy membership element u_{ij} belongs to [0,1] and is subject to the constraints:

$$\sum_{j=1}^{n}u_{ij}>0,\forall i \quad \text{and} \quad \sum_{i=1}^{C}u_{ij}=1,\forall j \tag{2}$$

Under such hypotheses the fuzzy membership matrix is called "probabilistic" since u_{ij} formally resemble the (posterior) probability $p(i|\mathbf{x}_j)$ that, given $\mathbf{x}_j$, the pattern came from class i [17].

The parameter m $(m>1)$ is called the fuzzifier or weighting exponent. The exponentiation of the memberships with m leads to a generalization of the classical least squared error functional used in k-means algorithm.[2]

The minimization of Eq. (1) is performed by the alternating optimization (AO) scheme [3]. Starting with an initial selection (usually random) of cluster centroids, AO encompasses two repeated steps: (i) update of the membership matrix; (ii) update of the centroids. The updating stops when the number of iterations exceeds a maximum allowed value (typically one hundred) or when the variation in the objective function is smaller than a prefixed accuracy (ε=10^{-5}). The update formulae are derived by setting the partial derivatives of J_m equal to zero under constraints (2), this leads to the well know equations:

$$u_{ij}=\frac{\|\mathbf{x}_j-\mathbf{v}_i\|^{-2/m-1}}{\sum_{i=1}^{C}\|\mathbf{x}_j-\mathbf{v}_i\|^{-2/m-1}} \tag{3a}$$

$$\mathbf{v}_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}\mathbf{x}_j}{\sum_{j=1}^{n}u_{ij}^{m}} \tag{3b}$$

A detailed description of the classical FCM algorithm was given by Bezdek [4], while a recent overview with emphasis on FCM variants is offered by Höppner's [3]. Here we will limit to highlight some key aspects of the algorithm, namely: the effect of centers initialization, the role of the fuzzifier, and the influence of the distance measure.

It has been proven that, under the AO scheme, the FCM algorithm converges to local minima of the objective function [4]. Hence, the final partition is dependent on the random initial guesses of the clusters' centers, and simply relying on the solution of FCM on a single run might lead to unstable outcomes. For this reason the algorithm is run several times (e.g. one hundred) prior to calculating any validity index. This strategy alleviates the problem, but also requires a much longer computation time.

The weighting exponent determines the degree of fuzziness of the classification, thus significantly influencing the FCM clustering outcome, and consequently, affecting the cluster validity results. Many heuristic strategies are suggested in the literature [4,14,17–19]. A theoretical method for selecting the correct value of the weighting exponent providing applicable result was proposed by Jian Yu et al. [20]; this criterion allows setting an optimal upper bound for the m value that can be calculated *a priori* directly from the data matrix.

The distance metrics used in Eq. (1) leads to different definitions of the FCM algorithm. Euclidean distance is optimal for recovering hyperspherical clusters but it is suboptimal when clusters covariance is not negligible, e.g. in presence of high features correlation. Several variants have been proposed to improve FCM (see Ref. [3]), for example: other distances between cluster centers and data points (such as the Gustafson–Kessel

[1] The evaluation of similarity measure between partition matrices requires maximizing the similarity over all possible class permutations. The *Hungarian method* is used for solving the optimum weighted bipartite matching problem ,[28], however the time complexity still increases proportionally to C^3 being C the number of clusters.

[2] It has been proven that, for the case m=1, cluster assignments remain strictly hard when minimizing the objective function J_m, even though the membership degrees are allowed to be fuzzy [3].

algorithm based on Mahalanobis distance), fuzzy shell clustering algorithms, or kernel-based variants.

In this paper we focus our attention on cluster validity rather than on clustering strategies, therefore we use the classical FCM with Euclidean distance leaving distance optimization out of the scope of the work.

2.2. Classical validity methods

Classical CV methods [10–24] consider the optimization of an objective function based on intra-cluster compactness and inter-class separation which are defined in fuzzy sense, i.e. by taking into account the fuzzy membership function of patterns. CV rules can be classified into three main categories: the first category comprises indices that use only the pattern membership values; the second one involves both the membership values and the feature vectors, thus considering also the spatial distribution of patters; the third group covers the indices based on fuzzy hypervolume and density.

For a comparison study we selected some of the best performing indices in literature which cover the previous three categories. These indices were preferred among the different proposals because they are easy to interpret from the statistical point of view and offer a straightforward numerical implementation.

Two indices were chosen from the first category. The partition coefficient (PC) was early proposed by Bezdek [4]; here we used the modified partition coefficient (MPC) adapted by Dave [39] in order to compensate the monotonic tendency of PC to decrease when the number of clusters increases. The second index has been defined by Chen and Linkens (CHEN) [22]; it behaves much like to the original partition coefficient (PC).

Among the indices involving both the membership values and the feature vectors three were chosen: the Xie and Beni (XB) index [23] which is considered a benchmark in fuzzy clustering literature, the SC index proposed by Zahid et al. [16], and the method defined by Pakhira et al. [15]—here indicated as PBMF.

The last index, belonging to the third category, is called fuzzy hypervolume (FH) and was proposed by Gath and Geva [12].

2.3. Stability based validity method

Before introducing the generalized fuzzy version of BPSE validity method, we will shortly review the original index definition given by Law and Jain [33] that applies to hard partitions.

The BPSE procedure is based on the following key points: (1) any clustering algorithm can be regarded as a estimator for the partition of data space; (2) the bootstrap procedure can be used to estimate the stability of this estimator; (3) the stability represents an index of consistency of the cluster partition and can be used for determining the best number of clusters in the data.

Thus the BPSE algorithm for model order selection proceeds as follows:

1. The number of clusters is set to K. The starting point is usually K=2—the BPSE strategy, like almost every CV method, is not applicable for K=1—then K is incremented of one unit up to a maximum value which is typically chosen equal to the square root of the number of patterns (see Refs. [14,17]).
2. B bootstrap samples of the same size of the original data sample are generated by sampling the patterns with replacement. In our calculations B=20 was used; such relatively small number of samples was anyway sufficient for achieving a good estimation for the number of clusters, larger B values can improve the accuracy of results but also dramatically increase the computation time.
3. The fuzzy clustering algorithm is run on each bootstrap sample, as illustrated in Section 2.1, and the corresponding partition matrix is obtained.
4. The variability $V(K)$ of the partition with K clusters is evaluated by averaging the similarity values across all pairs of bootstrapped sets (see Eq. (4)).
5. The best number of clusters is found by searching for the global minimum of V across the spanned range of K.

The partition variability, which is formally equivalent to the empirical bootstrap estimate of the variance, was introduced by Law and Jain [33] for hard partitions. Here we extend this definition to fuzzy partitions as follows:

$$V(K)=\frac{1}{B(B-1)}\sum_{i=1}^{B}\sum_{j=1}^{B}[1-s(U_K(Y(i)),U_K(Y(j)))] \tag{4}$$

where $U_K(Y(i))$ stands for the membership matrix achieved by running the fuzzy clustering on the i-th bootstrap sample $Y(i)$, and $s(.,.)$ represents a similarity measure between two fuzzy partitions. By definition V is comprised in [0,1]; the partition stability is simply S=1−V, thus minimizing the variability is equivalent to obtain maximum stability.

Similarity measures for hard partitions have been introduced for the first time in early seventies by Rand [36]. The Rand index quantifies the similarity between two partitions by counting the fraction of pattern pairs which are simultaneously assigned to the same cluster or to different clusters in both partitions (see Ref. [1, Chapter 4]). Other relevant partition similarity measures are for example the Jaccard coefficient, the Fowlkes and Mallows index, Hubert index and the Gamma statistics [1].

The evaluation of the fuzzy stability index by Eq. (4) requires measuring the similarity between two fuzzy partitions. The simplest way to obtain this is to convert the fuzzy membership function into a hard membership matrix, e.g. by assigning the patterns to the maximum membership cluster, and then to use one of the binary measures mentioned above. We will call this procedure hard BPSE (hBPSE). The drawback of this procedure is that different fuzzy partitions (describing data structures with different spatial distributions) may result into the same crisp partition, and therefore, into the same value for the similarity measure. This could lead to biased variability values—since the bootstrap partitions will appear more similar than they actually are, and therefore more stable—by preventing the hBPSE to predict the correct number of clusters.

The most direct generalization of BPSE to fuzzy partitions (thereafter named fBPSE) consists of using a similarity measure appositely defined for comparing fuzzy membership functions. Such fuzzy similarity measures have been recently proposed by Campello [37].

The basic ingredients for evaluating any similarity measure are the elements of the contingency table for the two partitions. First define two sets for each individual partition:

- the set of pairs of data belonging to the *same* class in the i-th partition ($\Omega\times Sp,\ p=1,2$);
- the set of pairs of data belonging to *different* classes in the i-th partition ($\Omega\times Dp,\ p=1,2$).

Here, Ω denotes the whole set of data pairs, while S and D are two indicator functions which, for hard partitions, take only discrete values (0 or 1).

By following Campello, the individual terms of the contingency table for two hard partitions can be written in terms of the cardinality of the intersection between the previous sets:

$$\begin{aligned} a &= |\Omega \times S_1 \cap \Omega \times S_2| \\ b &= |\Omega \times S_1 \cap \Omega \times D_2| \\ c &= |\Omega \times S_2 \cap \Omega \times D_1| \\ d &= |\Omega \times D_1 \cap \Omega \times D_2| \end{aligned} \tag{5}$$

The fuzzy generalization of the indicator functions S and D is straightforward by using the definitions of fuzzy union (Max norm) and fuzzy intersection (Min norm) [40]:

$$\begin{aligned} s_{ij} &= \max\Big(\min_{c}(u_{ci}, u_{qj})\Big) \\ d_{ij} &= \max\Big(\min_{c \neq q}(u_{ci}, u_{qj})\Big) \end{aligned} \tag{6}$$

Finally, the elements of the fuzzy contingency table for the two fuzzy membership matrices can be calculated by applying the relationships (5), provided to interpret the intersection in fuzzy sense (i.e. through Min norm):

$$\begin{aligned} a &= \sum_{ij} \min(s_{1ij}, s_{2ij}) \\ b &= \sum_{ij} \min(s_{1ij}, d_{2ij}) \\ c &= \sum_{ij} \min(s_{2ij}, d_{1ij}) \\ d &= \sum_{ij} \min(d_{1ij}, d_{2ij}) \end{aligned} \tag{7}$$

Once calculated the elements of the contingency table, similarity measures can be obtained by applying the corresponding definition (see Ref. [1, Chapter 4]).

In Eq. (4) we used the corrected Rand's index. This index is normalized in order to yield values close to zero when the two partitions are taken under the random label hypothesis (Ref. [1, p. 175]). This procedure permits achieving statistically unbiased results.

The similarity measure in Eq. (4) assumes that the two partitions come from the same data set, however, the bootstrap sampling, due to its randomness, produces a different data set at each iteration. It is therefore necessary to make the bootstrap partitions homogeneous, and then comparable. A way to proceed is to consider only the fraction of patterns common to both bootstrap sets $Y(i)$ and $Y(j)$, thus we calculate the adjusted Rand's index over the sets intersection.

Different strategies are also possible. One consists of using the clustering results achieved in each bootstrap iteration to build a classifier of the missing patters, thus obtaining a set of homogeneous partitions of the original data [33]. The drawback of this approach is that the results strongly depend on the classifier; on the contrary our method does not contain free parameters and it is computationally more efficient (results not shown).

3. Data sets

3.1. Artificial data sets

For a preliminary comparison study we test the CV algorithms on simulated data composed by a variable number of independent, uncorrelated and randomly distributed Gaussian clusters. The Gaussian model is often regarded as a benchmark in literature for studying new validity rules; indeed it provides an effective (though rather simple) platform for testing the validity method in a case where the FCM clustering algorithm gives optimal performance. This permits to focus on the appropriateness of the validity method more than on the fitness of the clustering algorithm to the data. If other variables are present (cluster shape, size, orientation) the final recovery performance could depend of the implemented clustering strategy more than of the validity approach.

Simulations allow controlling the parameters influencing cluster recovery performance, such as: the feature space dimensionality D, the true number of clusters C, and the separation among clusters α (given in standard deviation units).

In our simulation study the C cluster centers are randomly drawn according to a multivariate normal distribution in D dimensions:

$$N\left(0, \frac{\alpha^2}{2D} I_{D\times D}\right)$$

Using $\alpha^2/2D$ as scaling factor of the variance, the expectation value of the square distance between any two centers is equal to α^2, independently of D. In order to control the minimum clusters separation we discard simulations where, due to the randomness of the process, any two centers are closer than $\alpha/2$. Finally, for each center, we generate 50 Gaussian distributed patterns with unit variance; with this design FCM coupled with Euclidean distance provides optimal clustering results.

We designed 36 data sets by varying three types of spatial parameters:

- increasing dimensionality (D=2, 5, 10, 15), which means increasing data sparseness;
- decreasing the degree of overlap between the clusters (α=3, 6, 9);
- increasing the number of clusters (C=4, 7, 10) so that data are more structured and then more complex;

To provide an insight into the spatial distribution of the simulated clusters we performed feature extraction and dimensionality reduction by principal component analysis (PCA). PCA score plots (Fig. 1) show the clusters distribution of the 12 data sets with C=4 in different dimensions as the parameter α spans from 3 to 9. As anticipated, it can be noted that the average spatial separation among the clusters for a given α remains constant across different D values (see e.g. the plots on the first row in Fig. 1).

If α=9 the four clusters are perfectly separated from each other, and there is not overlap. When α=6, due to "3-sigma law" for Gaussian distributions, the clusters "touch" each other only by a small amount on the borders. Finally, for α=3 the clusters overlap almost completely; in this last scenario we may expect to have few chances of recovering the clustering structure.

3.2. Experimental data sets

In order to test the performance of the cluster validity methods we use four popular data sets from the UCI Machine Learning Repository [38], namely: IRIS, WINE, IMAGE, and MFDS.

IRIS data set has 3 clusters, each of which contains 50 observations, in 4 dimensions. Two clusters are strongly overlapped and non-linearly separable, thus most clustering strategies recover only two clusters. Due to the assumptions of FCM clustering coupled with Euclidean distance, we also expect to recover two clusters at the best.

WINE data set consists of a chemical analysis of 3 wines, grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13

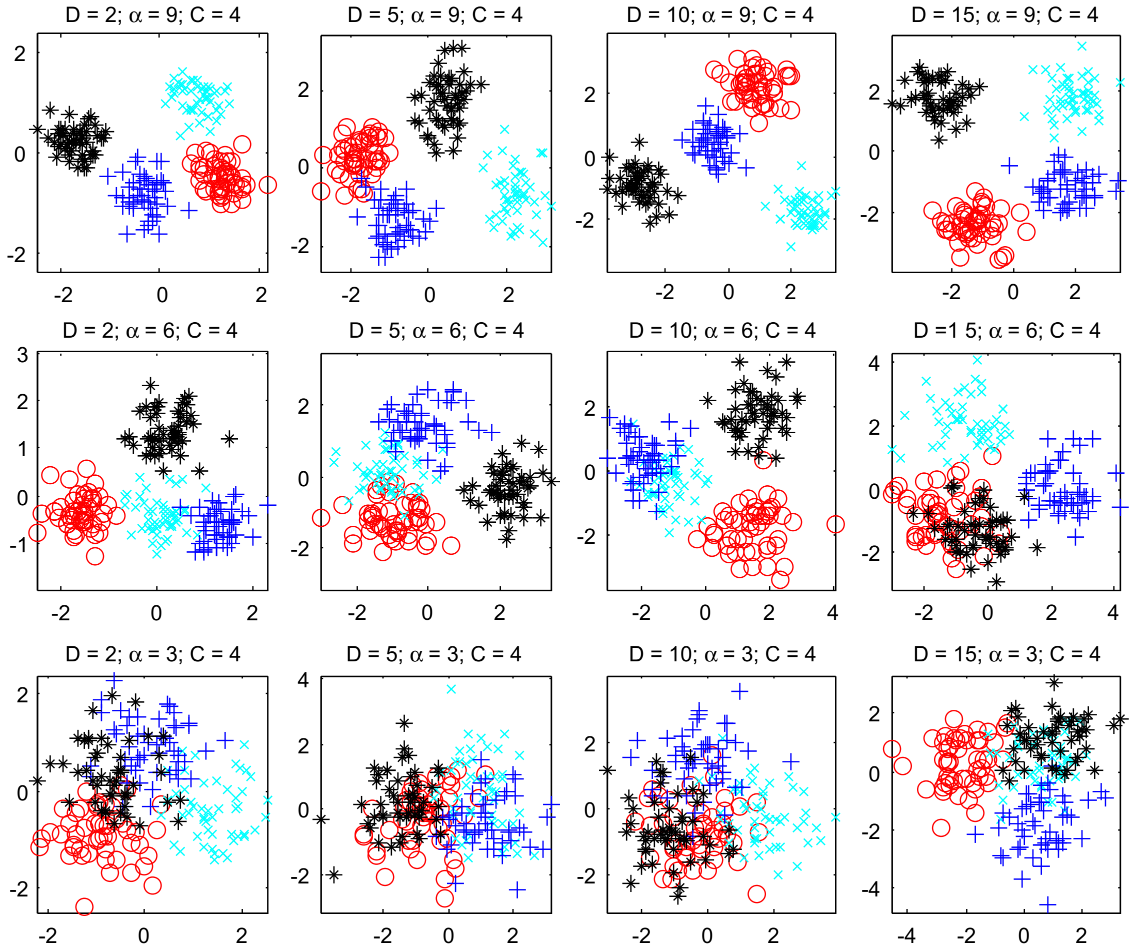

Fig. 1. PCA plots of simulated data sets with C=4 clusters for different feature space dimensions (D=2, 5, 10, 15 increasing across rows) and for increasing clusters overlap (α=9, 6, 3 decreasing across columns). The distribution of clusters clearly represents only one of the possible outcomes due to random simulation procedure.

constituents found in each of the three types of wines. The three classes are almost hyperspherical and well separated in 13 dimensions; therefore this is a good experimental data set for the testing FCM classifier.

A more difficult problem is represented by the IMAGE segmentation data set. It contains 2100 measurements of 7 different outdoor images (we used only the "test data" subset, with 300 instances per class). The images were segmented and 19 attributes were defined. We selected only 13 attributes: uninformative features (1, 3, 4, 5, 7, and 9) were not considered, as in [41]. The classes are not linearly separable.

The Multiple Features Data Set (MFDS) is very challenging. Details of this data set are available in [42]. The data set consists of handwritten numerals (0–9) extracted from a collection of Dutch utility maps. Two hundred patterns per class (for a total of 2000 patterns) are available in the form of 30 times 48 binary images. These characters are represented in terms of six sets of features; each set has a different size for a total of 649 features. Before performing cluster analysis we have reduced dimensionality in order to reduce data sparseness. We performed a ranking of the features with a paired t-test. There are 45 possible binary groups among ten classes and we selected the highest scoring 10 features in each group. Then we merged the features without considering replicates, obtaining a best set with 171 features.

Data matrices were normalized by zscore (i.e. mean centering with unit standard deviation constraint) prior evaluating the theoretical upper bound of the weighting exponent and before clustering the data.

4. Results and discussion

4.1. Artificial data sets

The optimal upper bound of the weighting exponent for simulated data sets has been calculated according to Ref. [20]. The theoretical values are reported in Table 1. We may note that—as already pointed out in [20]—the upper bound of the fuzzifier value decreases when the dimensionality increases (from D=5 to 15). We also note here that it tends to become smaller for an increasing number of clusters (from C=4 to 10), which might be related to the increasing patterns sparseness [19].

The CV results obtained on the 36 artificial data sets are reported in Table 2. The weighting exponent values were set according to Table 1.

When clusters strongly overlap (Table 2; α=3), independently of other hyper-parameters, none of the validity indices can recover the true number of clusters. In some cases C is underestimated; in other cases it is strongly overestimated. Therefore, we can compare the merits of the indices only when the clusters are perfectly separated (α=9) or slightly overlapped (α=6).

In two dimensions there is no upper bound for m (as noted in Ref. [20]) thus the FCM convergence is almost independent on m;

Table 1
Theoretical values of the upper bound of the weighting exponent, calculated according to Ref. [20], for different combinations of the spatial parameters.

α	Feature space dimension								
	***D*=5**			***D*=10**			***D*=15**		
	Number of clusters			Number of clusters *C*			Number of clusters *C*		
	C=4	**C=7**	**C=10**	**C=4**	**C=7**	**C=10**	**C=4**	**C=7**	**C=10**
9	13.0554	3.0905	2.0516	4.9066	2.3386	1.5777	2.2851	1.9997	1.3360
6	4.4861	2.3333	1.5119	2.8781	1.8657	1.6996	4.4002	1.6418	1.4336
3	3.2831	1.5113	1.4228	2.1595	1.4907	1.3009	2.0961	1.4707	1.2651

Table 2
Cluster validity results for the simulated data sets.

α	Index	Feature space dimension											
		***D*=2** True number of clusters			***D*=5**			***D*=10**			***D*=15**		
		***C*=4**	***C*=7**	***C*=10**	***C*=4**	***C*=7**	***C*=10**	***C*=4**	***C*=7**	***C*=10**	***C*=4**	***C*=7**	***C*=10**
		Weigthing exponent[a]											
		***m*=2**	***m*=2**	***m*=2**	***m*=13; 2**	***m*=3**	***m*=2**	***m*=4**	***m*=2**	***m*=1.2**	***m*=2; 1.2**	***m*=1.8**	***m*=1.2**
9	*MPC*	**4**	**7**	9	2; **4**	**7**	**10**	**4**	**7**	**10**	6; **4**	**7**	9
	CHEN[b]	**4**	**7**	2/**10**	2; 3/**4**	**7**	2/**10**	**4**	**7**	**10**	2; **4**	6	9
	XB	**4**	5	9	20; 3	**7**	**10**	3	**7**	9	6; **4**	5	9
	SC	18	20	19	20; 20	20	**10**	**4**	**7**	**10**	2; **4**	**7**	**10**
	PBMF	**4**	**7**	11	20; **4**	**6**	7	3	8	6	6; 3	5	4
	FH	**4**	**7**	11	20; **4**	**7**	**10**	20	8	20	8; 18	20	20
	hBPSE	**4**	5*	2*	3; 3	**7**	**10**	**4**	**7**	2	**4**	6**	8**
	fBPSE	**4**	**7***	**10***	**4; 4**	**7**	**10**	**4**	**7**	11	**4**	6**	12**
		***m*=2**	***m*=2**	***m*=2**	***m*=4**	***m*=2**	***m*=1.2**	***m*=2.5**	***m*=1.8**	***m*=1.2**	***m*=4**	***m*=1.5**	***m*=1.2**
6	*MPC*	**4**	**7**	9	3	5	**10**	17	**7**	10	**4**	6	**10**
	CHEN	3/**4**	2/**7**	9/**10**	3	2/**7**	2/**10**	20	**7**	10	**4**	6	**10**
	XB	3	5	8	2	**7**	**10**	2	4	10	**4**	4	9
	SC	20	20	19	5	**7**	**10**	2	5	2	**4**	4	2
	PBMF	**4**	8	9	5	6	9	20	5	6	3	4	5
	FH	**4**	**7**	11	2	**7**	**10**	2	20	19	20	20	20
	hBPSE	**4**	5*	5*	**3**	**7**	2	2	8	**10**	**4**	4	**10**
	fBPSE	**4**	**7***	**10***	**4**	**7**	11	**4**	**7**	**10**	**4**	**7**	**10**
		***m*=2**	***m*=2**	***m*=2**	***m*=3**	***m*=1.5**	***m*=1.2**	***m*=2**	***m*=1.4**	***m*=1.2**	***m*=2**	***m*=1.4**	***m*=1.2**
3	*MPC*	3	3	4	2	2	2	2	2	3	2	2	2
	CHEN	2	2	2	2	2	2	2	2	2	2	2	2
	XB	3	12	6	2	2	2	2	2	2	2	16	2
	SC	19	20	20	2	2	2	2	2	2	2	2	2
	PBMF	19	11	6	10	10	18	16	19	16	19	15	12
	FH	19	17	4	2	2	2	2	2	2	2	2	2
	hBPSE	2	3	2	2	2	2	16	4	7	18	16	9
	fBPSE	18	20	20	2	2	2	18	20	20	13	17	20

Shaded columns indicate the data sets for which replicated analyses have been performed (see Table 3). Bold numbers mean match between true and predicted number of clusters. The cases marked with an asterisk refer to the plots in Fig. 2, with a double asterisk in Fig. 3.

[a] Except than in two dimensions where m was set equal to the "classical" value, the fuzzifier value was set lower but close to the theoretical upper bound value reported in Table 1. In two cases (*D*=5; *C*=4 and *D*=15; *C*=4) we have reported the results for two very different values of *m* (separated by semicolon) in order to show the different performance of the indices.

[b] Where CHEN index presents two predicted values the left one refers to the original definition while the right one to the compensated index. Where only one value is given the two predictions coincide.

in this case the weighting exponent was set equal to the "classical" value $m=2$ (Table 2). The cluster recovery results were found to be rather independent of the fuzzifier value (data not shown). The only exception is represented by the SC index which shows very poor performance for $D=2$ and also inconsistent results throughout Table 2. SC shows a critical behavior with increasing dimensionality and also with particular combinations of hyper-parameters (for example it depends on how the feature space dimensionality competes with clusters' overlap). Its peculiar behavior will be explained later on.

The simplest two-dimensional case permits to better understand the different behavior of stability indices. The fBPSE clearly

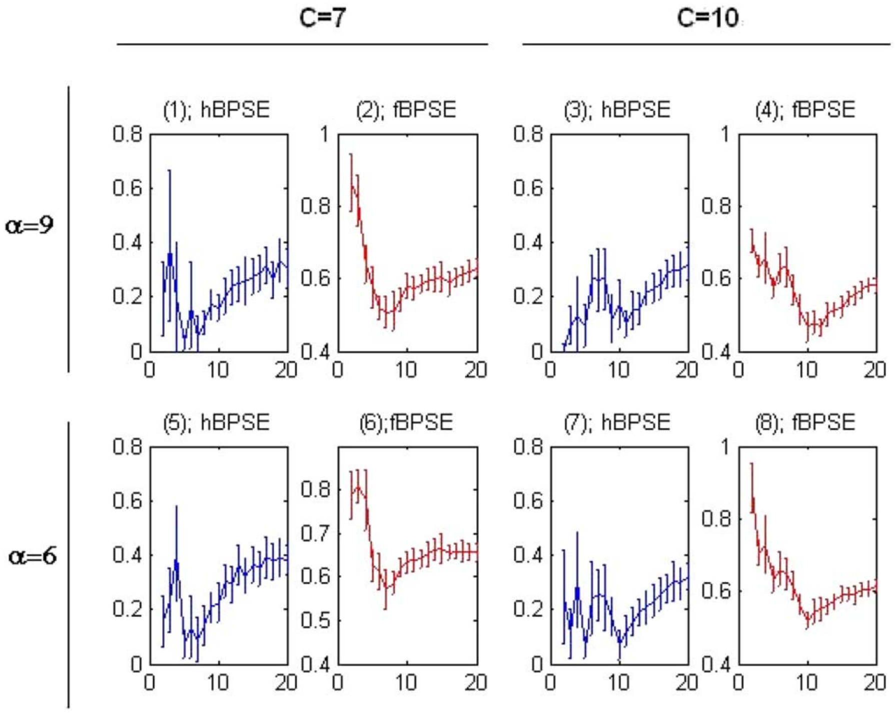

Fig. 2. Variability curves of hBPSE and fBPSE versus the number of clusters for four data sets marked in Table 2: D=2, α=9, C=7 (1); D=2, α=9, C=10 (2); D=2, α=6, C=7 (3); D=2, α=6, C=10 (4). Error bars represent the standard deviation evaluated over the 20 bootstrap iterations.

outperforms hBPSE which in several cases estimates too low C values. We have focused on four situations (marked with a single asterisk in Table 2) for which we plotted variability values versus number of clusters (Fig. 2).

The fBPSE shows, for every case, a global minimum in correspondence to the true number of clusters. On the contrary, the global minimum of the hBPSE curve often does not coincide with the true C, although the index shows always a local minimum at that value, but appears systematically biased towards smaller numbers (see plots (1)–(3)–(5)–(7) in Fig. 2). A direct comparison of hBPSE and fBPSE (e.g. plots (3) and (4) in Fig. 2) shows that, while fBPSE takes its maximum at C=2 and then decreases until C=10, hBPSE increases first (up to C=5) and then decreases.

Since this occurs only for hBPSE, we argue that it is due to the conversion of the membership function into a hard index. Actually, for patterns belonging to core areas of the clusters, there is not much difference between hard and fuzzy similarity measures across bootstrap partitions. On the contrary, the assignment of patterns belonging to the boundary regions of clusters appears to be more stable by taking the hard membership than the fuzzy membership matrix. Indeed the fuzzy membership of border patterns is subject to a higher variability because it takes into account the actual membership values whereas the class assignment often remains unchanged. Therefore, the reduction to a hard membership, by forcing the patterns to belong to only one class, causes a loss of resolution when calculating bootstrap variability by Eq. (4). Evidence of this phenomenon can be quantitatively obtained by comparing the four terms a, b, c, d in the Rand's index definition[3] (data not shown). By considering the two cases reported in Fig. 2(3) and (4) and by looking at the partition C=2 for example, we observed that for hBPSE the terms b and c are negligible with respect to the terms a and d, thus leading to a Rand's similarity value close to one. Conversely for fBPSE the four terms are almost comparable in magnitude, therefore meaning a lower similarity value among the bootstrap partitions and providing a higher variability. Since b and c terms are associated with the misclassified pattern pairs they basically accounts for the contribution of the patterns on the clusters border for which the variations of membership values are higher over the bootstrap samples.

In higher dimensions ($D \geq 5$) it is very important to set a proper value of the weighting exponent. In particular, there were two cases, i.e.: D=5 with C=4, and D=15 with C=4, in which we found that smaller values of the weighting exponent led to much better validity results. We argue that, although the fuzzifier is optimized to guarantee FCM convergence, its actual value might play a crucial role in the recovery performance of the cluster validity indices. Indeed, there could be a direct or indirect dependence of recovery performance of the index on the fuzzifier: a direct dependence can exist when the weighting exponent is included in the index's definition (such as for XB index [23]), an indirect connection can be due to the unavoidable use of the membership matrix which in turns depends of the fuzzifier itself.

In the two mentioned cases, it can be noted that the fBPSE can predict the true number of clusters even for larger values of the weighting exponent. This indicates that, with respect to classical indices, the new validity rule is more robust towards variations of the fuzzifier, allowing a better generalization capability.

As we already pointed out, due the randomness of clusters generation process, CV results might depend on the realized distribution of clusters. A small mismatch between the predicted and the true value of C, on the order of one unit is acceptable. To further investigate the effect of the randomness , in four specific cases (shaded columns in Table 2) characterized by different clusters overlap and different feature space dimension, the cluster simulation and then the CV procedure have been replicated 20 times (Table 3).

[3] The comparison among the four terms is performed by taking the average over the bootstrap partitions.

Table 3
Cluster validity results achieved on four data sets characterized by different combinations of the spatial parameters over 20 simulation trials (the data generation procedure is repeated in each trial).

Index	Predicted number of clusters																			
	1	*2*	*3*	*4*	*5*	*6*	***7***	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
(a) α=9, *C*=7, *D*=5, *m*=3																				
MPC	–	0	1	1	0	4	**14**	0	0	0	0	0	0	0	0	0	0	0	0	0
CHEN	–	**11**	1	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0
XB	–	0	0	2	4	**7**	**7**	0	0	0	0	0	0	0	0	0	0	0	0	0
SC	–	0	0	0	0	0	**0**	0	0	0	0	0	0	1	1	2	1	5	4	**6**
PBMF	–	0	0	0	2	5	**13**	0	0	0	0	0	0	0	0	0	0	0	0	0
FH	–	0	0	0	0	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0
hBPSE	–	3	1	0	0	0	**15**	1	0	0	0	0	0	0	0	0	0	0	0	0
fBPSE	–	0	0	0	0	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0
(b) α=9, *C*=7, *D*=15, *m*=1.8																				
Index	*1*	*2*	*3*	*4*	*5*	*6*	***7***	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
MPC	–	0	0	1	0	2	**16**	1	0	0	0	0	0	0	0	0	0	0	0	0
CHEN	–	1	0	1	0	2	**15**	1	0	0	0	0	0	0	0	0	0	0	0	0
XB	–	0	1	1	1	6	**11**	0	0	0	0	0	0	0	0	0	0	0	0	0
SC	–	0	0	0	0	2	**15**	1	1	0	0	1	0	0	0	0	0	0	0	0
PBMF	–	1	**6**	4	**6**	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FH	–	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	2	**18**
hBPSE	–	2	1	0	0	1	**16**	0	0	0	0	0	0	0	0	0	0	0	0	0
fBPSE	–	0	0	0	0	0	**19**	1	0	0	0	0	0	0	0	0	0	0	0	0
(c) α=6, *C*=7, *D*=5, *m*=2																				
Index	*1*	*2*	*3*	*4*	*5*	*6*	***7***	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
MPC	–	2	1	0	3	4	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0
CHEN	–	**17**	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0
XB	–	0	0	3	3	**6**	**8**	0	0	0	0	0	0	0	0	0	0	0	0	0
SC	–	0	0	0	0	4	**14**	2	0	0	0	0	0	0	0	0	0	0	0	0
PBMF	–	0	0	1	3	5	**11**	0	0	0	0	0	0	0	0	0	0	0	0	0
FH	–	0	0	0	0	2	**18**	0	0	0	0	0	0	0	0	0	0	0	0	0
hBPSE	–	4	1	0	0	0	**14**	1	0	0	0	0	0	0	0	0	0	0	0	0
fBPSE	–	0	0	0	0	0	**18**	1	1	0	0	0	0	0	0	0	0	0	0	0
(d) α=6, *C*=7, *D*=15, *m*=1.5																				
Index	*1*	*2*	*3*	*4*	*5*	*6*	***7***	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
MPC	–	1	0	0	1	6	**11**	1	0	0	0	0	0	0	0	0	0	0	0	0
CHEN	–	7	1	0	0	0	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0
XB	–	0	2	3	**6**	**6**	**2**	0	1	0	0	0	0	0	0	0	0	0	0	0
SC	–	**8**	4	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
PBMF	–	1	**4**	**5**	**5**	2	2	0	1	0	0	0	0	0	0	0	0	0	0	0
FH	–	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	20
hBPSE	–	5	0	1	1	0	**13**	0	0	0	0	0	0	0	0	0	0	0	0	0
fBPSE	–	0	0	0	1	0	**19**	0	0	0	0	0	0	0	0	0	0	0	0	0

Column with the true number of clusters is shaded, while the estimated number of cluster (most recovered) is bold.

In spite of its simplicity, MPC shows good recovery ability even for partially overlapping clusters and across different features sets and different numbers of clusters present in the data set. Replication analysis shows a reduction of the performance in case of higher overlap (Table 3(c) and (d)), while the cluster recovery ability seems independent of the dimensionality. Our findings are consistent with previous comparison studies which report good performance of this index on data sets formed by standard normal classes [11,14].

CHEN correctly recovers the number of clusters in high dimensions, while in low dimensions it tends to strongly underestimate the number of predicted clusters (in Table 2 see e.g. the case D=2, C=10 while the predicted value is 2). Replicated simulations also confirm this result: in five dimensions CHEN most often recovers two clusters (see Table 3(a) and (c)) while in high dimensions the most recovered value coincides with the true number of clusters (Table 3(b) and (d)). This is a consequence of the monotonic decreasing trend of the index with the number of clusters, hence in many cases the index global maximum occurs at C=2 while there is only a local maximum at the true value. This problem can be solved by taking into account also local maxima, although this approach can be hardly applied for automated clusters identification. A second strategy could be to modify the CHEN index by introducing a punishing term for low C values, or by compensating the index as proposed by Dave for PC index [39]. For comparison, along with the results of the original index, we have reported also the C values predicted by the compensated index (Table 2). Indeed, the results of compensated CHEN index are much better and similar to those obtained by MPC.

XB is regarded as a benchmark in the field of fuzzy clustering validation. We found that XB performs a little worse than MPC, corrected CHEN and fBPSE. It often estimates a lower number of clusters (in many cases two units less than the true value). Exploring the XB curves against the number of clusters (data not shown), we observed in many cases the presence of multiple minima, where often the global minimum does not correspond to the true number of clusters. XB values are mostly determined by the minimum squared Euclidean distance between clusters centers (definition [23]), which is dependent on the spatial distribution of centroids. For this reason the predictions made by XB index are strongly affected by the specific distribution of clusters achieved in the simulation trial. This explains why over replicated simulations the XB index provides very spread predicted values (Table 3).

SC provides scattered results. Especially in low dimensions, it gives wrong estimates even for the largest cluster separation. Xu and Brereton [14] have already pointed out one problem with SC: the two terms occurring in its definition [16], namely the geometrical term SC1 and the membership term SC2, have different scales depending on D. In very low dimensionality (D=2), SC1—which measures the geometrical ratio between clusters' separation and compactness—assumes much larger values than SC2—which provides the same ratio but only in terms of fuzzy membership elements. Therefore, globally, SC is monotonically increasing with C, leading to large predicted C values. When D gets bigger, SC2 starts to be comparable to SC1, and terms can compensate, providing a correct estimate of C (Table 2: α=9, D=10 and 15; Table 3(b)). However, if clusters separation diminishes, SC1 becomes smaller and SC2 can dominate, leading now to a monotonic decreasing trend of SC that will underestimate the number of clusters (Table 3(d)). By diminishing clusters separation, the two terms can compensate in lower dimensionality; this explains why, for example, in five dimensions SC correctly predicts the number of clusters for slightly overlapped classes but not for perfectly separated groups (compare Table 3(a) and (b)).

The drawback of PBMF is similar to that of XB, i.e. the wrong estimated values in Table 2 are related to the presence of multiple maxima. The PBMF index definition contains the maximum value of the squared Euclidean distance among the clusters centers [15]. By replicating simulations, we observe that PBMF performs better in low dimensions (Table 3(a) and (c)). In high dimensions the predicted values are much more spread and often the number of clusters is underestimated.

FH index provides a correct estimate for the number of clusters only for low dimensional data sets (Table 2: D=2 and 5) while it fails in higher dimensions. Over repeated simulations, FH shows good recovery performance in five dimensions (Table 3(a) and (c)) while in 15 dimensions it returns always the maximum explored number of clusters.

FH behavior can be understood by going back to its definition. FH is defined as the sum (over all clusters) of the fuzzy covariance matrix determinant of each cluster, which is proportional to the cluster volume in feature space. A good partition should yield compact classes and then a low FH. The major difficulty lies in the calculation of the fuzzy covariance matrix in high dimensions, since some variables might be correlated to each other, making the covariance matrix of each cluster singular, and therefore, the corresponding determinant equal to zero. In this scenario FH index will always be equal- or very close-to zero and hence useless. By increasing the m value the clusters become softer, membership elements are higher on average, and the determinant of the fuzzy covariance matrix gets bigger allowing to compensate the shrinking due to dimensionality. This compensation works only for moderate dimensionality values, whilst in high dimensions the index has a monotonic increasing trend with the number of clusters.

Finally, to understand how the fuzzifier value affects the trend of fBPSE and hBPSE four data sets have been selected from Table 2 (marked with a double asterisk) and the corresponding variability plots are reported in Fig. 3.

When the weighting exponent is not optimized the variability values against the number of clusters do not provide any useful information for determining the optimal C value. The fBPSE curve evidences constant values (Fig. 3, plots (2) and (6)), while for hBPSE the differences among values are not statistically significant when we take into account the error bars (Fig. 3, plots (1) and (5)). This occurs because the partition fuzziness is too high and the membership function elements are almost identical across all clusters—each pattern is equally assigned to every cluster—leading to constant fuzzy similarity across all C values. The procedure of converting the fuzzy membership into a hard membership forces the patterns to belong to only one cluster, causing small fluctuations of the partition similarity values, and insignificant (comprised within the error bars) hBPSE differences.

With a correct value of the fuzzifier, the FCM converges to the proper membership function. The clustering becomes harder, and

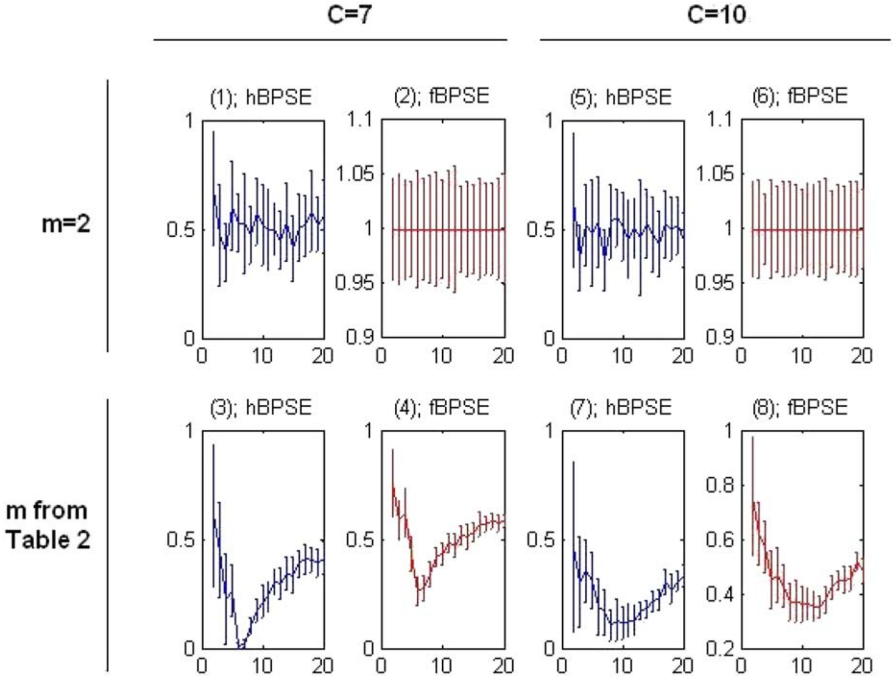

Fig. 3. Behavior of hBPSE and fBPSE versus the number of clusters for two selected data sets with D=15, a=9, and C=7 (plots 1–4) or C=10 (plots 5–8) for different values of m: (upper plots) m is not optimized, being above the theoretical upper bound; (lower plots) m is correctly set to the values in Table 2. Error bars represent the standard deviation of the variability over the 20 bootstrap iterations.

then, the stability index starts to become more sensitive to the partitions changes. The variability curves show a clear global minimum; this is quite sharp for $C=7$ (Fig. 3, plots (3) and (4)) and becomes smoother for $C=10$ (Fig. 3, plots (7) and (8)). This is reasonable if we imagine that the higher is the number of clusters the more difficult is to identify every individual cluster.

The fuzzy stability method fBPSE gives globally the best recovery performance among the tested indices being very robust towards changes of the spatial clustering parameters and to variations of the weighting exponent.

4.2. *Experimental data sets*

In the previous section we have considered simulated data sets that include only standard normal clusters, i.e. each class has unitary covariance and then clusters are hyperspherical in the feature space. In this scenario, if the number of clusters matches the true one and provided that the fuzzifier value is correctly chosen, the standard FCM objective function (with Euclidean distance) should converge to the best partition.

Experimental data are generally more complex. Features are usually correlated to each other leading to the presence of clusters with different shapes and different spatial orientation. For this reason we should expect a mismatch between the predicted structure and the true one which is intrinsically related to the use of squared Euclidean distance in Eq. (1). Indeed, by using Euclidean metrics, we can only find a subdivision of data in a number of hyperspherical clusters, and clearly, this subdivision can be very different from the intrinsic data structure.

Therefore, the distance function is another hyper-parameter, such as the fuzzifier value, that should be optimized before proceeding with cluster validity. Distance optimization can be laborious and prone to overfitting; the optimal distance can also depend of many other choices such as the selected features and the type of kernel used in the FCM objective function. Hence, we decided not to address distance optimization in this work but to use the standard FCM objective function.

Nevertheless, we might ask whether the Euclidean distance is appropriate for clustering the experimental data sets. This can be judged by comparing across the entire range of partitions the cluster labels with the true data labels (also known as *external validity*). The similarity is evaluated in this case by means of Jaccard index (Ref. [1, Chapter 4]), i.e. by counting only the fraction of patterns pairs that are correctly assigned in both partitions. The index spans from zero to one, where values closer to one mean higher similarity between the two partitions. The external validity results are displayed in Fig. 4. For each data set the theoretical

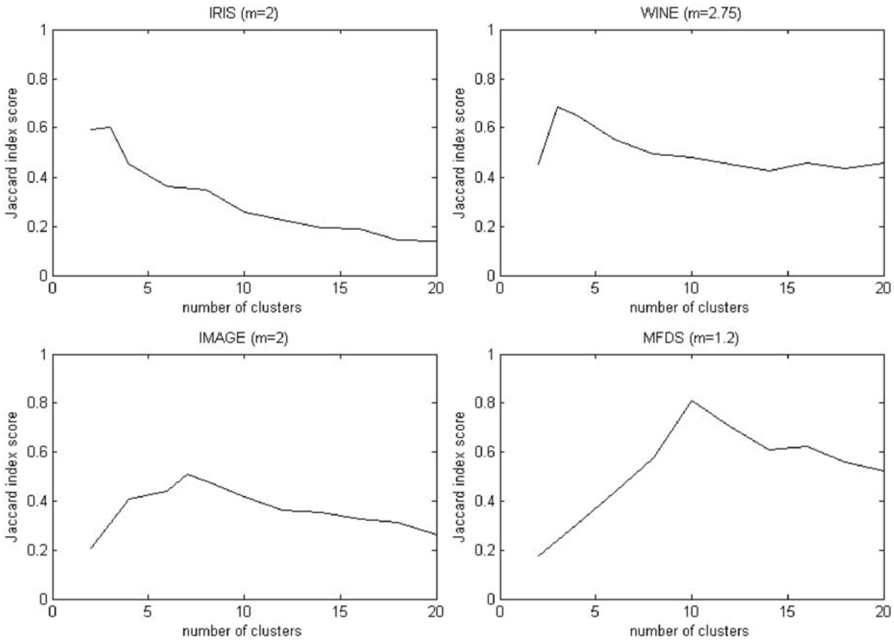

Fig. 4. Results of external validity achieved on the experimental data sets. The plots report the Jaccard similarity value between the true data labels and the partition labels (partitions were obtained by FCM algorithm with Euclidean distance). The maximum of each curve corresponds to the best matching.

Table 4
Cluster validity results for the experimental data sets.

Data set	Feature space D	λ max[a]	m upper bound[a]	m selected	True number of clusters	Predicted number of clusters							
						MPC	*CHEN*	*XB*	*SC*	*PBMF*	*FH*	*hBPSE*	*fBPSE*
WINE	13	0.3581	3.52	2.75	3	**3**	**3**	**3**	**3**	**3**	n.d.	**3**	**3**
IRIS	4	0.6652	$+\infty$	2	2–3	**2**	**2**	**2**	n.d.[b]	**3**	n.d.	**2**	**2**
MFDS	171[c]	0.2174	1.77	1.2	10	**10**	2	6	2	3	2	3	**10**
IMAGE	13[c]	0.4979	$+\infty$	2	7	3	2	3	n.d.	3	4	4	**6**

[a] Calculated according to Ref. [20] after zscore normalization of data.
[b] n.d.=not determined.
[c] After feature selection.

upper bound of the weighting exponent was calculated, and then the m value was set in agreement with that (see Table 4). When there was no superior limit, the traditional value $m=2$ was selected.

WINE, which comprises three nearly hyperspherical classes, shows the maximum Jaccard index value (≈ 0.7) at $C=3$. If the m value was higher than 2.75, we experimentally observed that the Jaccard index became almost constant across all C values, therefore we selected that value for evaluating the recovery performance of the validity rules.

MFDS data, for which the ten classes are almost Gaussian shaped, show also high values of Jaccard index. At $C=10$ the 80 percent of patterns is correctly classified, provided that the fuzzifier was set equal to $m=1.2$.

For IRIS and IMAGE, the low values of Jaccard index—whatever is the number of clusters—indicate that there is a large mismatch between the true labels and the clustering labels. This happens because these data sets contain non-linearly separable classes, therefore the Euclidean distance leads to highly suboptimal partitions in terms of patterns assignment. Nevertheless, there is not an evident mismatch between the true number of clusters and the number that maximizes the Jaccard index. For IRIS data there is not significant difference between $C=2$ or $C=3$ clusters (Jaccard is about 0.6 in both cases); for IMAGE the best match between cluster labels and true labels occurs just at $C=7$. Therefore, at least with regard to the number of clusters, Euclidean distance provides the correct values.

The predicted values of the number of clusters for the experimental data sets are summarized in Table 4. The individual trends of some selected indices against the number of clusters are displayed in Fig. 5; for space constraints only the best indices were reported.

WINE is the simplest case since the true clusters are approximately Gaussian. In fact almost all the indices, except FH, provide the correct answer. FH has a shallow local minimum at three clusters (plot not shown) but then its monotonic decreasing trend with increasing the number of clusters dominates by making impossible to predict the C value. FH failure could be due to the high dimensionality or, most probably, to the features correlation which plays a crucial role in the calculation of fuzzy covariance matrix. This also gives reason for the small values of the index (<0.01). FH index shows the same behavior also for the other data sets.

Some indices, e.g. XB (see Fig. 5) as well as CHEN and SC, do not show a remarkable difference between the values at $C=2$ and 3. The remaining indices, MPC, PBMF, and fBPSE, show similar trends against the number of clusters and have a rather deep minimum at $C=3$ (Fig. 5). The hBPSE behaves similarly to fBPSE.

For IRIS data we assume it is equivalent to predict two or three clusters. Indeed five methods out of eight predict two clusters, while PBMF is the only index predicting three clusters (Table 4).

SC and FH do not work. The problem with SC index is connected with the low feature space dimensionality ($D=4$). Therefore, as already remarked with simulated data, the geometrical term SC1 dominates over the membership term SC2 and the index assumes a monotonic increasing trend with the number of clusters and it predicts a large number of clusters.

Concerning PBMF, together with the global maximum at $C=3$ there is a comparably large local maximum at $C=9$ (Fig. 5). To explain this we must consider that the index values across different C numbers critically depend on how the various factors occurring in its definition compete with each other (see Ref. [15]). The original definition considers the first factor equal to $1/C^2$, however the results on IRIS data dramatically change if we take, for example, this factor equal to $1/C$ or instead to $1/C^3$: in the former case PBMF has a monotonic increasing trend, since the linear decay of the factor cannot compensate the power law

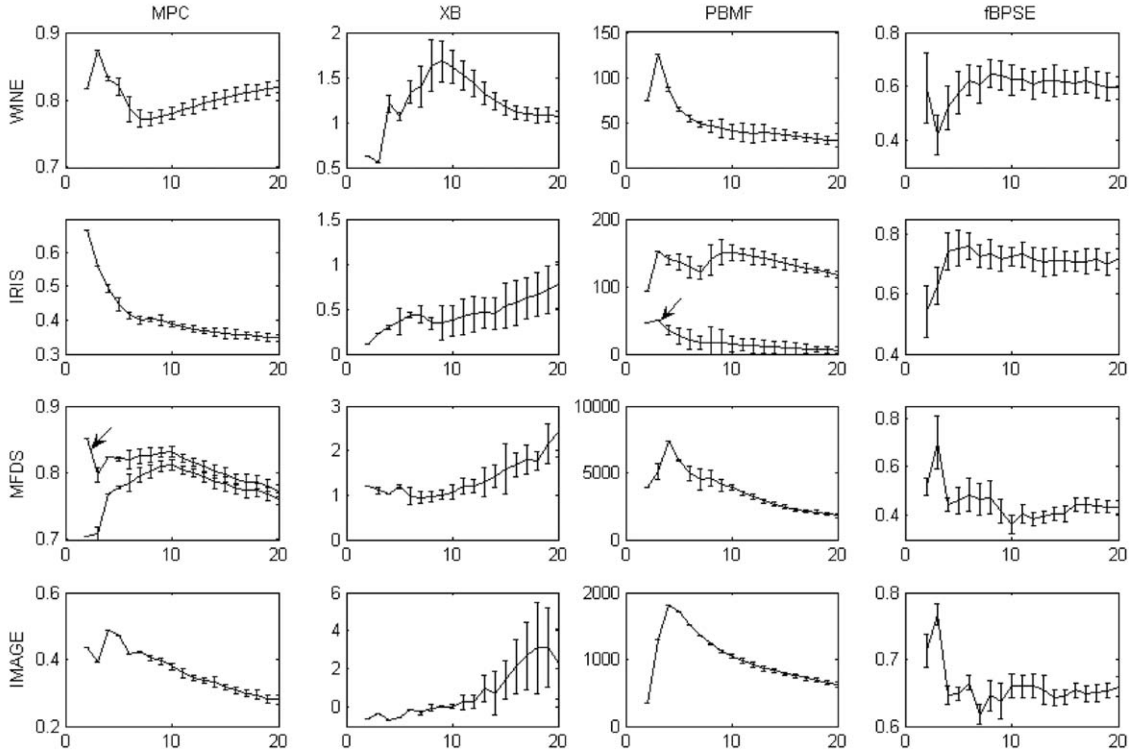

Fig. 5. Trends of four selected indices (MPC, XB, PBMF, and fBPSE) against the number of clusters for the experimental data sets. For the classical indices standard deviations are evaluated over the FCM algorithm repetitions, for fBPSE error bars are evaluated over the bootstrap iterations.

increase of the maximum separation between the pairs of centers; in the latter case (see the curve marked with an arrow in Fig. 5) the trend is monotonically descending and PBMF predicts only two clusters.

For more complex data sets, such MFDS and IMAGE, the fuzzy stability method fBPSE outperforms both the classical CV techniques and the hBPSE version.

With regard to MFDS, both fBPSE and MPC recover the correct partition (Table 4). The MPC shows a well defined maximum value at $C=10$ (Fig. 5), however this nice behavior comes mainly from the compensation suggested by Dave for the index which considers a specific rational function of C (see Ref. [39]). In fact the original partition coefficient (see the MPC curve marked with an arrow in Fig. 5) does not work well, having a global maximum at $C=2$ and only a local peak for 10 clusters. This is basically the same effect that we have observed with PBMF, i.e. the recovery ability depends of specific parameterizations of the index with factors scaling with the number of clusters.

The superior performance of fBPSE method is demonstrated on the IMAGE data set. The fBPSE index predicts six clusters (Table 4), which is close to the true value, while none of the other indices is able to recover this partition.

4.3. Discussion

According to Buhmann [28], due to the lack of a general, data-independent formal objective in clustering, the question of identifying an appropriate cluster algorithm is ill posed. In line with all previous studies [28–33], we have limit ourselves to the problem of finding the correct model order, i.e. the number of clusters, given a fixed clustering algorithm (FCM). Obviously, this work could have been replicated using a different fuzzy clustering algorithm. It is our intuition, that the conclusions do not depend critically in this assumption, as long as the chosen algorithm is suited to the underlying structure of the data.

We also acknowledge the fact that there might be more than one useful answer to this second question. Yet, an indispensable requirement is the robustness of the clustering solution, that means, the result should be reproducible on other data sets drawn from the same source which the fundamental assumption of the stability based approach.

We showed that the combination of FCM with fBPSE has equal (or even better) performances than FCM with any of several other tested validity indices. This holds both for synthetic Gaussian-shaped datasets and for non-Gaussian experimental datasets, thus suggesting that the proposed validity method is quite generally applicable.

Taking such point of view, the main advantage of fBPSE compared to traditional indices lies on its generalization capability. Classical indices are based on a specific analytical function of the pattern memberships and spatial coordinates which may work or not depending on the data set; conversely the stability method depends only on the data, through the resampling process, and then is more flexible.

We also qualitatively saw that the extreme point in the graph reporting the fBPSE index versus the number of clusters can be peculiar of the data structure captured by FCM: (i) can be shallow (as e.g. for the IMAGE dataset in Fig. 4); (ii) can have a relatively small absolute value, e.g. the IMAGE and the IRIS datasets in Fig. 4 have a value of circa 0.6 on a 0 to 1 scale (while the other two datasets have a value of circa 0.8). Any of these two phenomena can be considered to a posteriori flag a weak confidence in the optimal cluster number. In particular, the low absolute value of the stability index means that even the best derived cluster structure is not very stable. This does not mean necessary that fBPSE per se is not applicable but rather than FCM is not the best clustering algorithm whilst it could well be that another clustering algorithm together with fBPSE gives a clear extreme value.

5. Conclusions

The main contribution of this work to the field of cluster analysis is a novel approach to fuzzy clustering validity based on partition stability. This approach—called fuzzy BPSE (fBPSE)—is based on the measure of partition stability under perturbation of the data by bootstrapping patterns. Through an accurate study, on both artificial and experimental data sets, we showed that by using bootstrap and by estimating the partition variance, the new method is particularly robust.

The index behavior has been studied, including the influence of the weighting exponent, and compared with some classical CV methods based on compactness-to-separation criteria. Our findings suggest that the fBPSE is well suited for validating fuzzy partitions, and in many cases outperforms existing validity techniques. Additionally, since the method is designed for fuzzy clustering it will show advantages for overlapping classes or naturally occurring partial membership on the object data.

More efficiently use of the information contained in the fuzzy membership function has also been proven by comparing the fBPSE with the corresponding hard index version (hBPSE). The increased computational complexity of fBPSE as compared to hBPSE might be a drawback for large data sets. However, our results demonstrate that the use of the entire fuzzy membership matrix for evaluating the fuzzy partition stability improves performance, e.g. by removing the bias towards low number of clusters typical of hard BPSE.

Of course, stability-based techniques, as any other validation technique, may also be misleading, e.g. for data sets in which the clusters shape cause the clustering algorithm to converge reliably to suboptimal solutions. This is the case for example of FCM with Gaussian metric applied to a data set characterized by non spherical clusters; however, this is due more to the intrinsic poor fitting ability of the adopted clustering model than to validity index limitations. Even then, the fBPSE index may provide a starting point for further refinements (e.g. a range of K values around which to explore the data by adopting a clustering strategy that better fits the structure of the clusters). Of course this is feasible since the principle underlying our proposed method is quite general, and therefore, it is applicable in combination with any clustering strategy. We think that this paper can be the seed for further investigations, for example the study of recovery properties of fBPSE in combination with different metrics (other than Euclidean) or clustering strategies (other than FCM).

Acknowledgments

This work was financially supported by the FP6-EU Network of Excellence GOSPEL “General Olfaction and Sensing Projects on a European Level” (IST-2002-507610) and by the “Short Mobility Program 2007” of the National Research Council (CNR), Roma, Italy. We would like to thank the comments of the anonymous reviewers that helped to improve the quality of the manuscript.

References

[1] A. Jain, R. Dubes, Algorithms for Clustering Data, ed, Prentice Hall, Englewood Cliffs, New Jersey, 1988.

[2] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Surveys 31 (1999) 264–323.

[3] F. Höppner F. Klawonn, R. Kruse, T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition, ed, John Wiley & Sons Ltd., WileyNew York, 1999.

[4] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, edition, Plenum Press, New York, 1981.

[5] J. Handl, J. Knowles, D.B. Kell, Computational cluster validation in post-genomic data analysis, Bioinformatics (Review) 21 (2005) 3201–3212.

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering validity methods, ACM SIGMOD Record, 31 2 (2002) 40–45;
M. Halkidi, Y. Batistakis, M. Vazirgiannis, Clustering validity methods, ACM SIGMOD Record, 31 3 (2002) 19–27.

[7] G.W. Milligan, M.C. Cooper, An examination on the procedures for determining the number of clusters in a data set, Psychometrica 50 (1985) 159–179.

[8] M. Falasconi, M. Pardo, M. Vezzoli, G. Sberveglieri, Cluster validation for electronic nose data, Sensors and Actuators B: Chemical 125 (2007) 596–606.

[9] A. Celikyilmaz, I.B. Türkşen, Validation criteria for enhanced fuzzy clustering, Pattern Recognition Letters 29 (2008) 97–108.

[10] W. Wang, Y. Zhang, On fuzzy cluster validity indices, Fuzzy Sets and Systems 158 (2007) 2095–2117.

[11] M. Bouguessa, S. Wang, H. Sun, An objective approach to cluster validation, Pattern Recognition Letters 27 (2006) 1419–1430.

[12] I. Gath, A.B. Geva, Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (1989) 773–781.

[13] R.J.G.B. Campello, E.R. Hruschka, A fuzzy extension of the silhouette width criterion for cluster analysis, Fuzzy Sets and Systems 157 (2006) 2858–2875.

[14] Y. Xu, R.G. Brereton, A comparative study of cluster validation indices applied to genotyping data, Chemometrics and Intelligent Laboratory Systems 78 (2005) 30–40.

[15] M.K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern Recognition 37 (2004) 481–501.

[16] N. Zahid, M. Limouri, A. Essaid, A new cluster-validity for fuzzy clustering, Pattern Recognition 32 (1999) 1089–1097.

[17] N.R. Pal, J.C. Bezdek, On cluster validity for the fuzzy *c*-means model, IEEE Transactions on Fuzzy Systems 3 (1995) 370–379.

[18] D. Dembéle, P. Kastner, Fuzzy *c*-means method for clustering microarray data, Bioinformatics 19 (2003) 973–980.

[19] C. Döring, C. Borgelt, R. Kruse, Effects of irrelevant attributes in fuzzy clustering, in: Proceedings of the 14th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Reno, Nevada, USA, May 22–25, 2005, pp. 862–866.

[20] J. Yu, Q. Cheng, H. Huang, Analysis of the weighting exponent in the FCM, IEEE Transactions on Systems, Man and Cybernetics-—Part B: Cybernetics 34 (1) (2004) 634–639.

[21] N.R. Pal, J.C. Bezdek, Correction to "On cluster validity for the fuzzy *c*-means model, IEEE Transactions on Fuzzy Systems 5 (1997) 152–153.

[22] M.Y. Chen, D.A. Linkens, Rule-base self-generation and simplification for data-driven fuzzy models, Fuzzy Sets and Systems 142 (2004) 243–265.

[23] X.L. Xie, G. Beni, A validity measure for fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 841–847.

[24] S.H. Kwon, Cluster validity index for fuzzy clustering, Electron. Lett.ics Letters 34 (22) (1998) 2176–2177.

[25] J. Yu, C.-X. Li, Novel cluster validity index for FCM algorithm, J. Comp. Sci. & Technol.ournal of Computer Science and Technology 21 (2006) 137–140.

[26] W. Wei, J.M. Mendel, Optimality tests for the fuzzy *c*-means algorithm, Pattern Recognition 27 (11) (1994) 1567–1573.

[27] J. Breckenridge, Replicating cluster analysis: method, consistency and validity, Multivariate Behavioural Research 24 (1989) 147–161.

[28] T. Lange, V. Roth, M.L. Braun, J.M. Buhmann, Stability-based validation of clustering solutions, Neural Computation 16 (2004) 1299–1323.

[29] S. Dudoit, J. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a data set, Genome Biology 3 (2002) 1–21;
J. Fridlyand and, S. Dudoit, Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method, Technical Report 600, Department of Statistics, Berkeley, 2001.

[30] R. Tibshirani, G. Walther, Cluster validation by prediction strength, Journal of Computational and Graphical Statistics 14 (3) (2005) 511–528.

[31] A. Ben-Hur, A. Elisseeff, I. Guyon., A Stability based Method for Discovering Structure in Clustered Data, Pacific Symposium on Biocomputing, vol. 7, World Scientific Publishing Co., New Jersey, 2002, pp. 6–17.

[32] E. Levine, E. Domany, Resampling method for unsupervised estimation of cluster validity, Neural Computation 13 (2001) 2573–2593.

[33] M. Law, A. Jain, Cluster validity by bootstrapping partitions, Technical Report MSU-CSE-03-5, Department of Computer Science and Engineering, Michigan State University, 2002.

[34] S. Ben-David, U. von Luxburg, D. Pal, A sober look at clustering stability, in: Proceedings of the 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA, June 22–25, 2006.

[35] C. Borgelt, Resampling for fuzzy clustering, International Journal of Uncertainty, Fuzziness and Knowledge-Based 15 (2007) 595–614.

[36] W.M. Rand, Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association 66 (1971) 846–850.

[37] R.J.G.B. Campello, A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment, Pattern Recognition Letters 28 (2007) 833–841.

[38] A. Asuncion, D.J. Newman, UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2007 ⟨http://archive.ics.uci.edu/ml/⟩.

[39] R.N. Dave, Validating fuzzy partition obtained through c-shells clustering, Pattern Recognition Letters 17 (1996) 613–623.

[40] L.A. Zadeh, Fuzzy sets, Information and Control 8 (1965) 338–353.

[41] K. H. Tung, X. Xu, B.C. Ooi, CURLER: finding and visualizing nonlinear correlated clusters, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, June 14–16, 2005, Baltimore, USA, pp. 467–478.

[42] M. van Breukelen, R.P.W. Duin, D.M.J. Tax, J.E. den Hartog, Handwritten digit recognition by combined classifiers, Kybernetika 34 (1998) 381–386.

About the Author—MATTEO FALASCONI received his degree in Physics in 2000 from the University of Pavia. He obtained his Ph.D. degree in Materials Engineering from the University of Brescia in 2005. At present he is member of SENSOR Lab, University of Brescia. His research interests include pattern recognition and statistical data analysis for artificial and biological olfaction.

About the Author—AGUSTIN GUTIERREZ GALVEZ received his BE degrees in Physics and Electrical Engineering from University of Barcelona in 1995 and 2000, respectively. He received his Ph.D. degree in Computer Science from Texas A&M University in 2005. In 2006 he has been a JSPS postdoctoral fellow at Tokyo Institute of Technology. Currently he is with Department of Electronics, University of Barcelona. His research interests include biologically inspired processing for gas sensor-arrays, computational models of the olfactory system, pattern recognition, and dynamical systems.

About the Author—MATTEO PARDO got a degree in Physics (summa cum laude) in 1996 with a thesis in theoretical surface physics at the University of Milano. In 2000 he obtained the Ph.D. in Computer Engineering. Since 2002 he is a researcher of CNR-INFM. He is currently on leave of absence at the Max Planck Institute for Molecular Genetics in Berlin on a Von Humboldt fellowship. His research interest is data analysis and in particular the applications of pattern recognition techniques to artificial olfaction and genomics. He has 25 journal papers and is the technical chair of the International Symposium on Olfaction and Electronic Nose 2009.

About the Author—GIORGIO SBERVEGLIERI received his degree in Physics cum laude from the University of Parma (Italy), where in 1971 he started his research activities on the preparation of semiconductor thin film solar cells. In 1994, he was appointed full professor in Physics. At present he is director of the CNR–INFM Sensor Lab established in 1988 at the University of Brescia. Sensor Lab is devoted to the preparation and characterization of materials for gas sensing and to research on artificial olfaction.

About the Author—SANTIAGO MARCO is associate professor (Profesor Titular) at the Departament d'Electronica of Universitat de Barcelona since 1995. He received the degree in Physics from the Universitat de Barcelona in 1988. In 1993, he received his Ph.D. (honor award) degree from the Departament de Física Aplicada i Electrònica, Universitat de Barcelona. Currently he is associate professor of Physics at the Department of Electronics, University of Barcelona. His research interests are chemical instrumentation based on intelligent signal processing and microsystem modeling.