Comparing Fuzzy Clusterings in High Dimensionality

Stefano Rovetta[1](✉) and Francesco Masulli[1,2]

[1] DIBRIS – Dipartimento di Informatica, Bioingegneria,
Robotica e Ingegneria dei Sistemi, Università di Genova, Genova, Italy
stefano.rovetta@unige.it
[2] Center for Biotechnology, Temple University, Philadelphia, USA

Abstract. Due to the specificity of clustering, a problem that is intrinsically ill-posed, there are several approaches to comparing clusterings. Comparison of clusterings obtained in different conditions is often the only affordable evaluation strategy, due to the lack of a ground truth. In this chapter we address a class of dimensionality-independent methods which can be applied in the presence of a high-dimensional input space. Specifically, we review some generalizations of this class of methods to the case of fuzzy clustering, in several variants.

1 Introduction

High-dimensional data is encountered in most fields of science and technology. Although progress in data analysis and processing methods constantly changes the concept of what constitutes high dimensionality, there are some aspects of the problem which are inherent and unescapable, since they are more related to the ratio between data dimensionality and cardinality than to absolute values of either.

The most well-known description of such phenomena is termed the curse of dimensionality, which expresses the consequences of volume growing exponentially with the number of dimensions. These consequences include for instance:

- Sparsification of data (the *empty space* phenomenon): In many cases the number of observations (cardinality) is comparable with or even lower than the number of observed variables (dimensionality).
- Exponentially growing number of model parameters, with corresponding growth of necessary observations to obtain a given level of confidence or precision.
- Concentration effect on distances: For metrics of a very general form, maximum and minimum observed values tend to take on the same value with a probability that grows very rapidly with dimensionality; therefore distances are not meaningful any more.

While supervised analysis (for instance, classification) can count on a rich set of methods to keep the effects of dimensionality under control, such as model

F. Masulli et al. (Eds.): CHDD 2012, LNCS 7627, pp. 50–71, 2015.
DOI: 10.1007/978-3-662-48577-4_4

complexity estimation and working with kernels, for unsupervised methods and especially for clustering the same tools are not always available. This is because, while classification works with the mapping from data to nominal labels, clustering works directly with the structure of the space where the data live. Nonetheless, several techniques are commonly encountered for enabling clustering in high dimensions. These include (see also [29] in this book):

- Variable selection: Retain only the variables that are deemed significant to the problem at hand, according to some criterion.
- Subspace clustering: Perform clustering in spaces defined by a subset of the variables, and then search for the most meaningful subset. Alternatively, find a subspace (a more general linear projection, not necessarily axis-parallel) where clustering is most satisfactory.
- Intrinsic dimensionality estimation followed by (possibly nonlinear) dimensionality reduction: Find the dimension of the subset of the data space where the data actually "live", which most often is much lower than the number of observed variables, and then map data onto a lower-dimensional structure, which can be linear (a subspace), locally linear (union of several subspaces, each restricted to a given region), or non-linear (a manifold).
- Specialized metrics: Find ways to measure (dis)similarity that are less affected by concentration effects, for instance with particular values of the exponent in Minkowski metrics, or by using ranks instead of primary measures.
- Working with an affinity matrix rather than directly with data, using specific methods that do not require the direct representation of objects: Agglomerative clustering, correlation clustering, shared neighbors clustering.
- Using kernel and spectral clustering, which start from data representations and map them into affinity-based representations by using specific measures (kernels).

In this contribution we describe some techniques to measure similarities between pairs of different clusterings, taking advantage of the added flexibility provided by fuzzy clustering. We will review a few existing clustering similarity indexes, and describe some possible generalisations and extensions that make them applicable even in the fuzzy case. Some applications, using benchmark data sets, will also be shown.

A recent extensive survey [12] cites 76 measures of similarity or dissimilarity developed over the last century. The same problem can be cast as measuring diversity among classifiers or clusterings, binary string similarity, categorical feature similarity. A more recent trend has been to incorporate more information than just the coincidence of binary/categorical attributes; this includes for instance the development of fuzzy variants [10,40].

2 Fuzzy Clustering

2.1 Some Notations and Definitions

The task of data clustering can be defined using set-theoretic concepts. A *clustering* of a given data sample, a set of N data points in a metric space

$X = \{\mathbf{x}_1, \dots \mathbf{x}_N\}$, can be defined as a partition of the sample itself. This identifies *partitional* clustering methods. So clustering seeks a K-partition $\mathit{\Pi} = \{C_1, \dots, C_K\}$ of X. Each "part" of the partition is a cluster C_j, represented by a centroid from a set $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_K\}$. Note that these definitions allow us to extend partitions from the finite sample at hand to the whole data space where the data "live", thus providing what is commonly called an "out-of-sample extension".

It is customary to express attribution of data point $\mathbf{x}_l$ to cluster C_j by means of an indicator function $u(\mathbf{x}_l, C_j)$, the *membership function.* In the presence of a finite data set X, the membership values are the entries $u_{lj} = u(\mathbf{x}_l, C_j)$ of the *membership matrix* U, whose rows are *membership vectors* $\mathbf{u}_l$.

2.2 Fuzzy Clustering

When we want to take into account more refined models, with more information made available through the representation of clusters, we can resort to a *fuzzy* formalism [45]. In general, there are several ways to represent uncertainty with fuzzy sets, but in the special case of fuzzy data clustering [42] "fuzzy" means specifically representing partitions by means of real-valued indicator functions. This implies that a fuzzy clustering is not a conventional partition, but rather a fuzzy partition allowing for partial overlap of clusters.

When compared to standard clustering, fuzzy clustering provides a more flexible and powerful data representation paradigm. Fuzzy partitional methods based on centroids share some model parameters with their closest non-fuzzy counterparts, the number of clusters being the most notable example. However, most of them also require setting some additional parameters, which often play the role of degrees of fuzziness. As basic examples, we can mention Bezdek's fuzzy c-means [7] which needs the exponent m to be set to control fuzziness, and Krishnapuram and Keller's possibilistic c-means [26] which requires a set of width parameters β_j, one per cluster. In [30], we have proposed a *graded possibilistic c-means* clustering technique (GPCM) that provides control over the degree of possibility, thus allowing a soft transition between the standard probabilistic and the possibilistic models. This is done through an additional parameter α.

2.3 Methods for Fuzzy Clustering

Now we briefly review some fuzzy clustering methods, characterized by the fact that cluster centroids are defined as follows:

$$\mathbf{y}_j = \frac{\sum_{l=1}^{N} u_{lj}\mathbf{x}_l}{\sum_{j=1}^{K} u_{lj}}. \tag{1}$$

This formulation characterizes all the methods derived from c-means and is obtained from the minimization of a suitable Lagrangian, but does not depend on the actual computation of the memberships.

However, the membership function will differ according to the specific method; therefore the resulting centroids, whose computation depends on memberships as per Eq. (1), in general will not be the same in different cases.

One taxonomy of methods uses the constraint imposed on the sum of all membership for any given data point $\mathbf{x}_l$,

$$\zeta_l = \sum_{j=1}^{K} u_{lj}, \tag{2}$$

as a discriminant feature. It is useful to think of a membership vector $\mathbf{u}_l$ as a point in the K-dimensional space of possible combinations of memberships for the given point. The different feasible sets of membership values characterise each specific method, as detailed in the following.

When the sum of memberships is constrained to $\zeta_l = 1$, we are in the standard "probabilistic" case. With the usual formulation for crisp clustering, where $u_x(C) \in \{0, 1\}$, subject to the sum-1 constraint, only a set of K possible configurations is available, namely, those corresponding to the membership vectors that lie on the coordinate axes, a subset of the vertices of the unitary K-hypercube. Here one and only one of the memberships can be 1, while all others are zero.

A more interesting and expressive case is that of fuzzy clustering. Here the memberships lie on a segment of the K-dimensional hyperplane

$$\zeta_l = 1. \tag{3}$$

More specifically, they are located on the diagonal of the K-hypercube $[0,1]^K \in \mathbb{R}^K$. Memberships obeying this constraint are formally equivalent to probabilities. This case is termed "probabilistic" to stress this analogy. Crisp clustering is a limit case of general probabilistic clustering, where "probabilities" correspond to certainty; crisp memberships can only be located at the vertices of the hypercube.

The Maximum Entropy (ME) approach [38,39] makes explicit use of the probabilistic interpretation of memberships. In ME, by imposing the necessary minimum condition on an objective function with an entropic penalty, the problem can be stated as a minimization of the following Lagrangian:

$$J_{\mathrm{ME}} = \sum_{l=1}^{N} \sum_{j=1}^{K} \left[u_{lj} d_{lj}^2 + \eta u_{lj} \log u_{lj} \right], \tag{4}$$

and as a result of computing the necessary minimum conditions, memberships can be obtained from:

$$u_{lj} = \frac{e^{-d_{lj}/\beta}}{Z_l}. \tag{5}$$

where $Z_l = \sum_{j=1}^{K} e^{-d_{lj}/\beta}$ is termed the partition function. Clusters C_j then obey the Gibbs distribution around the respective centroids $\mathbf{y}_j$:

$$\Pr\left(\mathbf{x} \mid \Pi_j\right) = \frac{e^{-\beta \|\mathbf{x}-\mathbf{y}_j\|^2}}{Z(\mathbf{x})}, \tag{6}$$

where $\mathbf{y}_j$ is the centroid representing cluster $\mathbb{C}_j$ and β is a resolution parameter that, in a thermodynamic analogy, plays the role of a temperature. The quadratic distortion is the "energy" of "particle" $\mathbf{x}_l$ and $Z(\mathbf{x}_l)$ is the corresponding "partition function" at the specific "temperature" value $1/\beta$:

$$Z(\mathbf{x}) = \sum_k e^{-\beta||\mathbf{x}_l - \mathbf{y}_k||^2}. \tag{7}$$

The optimization procedure proposed for this model includes an "annealing" schedule to gradually lower the system's temperature. Since at each step a stable state is reached before moving to the next step, this method is also called Deterministic Annealing.

2.4 Possibilistic Clustering Models

The Possibilistic c-Means (PCM) [3,25,26] can be seen as being located at the other end of the spectrum with respect to the probabilistic Maximum Entropy method. It is based on removing any equality constraint on the sum of memberships, replaced by a set of loose requirements, which essentially allow the memberships themselves to take any configuration within the hypercube $[0, 1]^K$, with the exception of two isolated points, those with all-zero and all-one memberships, respectively. These are excluded by design by means of additional checks to avoid trivial solutions. Note that now the memberships are not formally equivalent to probabilities any more.

In this *possibilistic* case, taking as a reference the second formulation presented in [26], the objective function has the form

$$J_{\text{PCM}} = \sum_{l=1}^{N} \sum_{j=1}^{K} \left[u_{lj} d_{lj}^2 + \eta_j \left(u_{lj} \log u_{lj} - u_{lj} \right) \right], \tag{8}$$

and, again per the necessary minimum conditions, memberships are computed as

$$u_{lj} = e^{-d_{lj}/\beta_j}. \tag{9}$$

If we set a single value β for all the β_j, the only difference with Eq. 5 is in the denominator. To take advantage of this fact, we generalize the membership function as follows:

$$u_{lj} = \frac{v_{lj}}{Z_l}, \tag{10}$$

where we have introduced the *free membership* v_{lj}, defined as follows:

$$v_{lj} = e^{-d_{lj}/\beta_j}. \tag{11}$$

These functions share the same term for penalizing the overall distortion, but each of them has different additional penalties. As a result, the centroid location update equations remain the same, resulting in centers being placed at the barycenter of clusters weighted by membership. The membership update equations, which as per Eqs. 5 and 9 express the dependence of memberships from distances, differ by the form of the term Z_l. For a general class of clustering formulations and associated objective functions, we may redefine ζ_l in terms of the free memberships

$$\zeta_l = \sum_{j=1}^{K} v_{lj}. \tag{12}$$

For instance, in the probabilistic approaches $Z_l = \zeta_l$, whereas in standard possibilistic approaches $Z_l = 1$.

2.5 Graded Possibilistic Models

The classic probabilistic membership model, be it either hard or fuzzy, implements the concept of partitioning a set into disjoint subsets with memberships formally equivalent to the probability of one out of K mutually exclusive events. In the possibilistic approach each membership is formally equivalent to the probability of one out of K mutually *independent* events. Of course they may retain the usual fuzzy interpretation as degrees of truth rather than probabilities.

The graded possibilistic membership model assumes instead that events may be independent to a certain degree, but not completely, so that, while intermediate cases will be treated as independent, extreme cases (with some very high or very low membership values) will be considered mutually exclusive. This provides the method with a notable expressive power in terms of fuzzy modelling.

The partition function characterizing the Graded Possibilistic c-Means (GPCM) is derived from the *interval* constraint $\sum_{j=1}^{K} {u_{lj}}^{[\gamma]} = 1$. Here we use an interval variable $[\gamma] = [\,\gamma^{(l)}, \gamma^{(u)}\,]$. Note that ${u_{lj}}^{[\gamma]} = [\,{u_{lj}}^{\gamma^{(l)}}, {u_{lj}}^{\gamma^{(u)}}\,]$ since an exponential with interval exponent $[A] = [\underline{A}, \overline{A}]$ is the interval $e^{[A]} = [e^{\underline{A}}, e^{\overline{A}}]$ [34].

An interval variable is commonly interpreted as the *admissible range* for the actual value of an unknown variable. Adopting this interpretation, the mixed-type equality between a non-interval variable a and an interval variable $[A]$ has been conventionally used with the following meaning: The equality $a = [A]$ is true when $\underline{A} \leq a \leq \overline{A}$, or $a \in [A]$. In most applications of interval arithmetic, from numerical error bracketing to type-2 fuzzy sets, this means that $[A]$ is the uncertain representation of a.

In Ref. [30] this is explained in some more detail; here we restrict ourselves to a particular choice of $\gamma^{(l)}$ and $\gamma^{(u)}$, for which we obtain the specific implementation that we study in this work: $\gamma^{(l)} = \alpha$ and $\gamma^{(u)} = 1$, where $\alpha \in (0, 1]$ controls the "possibility level." In other words, the interval parameter $[\gamma]$ has the form

$$[\gamma] = [\alpha, 1]. \tag{13}$$

In this specific, asymmetric implementation, memberships whose sum exceeds 1 are forbidden. Therefore clustering is effectively competitive among nearby centroids. However, for far-away centroids, the competition decreases with α. This allows us to obtain points which are not attributed to any cluster, thus providing a very natural representation for outliers. The partition function is in this case computed as follows:

$$\begin{cases} Z_l = \sum_{j=1}^{K} v_{lj} & \text{if} \quad \sum_{j=1}^{K} v_{lj} > 1 \\ Z_l = \left(\sum_{j=1}^{K} v_{lj}^{\alpha} \right)^{1/\alpha} & \text{if} \quad \sum_{j=1}^{K} v_{lj}^{\alpha} < 1 \\ Z_l = 1 & \text{otherwise.} \end{cases} \tag{14}$$

For $\alpha = 1$, the representation properties of the method reduce to those of ME, while in the limit case for $\alpha \to 0$, the representation properties are equivalent to those of PCM-II for low membership values, and to those of ME for higher values.

This method, the Asymmetric Graded Possibilistic c-Means (AGPCM), has several nice properties that make it worth studying and using in practice:

- In [30] this particular model has been shown to possess robustness properties. The rejection ability can be applied in robust clustering and outlier analysis, while its reverse, outlier identification, can be used in novelty detection and data distribution characterization (or one-cluster clustering) as in [15].
- While for the fully possibilistic method it is very difficult to attain convergence, the graded approach has better convergence for α sufficiently larger than 0. It is also possible to "play" with parameters along the optimization process, for instance applying an annealing schedule to α, so as to exploit the best values in the most appropriate phase of the convergence: higher α in the initial steps, when centroids need to break symmetry and diverge; lower α in the later steps, when a precise, outlier-insensitive placement is sought.
- From the point of view of data analysis, full membership to more than one cluster, as allowed by a symmetric formulation, may have a difficult interpretation; in contrast, a point which does not belong to any cluster is easily interpreted as an outlier.
- As a quantitative counterpart of the previous point, memberships summing up to *at most* one allow a much easier comparison between clusterings, since the range of values for fuzzy similarity indexes depends on the values of memberships. This point will be discussed further on in this paper.
- Outlier insensitivity presents advantages with respect to convergence as well, since, while centroids in non-fuzzy clustering are insensitive to points outside their cluster, centroids in fuzzy clustering have to account for the effect of all points. This is not the case with the possibilistic model. An illustration of this increased precision in locating cluster centres is provided in Fig. 1.
- On the other hand, with respect to PCM and symmetric-GPCM, AGPCM features an effective repulsion between nearby centroids, thus reducing the risk of overlapping clusters.

The main disadvantage of this method is the presence of relatively many parameters that need to be set. No criterion was given in the original work to

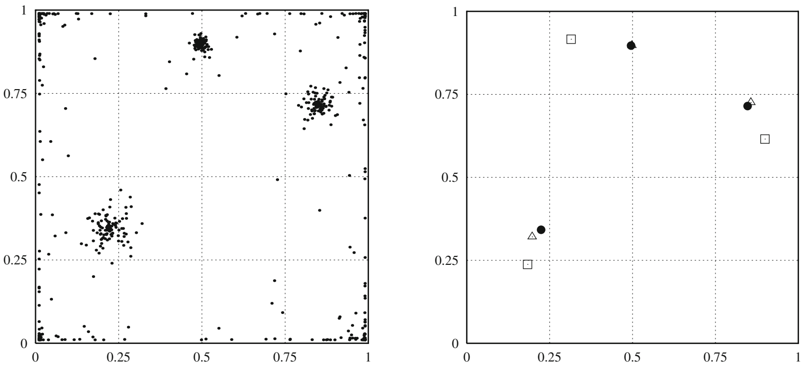

Fig. 1. Outlier rejection demonstration for AGPCM. Above: Data set. Below: Centroid locations. Black circles are the true cluster centres; triangles are centres found with $\alpha = 0$ (maximum rejection); squares are centres found with $\alpha = 1$ (no rejection, representationally equivalent to ME). Note that some triangles are hidden by true cluster centres since they are almost coincident; this is clearly not true for the "probabilistic" centroids. Figures from [30].

assess the value of α, while β was initialized following heuristics available in the literature; for both, "annealing" procedures were then applied, but without any measure of quality. The stability analysis described in the following sections was motivated by this lack of objective tools.

3 Comparing Fuzzy Clusterings

3.1 Approaches to the Comparison of Clusterings

Measuring the agreement between two clusterings amounts to measuring the similarity between two partitions, and there are several partition similarities available in the literature. It should be noted however that fuzzy clustering is not addressed very frequently in these works, even if it has several advantages over standard clustering from both representational and computational viewpoints [24,30]. The two main approaches include comparing clusters after matching them, and comparing co-association information.

The first approach consists in first identifying pairs of clusters, each composed of one cluster from the first partition and one from the second, which can be considered related, or, ideally the same cluster. A perfect match is difficult to obtain, and this *correspondence problem* may not have a satisfactory solution. The second step is to evaluate the degree of matching, and this is of course possible only if the first step succeeded.

A second approach is based on co-association. Two data items are co-associated if a partition puts them in the same cluster. The agreement or disagreement of partitions can be measured by coassociation, i.e., counting the number of pairs of data items on which both partitions agree, and comparing it with the number of pairs on which they disagree. Several classic indexes are based on this rationale. In the following we review some of them and describe some variations, including fuzzy and probabilistic versions.

This approach is followed in [16,21]. In [40] we defined a methodology to extend several indexes based on co-association to the fuzzy and possibilistic case, by means of co-association matrices [17]. Several indexes can simply be computed starting from the entries of a confusion or contingency matrix, which is readily obtained from the co-association matrix.

3.2 Notation

Suppose we select a partitional clustering method. For a specific choice of clustering parameters, possibly including selecting a given number of clusters and a random initialization, but excluding changes in the data sample, we obtain a given partition. If we repeat this process for a number of times, the i-th run will optimize a set $Y^i = \{\mathbf{y}^i_1, \ldots, \mathbf{y}^i_{K^i}\}$ of K^i centroids, which represent a K^i-partition $\Pi^i = \left\{C^i_1, \ldots, C^i_K\right\}$. We will not require that $K^i = K^k$ for $i \neq k$, i.e., it is not necessary to have the same number of clusters in different instances.

The i-th fuzzy indicator function will be similarly denoted as $u^i(\mathbf{x}_l, C^i_j)$, and likewise we will have $u^i_{lj} = u^i(\mathbf{x}_l, C^i_j)$. However, wherever we do not refer to specific instances nor need to differentiate among them explicitly, we will drop the sub/superscripts i, k to avoid a cumbersome notation.

We indicate the fact that partition Π^i puts two data items x_l and x_m in the same cluster by writing the indicator function $x_l \sim_i x_m$. The negation is expressed with the barred symbol: $x_l \nsim_i x_m$. This notation is borrowed from [5].

3.3 Co-association

In fuzzy clustering partitions are fuzzy, meaning that $\forall \mathbf{x}_l \in Xl : 1, \ldots, N$, the membership $u_{lj} = u(\mathbf{x}_l, C_j) \in [0, 1]$ for each cluster $C_j \in \mathcal{P}$ and $\forall l : 1, \ldots, N$ the constraint $\sum_{j=1}^{K} u_{lj} = 1$ holds as per Eq. (3); in addition, we allow possibilistic partitions by removing this last constraint. As discussed earlier, possibilistic partitions may be a meaningful extension to fuzzy partitions especially in the asymmetric constraint case (where $\sum_j u(\mathbf{x}_l, a_j) \leq 1$), although they can also be considered in symmetric graded cases and in fully possibilistic cases with a bit more interpretation effort.

Under a given partition Π, each data point is now represented by a membership vector. We define the *coassociation* ξ_{lm} between two data items $\mathbf{x}_l$ and $\mathbf{x}_m$ as the degree of similarity between the representation of the two items under the partition Π. Extending the notation of [5], we compute $\xi_{lm} = \mathbf{x}_l \sim \mathbf{x}_m$ as follows:

$$\xi_{lm} = \sum_{j=1}^{K} u_{lj} \wedge u_{mj}. \tag{15}$$

We can also define the *negative coassociation*, which is the logical complement of the coassociation:

$$\mathbf{x}_l \nsim \mathbf{x}_m = \overline{\xi_{lm}}. \tag{16}$$

Note that in the non-fuzzy case these definitions collapse to the propositions "partition Π puts/does not put $\mathbf{x}_l$ and $\mathbf{x}_m$ in the same cluster", but in the fuzzy case it is necessary to take all clusters into considerations because, in general, none of them will be exactly zero or one. On the other hand, in the fuzzy case, ξ_{lm} is a *degree* of coassociation rather than an integer value representing binary logic conditions.

3.4 Fuzzy Coassociation

To obtain a specific fuzzy instantiation of the general definition of coassociation just given, we have to appropriately define the conjunction connective [45]. We adopt the product t-norm [33], which provides uniformity with respect to other models of imprecision or uncertainty. The conjunction logical connective under the product t-norm is defined as $a \wedge b = ab$, and the negation operator as $\overline{a} = 1 - a$, so that $a \vee b = a + b - ab$ (the *probabilistic sum* t-conorm). The fuzzy coassociation between $\mathbf{x}_l$ and $\mathbf{x}_m$, therefore, is a real value ξ_{lm} computed as

$$\mathbf{x}_l \sim \mathbf{x}_m = \xi_{lm} = \sum_{j=1}^{K} u_{lj} u_{mj} = \mathbf{u}_l \cdot \mathbf{u}_m. \tag{17}$$

For a whole data set X, consider all possible pairs X^2. The coassociation of all pairs is a matrix Ξ, the *coassociation matrix* (also termed *bonding relationship* in [8]). This matrix is redundant, since by definition it is symmetric. In the following, as in [40], we *serialize* the coassociation matrix, taking only the upper triangular array corresponding to its elements above the diagonal, and we obtain a coassociation vector $\mathbf{s}$ of dimension $H = N(N+1)/2$, the N-th triangular number (the number of unique pairs of entries in the matrix *including* the diagonal).

The coassociation vector is defined as

$$\mathbf{s}_h = \xi_{lm} \quad \text{for} \quad h = l(l-1)/2 + m, \; h : 1 \ldots H. \tag{18}$$

when $l : 1 \ldots N$ and $m : 1 \ldots l$ (or, for C programmers: $h = l(l+1)/2 + m$, $h : 0 \ldots H-1$ when $l : 0 \ldots N-1$ and $m : 0 \ldots l$). As defined, the linear index h corresponds to a row-wise scan of the lower triangular part of Ξ, including the diagonal. Note that these diagonal entries, the self-co-associations (co-associations of points with themselves), are $\xi_{ll} = 1 \; \forall l$ only in non-fuzzy cases. In general, due to the triangle inequality,

$$\xi_{ll} = \sum_{j=1}^{K} u_{lj} u_{lj} = \sum_{j=1}^{K} {u_{lj}}^2 \leq \left(\sum_{j=1}^{K} u_{lj} \right)^2. \tag{19}$$

In the probabilistic case

$$\xi_{ll} \leq \left(\sum_{j=1}^{K} u_{lj} \right)^2 = 1, \tag{20}$$

where equality holds only for $u_{lj} = 1$ for some j (non-fuzzy case), whereas in the general possibilistic case

$$\xi_{ll} \leq \left(\sum_{j=1}^{K} u_{lj} \right)^2 \in \left(0, K^2\right) . \tag{21}$$

However, in the AGPCM case, again

$$\xi_{ll} \leq \left(\sum_{j=1}^{K} u_{lj} \right)^2 \leq 1. \tag{22}$$

We may note that the difference between the probabilistic case and AGPCM is in the lower bound, so that for AGPCM

$$\xi_{ll} \geq 0 \tag{23}$$

but in the probabilistic case

$$\xi_{ll} \geq \frac{1}{K^2}. \tag{24}$$

Coassociations $\xi_{lm} = \mathbf{x}_l \sim \mathbf{x}_m$ (between different points) can be shown to obey similar upper bounds, while the lower bound is 0 in all cases.

3.5 Comparing Two Partitions

To compare two partitions Π^i and Π^k we compute their respective coassociation vectors $\mathbf{s}^i$ and $\mathbf{s}^k$. Note that the dimension of these vectors is H (Eq. 18), so *it only depends on the number of points, not the number of clusters.* In other words, the proposed methodology can be applied without any problem to different-size partitions.

$$\mathcal{N} = \begin{bmatrix} \mathcal{N}_{00} & \mathcal{N}_{01} \\ \mathcal{N}_{10} & \mathcal{N}_{11} \end{bmatrix} \tag{25}$$

defined by

$$\begin{aligned} \mathcal{N}_{00} &= \text{number of items s.t. } x_l \nsim_i x_m \text{ and } x_l \nsim_k x_m \\ \mathcal{N}_{01} &= \text{number of items s.t. } x_l \nsim_i x_m \text{ and } x_l \sim_k x_m \\ \mathcal{N}_{10} &= \text{number of items s.t. } x_l \sim_i x_m \text{ and } x_l \nsim_k x_m \\ \mathcal{N}_{11} &= \text{number of items s.t. } x_l \sim_i x_m \text{ and } x_l \sim_k x_m \end{aligned} \tag{26}$$

or equivalently

$$\begin{aligned} \mathcal{N}_{00} &= \| \left(1 - \mathbf{s}^i\right) \wedge \left(1 - \mathbf{s}^k\right) \|_1 \\ \mathcal{N}_{01} &= \| \left(1 - \mathbf{s}^i\right) \wedge \mathbf{s}^k \|_1 \\ \mathcal{N}_{10} &= \| \mathbf{s}^i \wedge \left(1 - \mathbf{s}^k\right) \|_1 \\ \mathcal{N}_{11} &= \| \mathbf{s}^i \wedge \mathbf{s}^k \|_1 \end{aligned} \tag{27}$$

where $\| \cdot \|_1$ is the 1-norm. Many pairwise partition similarity indexes can be practically computed starting from the contingency matrix; reference [1] provides a table.

We will also refer to the normalized contingency matrix

$$\mathcal{F} = \frac{1}{\| \mathcal{N} \|_1} \mathcal{N}. \tag{28}$$

where in the crisp case $\| \mathcal{N} \|_1 = N$. Following [11], we will use an index chosen in $\{00, 01, 10, 11\}$ to refer to generic events, where 10 and 01 refer to disagreements, 11 to a positive agreement (coassociation in both partitions) and 00 to a negative agreement (non-coassociation in both partitions).

It is simple to verify that in the general case we can compute the entries of $\mathcal{N}$ as follows:

$$\begin{array}{l}
\mathcal{N}_{11} = \sum_{h=1}^{H} s_h^i \wedge s_h^k = \mathbf{s}^i \cdot \mathbf{s}^k \\
\mathcal{N}_{01} = \sum_{h=1}^{H} (\mathbf{1} - s_h^i) \wedge s_h^k = |\mathbf{s}^k|_1 - \mathbf{s}^i \cdot \mathbf{s}^k \\
\mathcal{N}_{10} = \sum_{h=1}^{H} s_h^i \wedge (\mathbf{1} - s_h^k) = |\mathbf{s}^i|_1 - \mathbf{s}^i \cdot \mathbf{s}^k \\
\mathcal{N}_{00} = \sum_{h=1}^{H} (\mathbf{1} - s_h^i) \wedge (\mathbf{1} - s_h^k) = H - |\mathbf{s}^i|_1 - |\mathbf{s}^k|_1 + \mathbf{s}^i \cdot \mathbf{s}^k,
\end{array} \tag{29}$$

where $\mathbf{1}$ is an H-vector of all 1, $\cdot$ is the usual dot product and $|\mathbf{v}|_1$ is the 1-norm of vector $\mathbf{v}$. This reduces to actual counts for proper partitions; the same direct interpretation is obviously not available for fuzzy partitions, but the above definitions still hold and can be used to derive the generalized indexes.

For unsupervised learning, similarity indexes combine the off-diagonal terms of M only in commutative operations, such as products or sums, because partitions should be analysed in a symmetric fashion, since no one of them plays the privileged role of a reference. Based on this observation, to make notations more compact, we can additionally define shorthand symbols:

$$\begin{array}{rl}
\pi = & \mathbf{s}^i \cdot \mathbf{s}^k \\
\sigma^i = & |\mathbf{s}^i|_1 \\
\sigma^k = & |\mathbf{s}^k|_1 \\
\sigma = & \sigma^i + \sigma^k,
\end{array} \tag{30}$$

so that

$$M = \begin{bmatrix} \pi & \sigma^i - \pi \\ \sigma^k - \pi & H - \sigma + \pi \end{bmatrix}. \tag{31}$$

4 Partition Similarity Indexes

As already noted, indexes of partition similarity based on co-association, and in particular on the contingency matrix M, can be computed by several approaches. Some of them are reviewed in [31] and some are experimentally compared in [27]. Here we use loosely the term *partition* to refer to crisp partitions, fuzzy partitions, and possibilistic clusters.

4.1 The Rand and Jaccard Indexes

The Rand index [36] is defined as

$$RI = \frac{\mathcal{N}_{00} + \mathcal{N}_{11}}{\mathcal{N}_{00} + \mathcal{N}_{11} + \mathcal{N}_{10} + \mathcal{N}_{01}} \tag{32}$$

The Rand index is known to have a higher sensitivity (lower false negative rate) than specificity (higher false positive rate). This is because the index does not incorporate a-priori assumptions on a given null hypothesis, therefore is not able to distinguish false negatives from true negatives. As a result, while the index is expected to output the value 1 for identical partitions, it will not necessarily output the value 0 for non-identical partitions. To cope with this known issue, a modified version of the Rand index was proposed by Hubert and Arabie [21] incorporating a "correction for chance" which provides the ability to compare partition diversity with the null model, a hypergeometric data assumption. The adjusted Rand index is another popular choice for comparing partitions.

The Jaccard index [22] is another well-known partition similarity measure. It is defined as the ratio of the size of the intersection of two sets A and B to the size of their union:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{33}$$

and the Jaccard distance is $D_J(A,B) = 1 - J(A,B)$.

When comparing two different partitions of the same set X, this index is usually computed from $\mathcal{N}$ as follows:

$$J(\Pi^i, \Pi^k) = \frac{\mathcal{N}_{11}}{\mathcal{N}_{11} + \mathcal{N}_{10} + \mathcal{N}_{01}}. \tag{34}$$

4.2 The Fuzzy Jaccard Index

Starting from the definition of the Jaccard index, the fuzzy generalization of 34 is straightforward:

$$J_f(\Pi^i, \Pi^k) = \frac{\pi}{\sigma - \pi}, \tag{35}$$

where Π^i and Π^k are fuzzy. We call this the *fuzzy Jaccard index* [40].

The choice of the Jaccard index over other possible measures is suggested by the conclusions drawn in [43] after analysing a set of 39 different measures. The Jaccard distance $1 - J$ is a metric; the value 0 is attained only for disjoint sets; and the value 1 if and only if the two compared sets are equal.

For the fuzzy Jaccard index, the bidirectional implication in the latter property holds only for non-fuzzy sets. Therefore self-co-association gives an indication about the degree of fuzziness of a clustering. We can also define the *normalized fuzzy Jaccard index* as:

$$J_{nf}(\Pi^i, \Pi^k) = \frac{J_f(\Pi^i, \Pi^k)}{\sqrt{J_f(\Pi^i, \Pi^i) J_f(\Pi^k, \Pi^k)}}, \tag{36}$$

which is 1 when comparing a partition with itself even in the fuzzy case. Therefore, an analysis based on both J_f and J_{nf} can evaluate partition similarity and partition confidence at the same time. Note that this very natural normalization is also applied in [16], although with a different aim (make the range of values comparable between measurements).

4.3 The Fuzzy Rand Index

Campello [10] and Brouwer [8] proposed fuzzy generalizations of the Rand [36], Adjusted Rand [21] and Jaccard [22] indexes. Brouwer's proposal is based on normalized dot products between bonding relationships, i.e., cosine similarities between fuzzy membership vectors.

Proceeding in a similar way to what we did with the Jaccard index, we can define a *fuzzy Rand index*:

$$R_f() = (\Pi^i, \Pi^k) = 1 + \frac{2\pi - \sigma}{H} \tag{37}$$

and a *normalized fuzzy Rand index*:

$$R_{nf}(\Pi^i, \Pi^k) = \frac{R_f(\Pi^i, \Pi^k)}{\sqrt{R_f(\Pi^i, \Pi^i) R_f(\Pi^k, \Pi^k)}}. \tag{38}$$

4.4 The Probabilistic Rand Index

We noted earlier that the Rand index suffers from a low specificity, and that the adjusted Rand index was designed to compensate this issue. In [11] another avenue was chosen to tackle the specificity problem, by including external information in the form of weights that change the relative importance of terms in the Rand index.

The rationale for this modification is that the terms of the contingency matrix should be given different levels of relevance, since they refer to cases providing different levels of information. In particular, there have been notable discussions among the practitioners [4,9,37] about whether the number of negative matches should be taken into account at all in similarity evaluation. The Jaccard index does not take this term into account. However the Rand index RI does. A weighted version of the Rand index was therefore defined by taking into account directly the a-priori probability of the four events of interest (prior to observing the data) $c \in \{00, 01, 10, 11\}$, namely, given a pair of arbitrary data items $(\mathbf{x}_l, \mathbf{x}_m)$, the probability that they are:

- in different clusters both in Π^A and in Π^B (event $h = 00$):

$$p_{00} = \Pr\left(\mathbf{x}_l \nsim_A \mathbf{x}_m \text{ and } \mathbf{x}_l \nsim_B \mathbf{x}_m\right);$$

- in the same cluster in Π^B but not in Π^A $(h = 01)$:

$$p_{01} = \Pr\left(\mathbf{x}_l \nsim_A \mathbf{x}_m \text{ and } \mathbf{x}_l \sim_B \mathbf{x}_m\right);$$

- in the same cluster in Π^A but not in Π^B ($h = 10$):

$$p_{10} = \Pr\left(\mathbf{x}_l \sim_A \mathbf{x}_m \text{ and } \mathbf{x}_l \nsim_B \mathbf{x}_m\right);$$

- in the same cluster both in Π^A and in Π^B ($h = 11$):

$$p_{11} = \Pr\left(\mathbf{x}_l \sim_A \mathbf{x}_m \text{ and } \mathbf{x}_l \sim_B \mathbf{x}_m\right).$$

Note that this definition is empirically approximated by the quantities defined in Eq. 26.

These values were computed in a maximum uncertainty (maximum entropy) hypothesis, where all clusters are equiprobable, no spatial structure is known, i.e., the probability of assigning a point to a cluster does not depend on its location, and points are uniformly sampled:

$$\begin{aligned} p_{00} &= \frac{K^A - 1}{K^B}\frac{K^A - 1}{K^B}; \\ p_{01} &= \frac{K^A - 1}{K^A}\frac{1}{K^B}; \\ p_{10} &= \frac{1}{K^A}\frac{K^B - 1}{K^B}; \\ p_{11} &= \frac{1}{K^A}\frac{1}{K^B}. \end{aligned} \tag{39}$$

Given the probability p_h of event h, the authors define a corresponding weight:

$$w_h = -\log p_h. \tag{40}$$

The *probabilistic Rand index* is defined as:

$$PRI = \frac{w_{00}\mathcal{N}_{00} + w_{11}\mathcal{N}_{11}}{w_{00}\mathcal{N}_{00} + w_{11}\mathcal{N}_{11} + w_{10}\mathcal{N}_{10} + w_{01}\mathcal{N}_{01}} \tag{41}$$

We can compute maximum likelihood a-posteriori estimates (given the data) of the probability of each of the four events of interest by approximating them with the observed relative frequencies:

$$q_h \approx f_h = \mathcal{F}_h. \tag{42}$$

By dividing numerator and denominator by the total sum, the indexes RI and PRI can be expressed using the observed frequencies f_h:

$$RI = \frac{f_{00} + f_{11}}{f_{00} + f_{11} + f_{10} + f_{01}} = f_{00} + f_{11} \approx q_{00} + q_{11} \tag{43}$$

and

$$\begin{aligned} PRI &= \frac{w_{00}f_{00} + w_{11}f_{11}}{w_{00}f_{00} + w_{11}f_{11} + w_{10}f_{10} + w_{01}f_{01}} \qquad (44) \\ &\approx \frac{w_{00}q_{00} + w_{11}q_{11}}{w_{00}q_{00} + w_{11}q_{11} + w_{10}q_{10} + w_{01}q_{01}} \qquad (45) \end{aligned}$$

This formulation makes it clear that the Rand index is the probability of agreement between two partitions, after observing the data, while the probabilistic Rand index also includes correction weights that depend on the a priori probability of agreement, *before* observing the data.

4.5 The Probabilistic Jaccard Index

At this point it could be noted that the same procedure can be applied to other contingency-matrix-based indexes, like the Jaccard index J defined in Eq. 34, which can be expressed in terms of the observed frequencies/probabilities:

$$JI = \frac{f_{11}}{1 - f_{00}} \approx \frac{q_{11}}{1 - q_{00}}. \tag{46}$$

A "probabilistic" (weighted) version analogous to PRI can be defined: The *probabilistic Jaccard index* is

$$\begin{aligned} PJI &= \frac{w_{11}\mathcal{N}_{11}}{w_{11}\mathcal{N}_{11} + w_{10}\mathcal{N}_{10} + w_{01}\mathcal{N}_{01}} \\ &\approx \frac{w_{11}q_{11}}{w_{11}q_{11} + w_{10}q_{10} + w_{01}q_{01}} \end{aligned} \tag{47}$$

with the same weight definitions as per Eq. 40.

5 Applications of Fuzzy Similarity Indexes

This section illustrates some applications of the dimensionality-independent fuzzy clustering similarity indexes discussed so far. The applications include a visual technique for stability analysis and monitoring the progress of clustering by deterministic annealing.

5.1 Visual Stability Analysis Based on Comparing Fuzzy Clusterings

Stability is the tendency of a learning system to be insensitive to changes in data or in model parameters. It is related to robustness [2] and to generalisation ability [23]. As already stated, in the context of clustering it is an important quality criterion to make up for the absence of supervised information for objective evaluation. Many applications of stability in this role have been proposed [6,28,44].

Cluster model selection it is one of the most studied issues in unsupervised pattern recognition, with a long history starting in cluster analysis [36], and then borrowing ideas from robust statistics [14,18–20,32,35].

In general (see Subsect. 3.4), fuzzy similarity indexes have the property that the level of fuzziness in the partitions is reflected in the maximum value that the index can reach. This is true also for possibilistic clusters, where an added feature is that the "best" clusterings are not only the stablest, but also those with the highest degree of self-similarity (value of the similarity index when comparing a clustering to itself). Self-similarity, therefore, acts as a measure of confidence. On the other hand, we have seen also normalized indexes for possibilistic clustering, so as to eliminate this sensitivity. In this case the analysis proceeds similarly to that of probabilistic and non-fuzzy clustering. Finally, pairing the normalized and unnormalized versions of an index makes it possible to perform both sensitivity and confidence evaluations simultaneously.

Here we discuss a visual and interactive procedure, allowing the user to perceive the effect of varying one or few parameters, in this case α and β in AGPCM. Visual analysis is effective for parameters with a smooth effect on clustering performance, so we don't suggest it, for instance, to choose between different initial conditions.

We resort to a graphic representation, a heat map which includes the complete information about the distribution of the index as a function of the parameters, and suggest some criteria to evaluate this information in a visual way. The clustering similarity matrix compares every possible pair of clusterings. The matrix is symmetric, but in the possibilistic case the diagonal may contains values lower than 1: usually this indicates that the cluster centres are not significant, i.e., that during training we found a bad local minimum. Therefore, we look for values for which the diagonal is brighter. To facilitate this search when self-similarity is particularly low, we can use J_{nf}, the normalized index. However, we monitor the maximum value of the self-similarity to keep the quality of the clustering under control.

For this experiment we have chosen a data set that has a good degree of structure, but at the same time is not clearly clustered. This results in a visible instability, for instance when starting from different initialization points. The problem is provided in the base data set package of the R language and environment (www.R-project.org) as "quakes". It consists of a subset of 1000 observations of quakes (seismic events with magnitude MB > 4.0) from a larger database of 5000 observations. These quakes occurred around Fiji, starting in 1964, and are described by three-dimensional coordinates (latitude, longitude and depth of event), plus the Richter magnitude and the number of stations that reported it, for a total of 5 variables.

Since setting a large number of cluster centroids reduces instability, we kept this number relatively small, fixing it at 7. The model parameter β was swept in 9 steps in the interval $[3.9 \times 10^{-3}, 3.2 \times 10^{-2}]$. The training was performed by 9 individual runs, each with a fixed value of β.

Each individual run consisted of one random initialization, and 30 complete optimizations, each one initialized with the output of the previous one, and α sweeping from 0.1 to 1 geometrically. Another parameter is varied across the 9 individual runs of each experiment. α, starting at 0.1 and progressing up to 1, so that we obtain 30×30 similarity matrices.

The visual output of the method is shown in Fig. 2 [41]. These are plots of the value of the fuzzy Jaccard index visualized as a heat map, the clearer the higher.

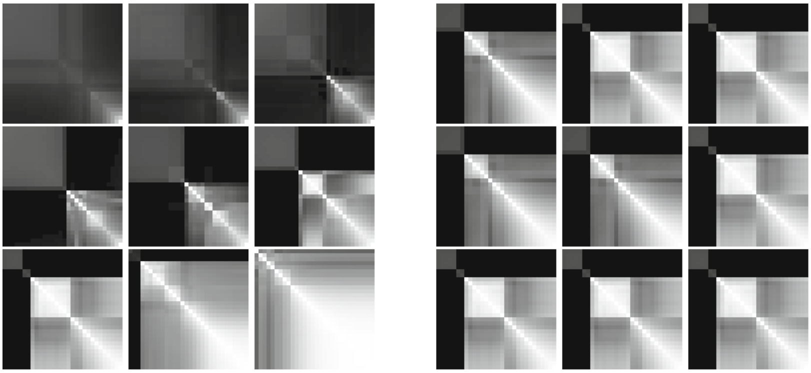

Fig. 2. Visual representation of the similarity index. Each individual heat map represents 30 experiments with different values of α ranging from 0.1 to 1. Across the 9 heat maps on the left (left to right, top to bottom): same initialization, β ranging from 0.0039 to 0.032. Across the 9 heat maps on the right: different random initializations, $\beta = 0.011$. From [41]

The first set of heat maps shows the variation as a function of β. The most stable, significant patch is attained in the seventh step ($\beta = 0.011$). The large, blurred patches in the last steps are due to the excessive value of the width parameter β. In this case, all points were attributed to a single, large cluster. On the other hand, when the width β is too small, even the diagonal has low values and only for the extreme values of α (lower right corner) data points are attributed to clusters with some confidence.

From this analysis, the best value for α is the one corresponding to the (row or column) coordinate of the center of the most stable area in the heat map. In this particular instance, the best value for α is not at the possibilistic or probabilistic extremes, but settles around an intermediate value, between 0.17 and 0.24. We select $\alpha = 0.21$.

This intermediate value is a confirmation that the Graded Possibilistic approach proposed in [30] actually provides a more flexible model, in terms of representation, than either the standard fuzzy or possibilistic methods.

If we now consider the experiments with fixed $\beta = 0.011$ and different random initializations (Fig. 2), we can see that, despite random variations in the results, the stable patch recurs in most experiments in about the same location, confirming the selected values of $\beta = 0.011$ and $\alpha = 0.21$.

5.2 Tracking Deterministic Annealing

As already noted, the Maximum Entropy clustering model described in Subsect. 2.3 is usually fit by an optimization procedure that involves gradual lowering of the model parameter. However, in contrast to the traditional Simulated Annealing [13] approach to minimization of functions of continuous variables, in this case a new annealing steps occurs only after a stable state has been reached

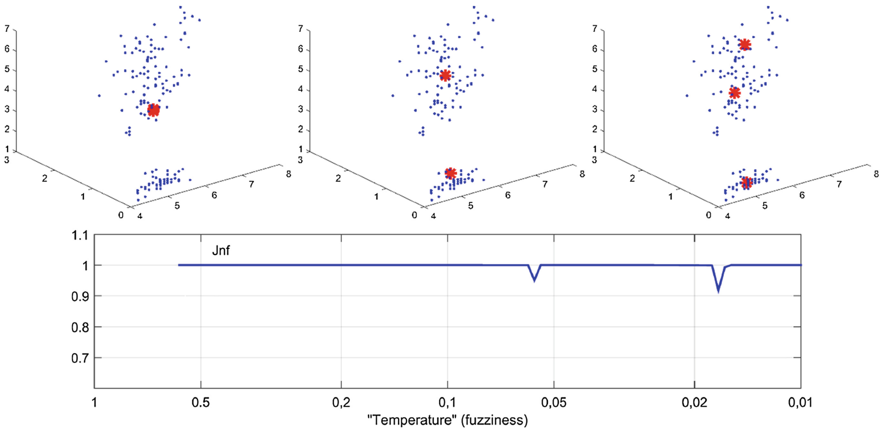

Fig. 3. Tracking deterministic annealing: Iris data.

at the previous step. This removes the main source of stochasticity, so that the method is termed Deterministic Annealing.

One peculiarity of this method is that the computational temperature parameter acts as a fuzzifier. This implies that, as the optimization progresses, less and less fuzzy clustering solutions are found, and for certain critical values of the temperature a phenomenon that parallels "phase transitions" [38] occurs. At phase transitions, cluster centroid that were overlapping because of the level of fuzziness are taken apart by the optimization. Without changing the number of centroids, this gives a hint about the number of clusters present in the data *at different resolutions*.

Fuzzy clustering comparison indices can be used to illustrate this phenomenon in high dimensionality, where the position of centroids is not easy to appreciate. But before applying the method in high dimensionality, we illustrate its operation in a lower-dimensional, well-known case, Iris data. Referring to Fig. 3, the top diagrams illustrate in three dimensions the position of centroids with respect to data in three stable states: The three centroids define one, two and three clusters depending on temperature.

The bottom diagram is a trace of the similarity of each pair of consecutive solutions, as measured by J_f and J_{nf}. In the stable states, solutions stay very similar to each other; this corresponds to flat areas where $J_{nf} = 1$. But at phase transitions, there is a sudden variation in clustering solutions. This is clearly pointed out by the notches in the graph of J_{nf}.

The high-dimensional problem chosen is the 20 Newsgroups data set. This is a collection of about 20000 Usenet posts from 20 different newsgroups. The selected version has been obtained from http://qwone.com/~jason/20Newsgroups/ already encoded by the vector space model.

To make the set more manageable only 1000 randomly selected samples from the first 5 newsgroups have been used. Due to the encoding adopted, the dimensionality has also been reduced from the original nearly 54 K to about 15 K dimensions. The number of terms in the dictionary depends logarithmically on the collection size, therefore the 1:20 reduction in data set cardinality results in less than 1:4 reduction in dimensionality and the reduced data set can still easily be categorized as high dimensional.

Figure 4 shows the result, obtained with 25 centroids. We expect about 5 clusters, so several centroids are going to overlap and we want to study the resolution (fuzziness) levels for which this overlapping is changed. Also in this case clear notches appear for quite well-defined values of the temperature parameter. By examining centroid positions in these configurations we can detect which centroids are overlapping, and how many effective clusters are there. This makes it possible to apply the multi-resolution analysis offered by the Deterministic Annealing method also in the presence of high dimensional data.

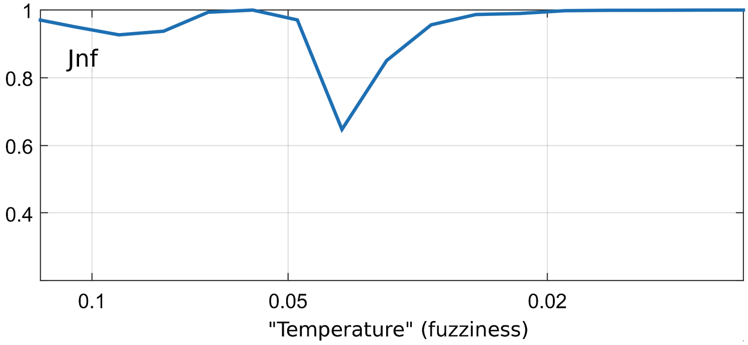

Fig. 4. Tracking deterministic annealing: 20 Newsgroups data.

6 Conclusion

In the presence of high dimensionality we face counter-intuitive situations, and visual inspection of clustering results would be beneficial. Some indices of mutual similarity between clusterings offer a way to perform this type of analysis in a dimensionality-independent way.

This chapter has presented some methods to extend these comparison indices to the cases of fuzzy and possibilistic methods. It turns out that comparing fuzzy clusterings reveals more information than in the crisp case.

In many cases we restricted the analysis to the Jaccard index, but a comparison between the possible choices from [40] could be performed.

References

1. Anderson, D.T., Bezdek, J.C., Popescu, M., Keller, J.M.: Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Trans. Fuzzy Syst. **18**(5), 906–918 (2010)
2. Anguita, D., Ridella, S., Rovetta, S.: Worst case analysis of weight inaccuracy effects in multilayer perceptrons. IEEE Trans. Neural Networks **10**(2), 415–418 (1999)
3. Barni, M., Cappellini, V., Mecocci, A.: Comments on ‘A possibilistic approach to clustering’. IEEE Trans. Fuzzy Syst. **4**(3), 393–396 (1996)
4. Baroni-Urbani, C., Buser, M.W.: Similarity of binary data. Syst. Biol. **25**(3), 251–259 (1976). http://sysbio.oxfordjournals.org/content/25/3/251.abstract
5. Ben-David, S., von Luxburg, U., Pál, D.: A sober look at clustering stability. In: Lugosi, G., Simon, H.U. (eds.) COLT 2006. LNCS (LNAI), vol. 4005, pp. 5–19. Springer, Heidelberg (2006)
6. Ben-Hur, A., Elisseeff, A., Guyon, I.: A stability based method for discovering structure in clustered data. In: Altman, R.B., Dunker, A.K., Hunter, L., Lauderdale, K., Klein, T.E. (eds.) BIOCOMPUTING 2002 Proceedings of the Pacific Symposium, pp. 6–17 (2001)
7. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
8. Brouwer, R.K.: Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. J. Intell. Inf. Syst. **32**(3), 213–235 (2009)
9. Buser, M.W., Baroni-Urbani, C.: A direct nondimensional clustering method for binary data. Biometrics **38**(2), 351–360 (1982). http://www.jstor.org/stable/2530449
10. Campello, R.J.G.B.: Generalized external indexes for comparing data partitions with overlapping categories. Pattern Recogn. Lett. **31**, 966–975 (2010)
11. Carpineto, C., Romano, G.: Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **34**(12), 2315–2326 (2012)
12. Choi, S.S., Cha, S.H., Tappert, C.C.: A survey of binary similarity and distance measures. J. Systemics Cybern. Inf. **8**, 43–48 (2010)
13. Corana, A., Marchesi, M., Martini, C., Ridella, S.: Minimizing multimodal functions of continuous variables with the “simulated annealing” algorithm. ACM Trans. Math. Softw. **13**(3), 262–280 (1987)
14. Davé, R.N., Krishnapuram, R.: Robust clustering methods: a unified view. IEEE Trans. Fuzzy Syst. **5**(2), 270–293 (1997)
15. Filippone, M., Masulli, F., Rovetta, S.: Applying the possibilistic c-means algorithm in kernel-induced spaces. IEEE Trans. Fuzzy Syst. **18**, 572–584 (2010)
16. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983). http://dx.doi.org/10.2307/2288117
17. Fred, A.L.N., Jain, A.K.: Data clustering using evidence accumulation. Int. Conf. Pattern Recog. **4**, 276–280 (2002)
18. Frigui, H., Krishnapuram, R.: A robust competitive clustering algorithm with applications in computer vision. IEEE Trans. Pattern Anal. Mach. Intell. **21**(5), 450–465 (1999)
19. Frigui, H., Krishnapuram, R.: A robust clustering algorithm based on m-estimator. In: Proceedings of the 1st International Conference on Neural, Parallel and Scientific Computations, Atlanta, USA, vol. 1, pp. 163–166, May 1995
20. Huber, P.J.: Robust Stat. Wiley, New York (1981)
21. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)

22. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bull. Soc. Vaudoise des Sci. Nat. **37**, 547–579 (1901)
23. Kearns, M., Schapire, R.: Efficient distribution-free learning of probabilistic concepts. J. Comput. Syst. Sci. **48**(3), 464–497 (1994)
24. Klawonn, F.: Fuzzy clustering: insights and a new approach. Mathware Soft Comput. **11**(3), 125–142 (2004)
25. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Trans. Fuzzy Syst. **1**(2), 98–110 (1993)
26. Krishnapuram, R., Keller, J.M.: The possibilistic C-Means algorithm: insights and recommendations. IEEE Trans. Fuzzy Syst. **4**(3), 385–393 (1996)
27. Kuncheva, L.I., Vetrov, D.P.: Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Trans. Pattern Anal. Mach. Intell. **28**(11), 1798–1808 (2006)
28. Lange, T., Roth, V., Braun, M.L., Buhmann, J.M.: Stability-based validation of clustering solutions. Neural Comput. **16**(6), 1299–1323 (2004)
29. Masulli, F., Rovetta, S.: Clustering High-Dimensional Data. In: Proceedings of CHDD 2012, Clustering High-Dimensional Data, Series Lecture Notes in Computer Science, LNCS 7627, 1, Springer-Verlag, Heidelberg, Germany (2015)
30. Masulli, F., Rovetta, S.: Soft transition from probabilistic to possibilistic fuzzy clustering. IEEE Trans. Fuzzy Syst. **14**(4), 516–527 (2006)
31. Meilă, M.: Comparing clusterings-an information based distance. J. Multivar. Anal. **98**(5), 873–895 (2007). http://dx.doi.org/10.1016/j.jmva.2006.11.013
32. Ménard, M., Courboulay, V., Dardignac, P.A.: Possibilistic and probabilistic fuzzy clustering: unification within the framework of the non-extensive thermostatistics. Pattern Recogn. **36**(6), 1325–1342 (2003)
33. Menger, K.: Statistical metrics. Proc. Natl. Acad. Sci. U.S.A. **28**(12), 535–537 (1942)
34. Moore, R.E., Kearfott, R.B., Cloud, M.J.: Introduction to Interval Analysis. Society for Industrial Mathematics, Philadelphia (2009)
35. Pal, N.R., Pal, K., Bezdek, J.C.: A mixed c-Means clustering model. In: FUZZIEEE97: Proceedings of the International Conference on Fuzzy Systems, pp. 11–21. IEEE, Barcelona (1997)
36. Rand, W.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
37. Real, R., Vargas, J.M.: The probabilistic basis of jaccard's index of similarity. Syst. Biol. **45**, 380–385 (1996)
38. Rose, K., Gurewitz, E., Fox, G.: A deterministic annealing approach to clustering. Pattern Recogn. Lett. **11**, 589–594 (1990)
39. Rose, K., Gurewitz, E., Fox, G.: Statistical mechanics and phase transitions in clustering. Phys. Rev. Lett. **65**, 945–948 (1990)
40. Rovetta, S., Masulli, F.: An experimental validation of some indexes of fuzzy clustering similarity. In: Di Gesù, V., Pal, S.K., Petrosino, A. (eds.) WILF 2009. LNCS, vol. 5571, pp. 132–139. Springer, Heidelberg (2009)
41. Rovetta, S., Masulli, F.: Visual stability analysis for model selection in graded possibilistic clustering. Inf. Sci. **279**, 37–51 (2014)
42. Ruspini, E.H.: A new approach to clustering. Inf. Control **15**(1), 22–32 (1969)
43. Shi, G.: Multivariate data analysis in palaeoecology and palaeobiogeographya review. Palaeogeogr. Palaeoclimatol. Palaeoecol. **105**(3–4), 199–234 (1993)
44. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. J. Roy. Stat. Soc. Ser. B Stat. Methodol. **63**(2), 411–423 (2001). http://dx.doi.org/10.1111/1467-9868.00293
45. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**(3), 338–353 (1965)