Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 833–841
Pattern Recognition Letters
www.elsevier.com/locate/patrec

A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment

R.J.G.B. Campello *

SCC/ICMC/USP, University of São Paulo at São Carlos, Av. do Trabalhador São-Carlense 400, São Carlos, SP, 13560-970, Brazil

Received 21 March 2006; received in revised form 21 August 2006
Available online 17 January 2007

Communicated by W. Pedrycz

Abstract

A fuzzy extension of the Rand index [Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc. 846–850] is introduced in this paper. The Rand index is a traditional criterion for assessment and comparison of different results provided by classifiers and clustering algorithms. It is able to measure the quality of different hard partitions of a data set from a classification perspective, including partitions with different numbers of classes or clusters. The original Rand index is extended here by making it able to evaluate a fuzzy partition of a data set – provided by a fuzzy clustering algorithm or a classifier with fuzzy-like outputs – against a reference hard partition that encodes the actual (known) data classes. A theoretical formulation based on formal concepts from the fuzzy set theory is derived and used as a basis for the mathematical interpretation of the Fuzzy Rand Index proposed. The fuzzy counterparts of other (five) related indexes, namely, the Adjusted Rand Index of Hubert and Arabie, the Jaccard coefficient, the Minkowski measure, the Fowlkes–Mallows Index, and the Γ statistics, are also derived from this formulation.

Keywords: Fuzzy clustering; Fuzzy classification; External validity indexes; Rand index; Adjusted Rand index; Jaccard coefficient; Minkowski measure; Fowlkes–Mallows index; Γ Statistics

1. Introduction

Clustering is essentially a problem in which the goal is to determine a finite set of categories (clusters) to describe a data set according to similarities among its objects (Jain and Dubes, 1988; Kaufman and Rousseeuw, 1990; Everitt et al., 2001). It is an unsupervised task since the categories are structures to be found in data, i.e., they are not predefined or previously known. Classification, in turn, is a problem in which the goal is to derive a system, called classifier, that is able to assign data objects to predefined categories (classes) (Mitchell, 1997; Duda et al., 2001). It is a supervised task since the classifier is derived from examples using the class labels of a subset of data objects whose labels are previously known. When a clustering or classification task is performed on a set of N data objects $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N\}$, where $\boldsymbol{x}_j$ are feature vectors consisting of n attributes, the final result is usually a partition of the data set into a certain number k of categories, such that:

$$P = [p_{ij}]_{k\times N}; \quad p_{ij} \in \{0,1\}; \quad \sum_{i=1}^{k} p_{ij} = 1 \quad \forall j \tag{1}$$

where P is a $k \times N$ partition matrix whose element p_{ij} is either 1 if the jth object belongs to the ith cluster/class or 0 otherwise.

Assessing the accuracy of classification results is not a difficult problem since any classifier is designed to represent the right (known) number of classes in a given data set. The most popular classification measure is possibly the misclassification rate, which is the ratio of the number of misclassified data objects to the total number of objects in a

* Tel.: +55 16 3373 9671; fax: +55 16 3373 9751.
E-mail address: ricardo.campello@pesquisador.cnpq.br

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.11.010

data set. Formally, given a reference partition $R = [r_{ij}]_{k\times N}$ according to the constraints in (1), with R representing the right classes known to exist in the data, and another partition $Q = [q_{ij}]_{k\times N}$ – also according to (1) – to be assessed, the misclassification rate can be written as:

$$m = \frac{1}{2N}\sum_{j=1}^{N}\sum_{i=1}^{k}|r_{ij} - q_{ij}| \tag{2}$$

Note that: (i) $m \in [0,1]$; and (ii) $m = 0$ if and only if the classification result under evaluation matches exactly the actual classes, i.e., iff $Q = R$.

When talking about clustering, it follows that the accuracy assessment problem is far more complicated. Cluster validity measures exist that are very useful in practice as quantitative criteria for evaluating the quality of clustering partitions (e.g., see (Halkidi et al., 2001) and references therein). These measures are important, for instance, to the estimation of the number of clusters in n-dimensional data sets, where visual inspection is prohibitive for "large" n (Bezdek and Pal, 1998). However, it is well-known that different measures may perform differently in different scenarios (Pal and Bezdek, 1995; Campello and Hruschka, 2006). This is expected since the concept of cluster itself is quite subjective, but makes it difficult to adopt a specific measure as an *absolute reference* for performance comparisons involving different clustering results (resulting from different algorithms or even from the same algorithm with different settings). Even more difficult is to compare the performances of the existing cluster validity measures themselves and, eventually, that of a new measure to be proposed, as this task requires an absolute (external) rather than subjective (internal or relative) referential standard.

The performance in terms of classification can be used as an absolute referential standard so as to get around the drawbacks mentioned above concerning the accuracy assessment problem in clustering tasks. In other words, the quality of the clustering partitions can be externally assessed by verifying the degree to which the obtained clusters match the classes in a classification data set (where the classes are known). In this case, classification benchmarks can be used that are widely available (e.g., see the UCI machine learning repository (Newman et al., 1998)). Indeed, the known classes in a data set not seldom are adequate approximations of the natural clusters in that data. More importantly, classification is often the final goal of clustering. In this context, the misclassification rate in (1) can in principle be used to evaluate the quality of clustering partitions, even if the number of clusters is different from the number of classes. To do so, it is assumed that each cluster is assigned to a class based on the majority class (the class to which most of the cluster's objects belong). Hence, a given object is "misclassified" if the class to which the object's cluster was assigned does not match the right class corresponding to that object. This procedure is a simple approximation that allows one to evaluate the quality of partitions in which the number of clusters is not necessarily the same as the number of classes, but it is not the most accurate approach to accomplish this task.

A more stringent external criterion for evaluating clustering partitions from a classification perspective is the so-called Rand index proposed by Rand (1971). This index, which will be described in details in Section 2, deals primarily with pairs of data objects and uses two different types of inconsistent classifications as well as two different types of consistent classifications, rather than the concepts of misclassification and correct classification only. The Rand index, as it was originally formulated, allows uniquely the evaluation of hard (so-called crisp or non-fuzzy) clustering partitions. It is well-known, however, that most data sets comprise ill-delineated subsets that cannot be adequately split this way. For instance, there are situations in which the data comprise categories that overlap with each other to some degree. In these cases, the use of clustering algorithms that are capable of dealing with such overlapping data clusters is recommended. Fuzzy clustering algorithms can naturally cope with this sort of problem since they aim to find fuzzy clusters to which all the data objects belong to some degree (Bezdek, 1981; Höppner et al., 1999). These algorithms have been shown to be one of the most successful approaches to data-driven design of fuzzy and neuro-fuzzy systems, namely, models, predictors, controllers and classifiers (Kosko, 1997; Babuška, 1998; Höppner et al., 1999; Dumitrescu et al., 2000). When applied to a set of N data objects, their final result is usually a partition of the data set into a certain number v of fuzzy clusters, such that:

$$F = [f_{ij}]_{v\times N}; \quad f_{ij} \in [0,1] \tag{3}$$

where F is a $v \times N$ fuzzy partition matrix whose element f_{ij} represents the fuzzy membership of the jth object to the ith fuzzy cluster.

The use of the original Rand index to evaluate fuzzy partitions as in (3) requires the partition matrix F to be converted into a hard partition as in (1) by assigning 1 to the highest membership value of each data object and 0 to the others. As will be shown in Section 3.5, by using this procedure it follows that different fuzzy partitions (describing data structures with notably different spatial distributions) may result into the same crisp partition and, accordingly, into the same value for the Rand index. This means that the loss of information due to the disposal of the fuzzy membership values causes the Rand index to be unable to discriminate between overlapped and non-overlapped data clusters. As such, this index is not appropriate for fuzzy clustering assessment.

A fuzzy extension of the Rand index, hereafter called Fuzzy Rand Index, is proposed in this paper in order to get around the information loss that is unavoidable when using the original index to evaluate fuzzy partitions. The extended index is obtained by first rewriting the original formulation of the Rand index in a fully equivalent form by using basic concepts from the set theory. This equivalent formulation is then extended to the fuzzy domain by using

analogous concepts from the fuzzy set theory (Pedrycz and Gomide, 1998; Zimmermann, 2001).

Although the Rand index is more stringent and reliable than the majority class method mentioned above, it has some shortcomings that will be further discussed in Section 3.3. In brief, the Rand index can be shown to have biases that may make its results to be misleading in particular application scenarios. Fortunately, there is a family of other external indexes that can be used in order to get more accurate results, such as the Adjusted Rand Index of Hubert and Arabie (Hubert and Arabie, 1985), the Jaccard coefficient (Halkidi et al., 2001), the Minkowski measure (Jiang et al., 2004), the Fowlkes–Mallows index (Fowlkes and Mallows, 1983), and the Γ statistics (Jain and Dubes, 1988). These indexes are closely related to the Rand index in that they can be constructed from the same building blocks based on pairwise comparisons of data objects. For this reason, the same set-theoretic formulation proposed to extend the Rand index to the fuzzy domain will also be used in the present paper to derive fuzzy extensions of these related indexes in a unified fashion.

The remaining of this paper is organized as follows. In Section 2, the original Rand index is reviewed. Section 3 is concerned with the proposed fuzzy extension of this index. Specifically, the Rand index is first rewritten in an alternative way by using basic concepts from the set theory in Section 3.1. Then, a Fuzzy Rand Index is derived in Section 3.2 by using analogous concepts from the fuzzy set theory. The fuzzy counterparts of five related indexes, namely, the Adjusted Rand Index of Hubert and Arabie, the Jaccard coefficient, the Minkowski measure, the Fowlkes–Mallows index, and the Γ statistics, are also derived from the same basic formulation in Section 3.3. A step-by-step algorithm for computing these fuzzy indexes is described in Section 3.4. For the sake of illustration, a conceptual example and the corresponding discussions are presented in Section 3.5. Finally, the conclusions and some perspectives for future work are addressed in Section 4.

2. Review of the original Rand index

The Rand index (Rand, 1971) can be seen as an absolute criterion or referential standard that allows the use of classification data sets for performance assessment not only of classifiers (which can produce different data partitions with the right number of classes), but of clustering results (in which different data partitions can be composed of different numbers of clusters) as well. This very simple and intuitive index handles two hard partition matrices (R and Q) of the same data set. The reference partition, R, encodes the class labels, i.e., it partitions the data into k known classes. Partition Q, in turn, partitions the data into v categories (classes or clusters), and is the one to be evaluated. The categories encoded by Q will be, from now on, called *clusters*, no matter whether they result from a clustering algorithm or from a classifier. This way, one can easily distinguish between them and the right *classes* encoded by R.

Given the above remarks, the Rand index is then defined as (Rand, 1971; Jain and Dubes, 1988; Everitt et al., 2001):

$$\omega = \frac{a+d}{a+b+c+d} \tag{4}$$

where:

- a: Number of pairs of data objects belonging to the same class in R and to the same cluster in Q.
- b: Number of pairs of data objects belonging to the same class in R and to different clusters in Q.
- c: Number of pairs of data objects belonging to different classes in R and to the same cluster in Q.
- d: Number of pairs of data objects belonging to different classes in R and to different clusters in Q.

Terms a and d are measures of consistent classifications (agreements), whereas terms b and c are measures of inconsistent classifications (disagreements). Note that: (i) $\omega \in [0,1]$; (ii) $\omega = 0$ iff Q is completely inconsistent, i.e., $a = d = 0$; and (iii) $\omega = 1$ iff the partition under evaluation matches exactly the reference partition, i.e., $b = c = 0$ ($Q = R$).

As an example, consider the data set in Fig. 1. It is composed of two classes of data objects with 4 objects each. Class 1 (in circles) is composed of objects 1, 2, 3, and 5, whereas Class 2 (in squares) is composed of objects 4, 6, 7, and 8. The reference partition R for this example is then given by:

$$R = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} \tag{5}$$

As shown in Fig. 1, the same data set has been partitioned into 3 clusters – Cluster 1 in black (objects 1, 2, 3, and 4), Cluster 2 in white (objects 5 and 6), and Cluster

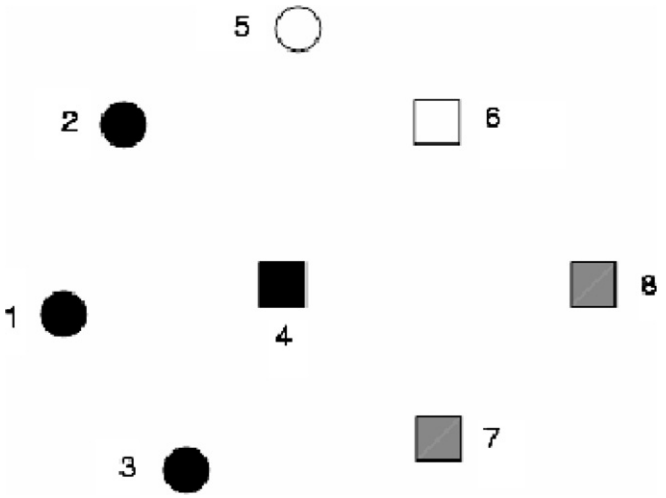

Fig. 1. Data set with $k = 2$ classes (class 1 in circles and class 2 in squares) partitioned into $v = 3$ clusters (clusters 1, 2 and 3 in black, white, and gray, respectively).

3 in gray (objects 7 and 8). The corresponding partition matrix is given by:

$$Q = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix} \tag{6}$$

The pairs of data objects belonging to the same class in R and to the same cluster in Q are therefore (1,2), (1,3), (2,3), and (7,8). Thus, the number of pairs of objects in the same class and in the same cluster is $a = 4$. Similarly, the other terms in (4) can easily be computed as $b = 8$, $c = 4$, and $d = 12$, which results in a Rand index of $\omega = 0.5714$.

3. Fuzzy Rand index and related indexes

A fuzzy extension of the Rand index reviewed in the previous section is derived here by first rewriting the formulation of the original index given by Eq. (4) in a fully equivalent set-theoretic form, as discussed in the sequel.

3.1. Set-Theoretic formulation for the Rand index

Let V, X, Y, and Z be the following crisp sets:

- V: Set of pairs of data objects belonging to the same class in R.
- X: Set of pairs of data objects belonging to different classes in R.
- Y: Set of pairs of data objects belonging to the same cluster in Q.
- Z: Set of pairs of data objects belonging to different clusters in Q.

From the above definitions, it is clear that the individual terms that compose the Rand index in (4) can be written as:

$$\begin{aligned} a &= |V \cap Y| \\ b &= |V \cap Z| \\ c &= |X \cap Y| \\ d &= |X \cap Z| \end{aligned} \tag{7}$$

where $|\cdot|$ and $\cap$ stand for cardinality and intersection of sets, respectively. In addition, it is straightforward to verify that sets V, X, Y, and Z can be written as:

$$\begin{aligned} V &= \bigcup_{i=1}^{k} V_i \\ X &= \bigcup_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{k} X_{i_1 i_2} \\ Y &= \bigcup_{l=1}^{v} Y_l \\ Z &= \bigcup_{\substack{l_1,l_2=1 \\ l_1 \neq l_2}}^{v} Z_{l_1 l_2} \end{aligned} \tag{8}$$

where $\cup$ stands for union of sets and:

- V_i: Set of pairs of data objects belonging to the ith class in R.
- $X_{i_1 i_2}$: Set of pairs of data objects belonging to different classes i_1 and i_2 in R, i.e., one object belonging to the i_1th class and the other object belonging to the i_2th class ($i_1 \neq i_2$).
- Y_l: Set of pairs of data objects belonging to the lth cluster in Q.
- $Z_{l_1 l_2}$: Set of pairs of data objects belonging to different clusters l_1 and l_2 in Q, i.e., one object belonging to the l_1th cluster and the other object belonging to the l_2th cluster ($l_1 \neq l_2$).

Then, the alternative equivalent formulation for the Rand index is derived simply by substituting (7) into (4), i.e.:

$$\omega = \frac{|V \cap Y| + |X \cap Z|}{|V \cap Y| + |V \cap Z| + |X \cap Y| + |X \cap Z|} \tag{9}$$

where V, X, Y, and Z are given by (8). The detailed computations with respect to the example displayed in Fig. 1, for instance, are provided in Appendix.

3.2. Fuzzy Rand index

Let Q be a fuzzy partition matrix resulting from a fuzzy clustering algorithm or from a classifier with fuzzy-like outputs,[1] such that $Q = [q_{ij}]_{v \times N}$ with $q_{ij} \in [0,1]$. Then, a fuzzy extension of the Rand index for accuracy assessment of Q with respect to a partition R encoding the known classes in a data set can be obtained simply by redefining the crisp sets V, V_i, X, $X_{i_1 i_2}$, Y, Y_l, Z, and $Z_{l_1 l_2}$ in (8) and (9) as fuzzy sets, in such a way that $|\cdot|$, $\cup$, and $\cap$ become the cardinality (so-called sigma count), union, and intersection of fuzzy sets, respectively. Note that, since the classes are known, then R can still be treated as a hard partition matrix (see Remark 3 below). For the sake of convenience, however, this matrix will also be treated as being fuzzy in the subsequent developments, i.e., $R = [r_{ij}]_{k \times N}$ with $r_{ij} \in [0,1]$.

Before proceeding with a more insightful analysis, it is worth remarking that, when dealing with a fuzzy partition Q, every object of the corresponding data set belongs to some (possibly null) degree to every fuzzy cluster in that partition. Accordingly, every pair of objects belongs to some degree to the fuzzy extension of every crisp set in (8). In other words, the fuzzy counterparts of the crisp sets defined in Section 3.1 are such that:

$$V, V_i, X, X_{i_1 i_2}, Y, Y_l, Z, Z_{l_1 l_2} : \Omega \to [0,1] \tag{10}$$

where Ω is the set of all non-ordered pairs of data objects ($|\Omega| = N(N-1)/2$). As such, they can be redefined as:

[1] Such as a fuzzy, neural, or neuro-fuzzy classifier with *one-of-k* class coding, where each output represents the membership degree of the input sample to a given class (Smith, 1993; Bigus, 1996; Duda et al., 2001).

- V_i: Fuzzy set of pairs of data objects belonging to the ith class in R. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the proposition "object j_1 belongs to the ith class and object j_2 belongs to the ith class", that is, $r_{ij_1}\mathrm{t}r_{ij_2}$, where "t" is a triangular norm (t-norm, such as min or product) used as a conjunction to implement the connective "and" of the proposition.
- $X_{i_1i_2}$: Fuzzy set of pairs of data objects belonging to different classes i_1 and i_2 in R. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the proposition "object j_1 belongs to the i_1th class and object j_2 belongs to the i_2th class", that is, $r_{i_1j_1}\mathrm{t}r_{i_2j_2}$.
- Y_l: Fuzzy set of pairs of data objects belonging to the lth cluster in Q. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the proposition "object j_1 belongs to the lth cluster and object j_2 belongs to the lth cluster", that is, $q_{lj_1}\mathrm{t}q_{lj_2}$.
- $Z_{l_1l_2}$: Fuzzy set of pairs of data objects belonging to different clusters l_1 and l_2 in Q. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the proposition "object j_1 belongs to the l_1th cluster and object j_2 belongs to the l_2th cluster", that is, $q_{l_1j_1}\mathrm{t}q_{l_2j_2}$.
- V: Fuzzy set of pairs of data objects belonging to the same class in R. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the proposition "(object j_1 belongs to the 1st class and object j_2 belongs to the 1st class) or (object j_1 belongs to the 2nd class and object j_2 belongs to the 2nd class) or $\cdots$ or (object j_1 belongs to the kth class and object j_2 belongs to the kth class)", that is:

$$V(j_1,j_2) = (r_{1j_1}\mathrm{t}r_{1j_2})\mathrm{s}\cdots\mathrm{s}(r_{kj_1}\mathrm{t}r_{kj_2}) \triangleq \mathop{\mathrm{S}}_{i=1}^{k}(r_{ij_1}\mathrm{t}r_{ij_2}) \tag{11}$$

  where "s" is a triangular co-norm (s-norm, e.g., max) used as a disjunction to implement the connective "or" of the proposition. This is the fuzzy counterpart of the union operation in (8).
- X: Fuzzy set of pairs of data objects belonging to different classes in R. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the disjunction (connective "or") of the propositions "object j_1 belongs to the i_1th class and object j_2 belongs to the i_2th class" for $i_1,i_2 = 1,\cdots,k$ ($\mathrm{i}_1 \neq i_2$), that is:

$$X(j_1,j_2) = \mathop{\mathrm{S}}_{\substack{i_1,i_2=1\\(i_1\neq i_2)}}^{k}(r_{i_1j_1}\mathrm{t}r_{i_2j_2}) \tag{12}$$

- Y: Fuzzy set of pairs of data objects belonging to the same cluster in Q. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the proposition "(object j_1 belongs to the 1st cluster and object j_2 belongs to the 1st cluster) or $\cdots$ or (object j_1 belongs to the vth cluster and object j_2 belongs to the vth cluster)", that is:

$$Y(j_1,j_2) = \mathop{\mathrm{S}}_{l=1}^{v}(q_{lj_1}\mathrm{t}q_{lj_2}) \tag{13}$$

- Z: Fuzzy set of pairs of data objects belonging to different clusters in Q. The membership degree of each pair of objects (j_1,j_2) is given by the truth value of the disjunction of the propositions "object j_1 belongs to the l_1th cluster and object j_2 belongs to the l_2th cluster" for $l_1,l_2 = 1,\cdots,v$ ($l_1 \neq l_2$), that is:

$$Z(j_1,j_2) = \mathop{\mathrm{S}}_{\substack{l_1,l_2=1\\(l_1\neq l_2)}}^{v}(q_{l_1j_1}\mathrm{t}q_{l_2j_2}) \tag{14}$$

Now, provided that the intersection of two fuzzy sets is generically computed using a t-norm as conjunction and since the cardinality of a fuzzy set is given by the sum of its membership values, it is straightforward to infer from the above definitions that:

$$\begin{aligned}
|V\cap Y| &= \sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{N} V(j_1,j_2)\mathrm{t}Y(j_1,j_2)\\
|V\cap Z| &= \sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{N} V(j_1,j_2)\mathrm{t}Z(j_1,j_2)\\
|X\cap Y| &= \sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{N} X(j_1,j_2)\mathrm{t}Y(j_1,j_2)\\
|X\cap Z| &= \sum_{j_1=1}^{j_2-1}\sum_{j_2=2}^{N} X(j_1,j_2)\mathrm{t}Z(j_1,j_2)
\end{aligned} \tag{15}$$

Eqs. (9), (11)–(15) constitute the Fuzzy Rand Index, which is a fuzzy extension of the original Rand index in (4).

Remark 1. Note that $V(j_1,j_2)$ and $V(j_2,j_1)$ refer to the same element of the fuzzy set V, since (j_1,j_2) and (j_2,j_1) refer to the same pair of data objects j_1 and j_2. The same holds with respect to X, Y, and Z. That's why the summations in (15) are such that $j_1 < j_2$.

Remark 2. It is clear by construction that the fuzzy index comprises the original index as a particular case. In spite of that, one can easily check this property from Eqs. (11)–(15) by using the boundary conditions of t-norms and s-norms (i.e., $0\mathrm{t}\alpha = 0$, $1\mathrm{t}\alpha = \alpha$, $0\mathrm{s}\alpha = \alpha$, and $1\mathrm{s}\alpha = 1$, where $\alpha \in [0,1]$) as well as by recalling that $r_{(\cdot)}$ and $q_{(\cdot)}$ take either 0 or 1 when dealing with the original index.

Remark 3. Since the cardinality of a fuzzy set is always a non-negative number, then it follows from Eq. (9) that the Fuzzy Rand Index keeps the crucial property of normality, i.e., $\omega \in [0,1]$. However, $Q = R$ (the partition under evaluation matches exactly the reference partition) is no longer a necessary and sufficient condition for $\omega = 1$. It is still necessary, but sufficiency is only guaranteed by ensuring that the reference partition R is hard. This has little impact on the practical usefulness of the Fuzzy Rand Index since a reference partition encoding known classes is always hard. Actually, the major impact is that the Fuzzy Rand Index cannot be seen as a general measure for comparing two fuzzy partitions. Instead, it is a measure for comparing a fuzzy partition against a hard partition (possibly with a different number of categories).

Remark 4. The Fuzzy Rand Index is not hooked on any particular class or instance of fuzzy clustering/classification algorithm. In fact, the only requirement is that the algorithm produces as its output a fuzzy partition Q of the data whose elements are such that $q_{ij} \in [0,1]$. It does not matter if the algorithm produces only probabilistic fuzzy partitions $\left(\sum_{i=1}^{v} q_{ij} = 1 \quad \forall j\right)$ or if possibilistic ones are also acceptable. Similarly, it does not matter if only non-degenerate fuzzy partitions $\left(\sum_{j=1}^{N} q_{ij} > 0 \quad \forall i\right)$ can be produced or if degenerated ones are also possible outcomes of the algorithm. Nevertheless, the Fuzzy Rand Index is expected to reflect the appropriateness of the choice of a particular algorithm given a specific problem (data set) in hand. For instance, if the geometric structure contained in the data is such that the natural clusters can be roughly described by hyperellipsoids,[2] a fuzzy clustering algorithm capable of finding hyperellipsoidal clusters (e.g., the Gustafson–Kessel algorithm (Gustafson and Kessel, 1979)) is expected to provide much more accurate results than those provided by an algorithm equipped with a fixed Euclidean distance norm (e.g., the traditional Fuzzy C-Means (Bezdek, 1981)). In this case, if a reference partition R is available for comparisons, the more accurate results of the Gustafson–Kessel algorithm will be characterized by higher values of the Fuzzy Rand Index.

Before proceeding with an example, a set of closely related indexes and their fuzzy counterparts are described in the sequel based strictly on the same concepts discussed above.

3.3. Related indexes

One of the main criticisms against the original Rand index is that it is not *corrected for chance*, that is, its expected value is not zero when comparing random partitions. Correcting an index for chance means normalizing it so that its (expected) value is 0 when the partitions are selected by chance and 1 when a perfect match is achieved (Jain and Dubes, 1988). Hubert and Arabie derived the following Adjusted Rand Index (Hubert and Arabie, 1985):

$$\omega_A = \frac{a - \frac{(a+c)(a+b)}{d}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{d}} \tag{16}$$

which is corrected for chance under the assumption that the number of groups (classes/clusters) in both partitions R and Q to be compared is the same.

Another common criticism against the original Rand index is that it gives the same importance to both the agreement terms a and d, thus making no difference between pairs of objects that are joined or separated in both partitions R and Q (Saporta and Youness, 2002; Youness and Saporta, 2004). This policy is arguable, particularly if a partition is interpreted as a set of groups of joined elements, the separations being just consequences of the grouping procedure (Denoeud et al., 2005). This interpretation suggests that term d should be removed from the formulation of the Rand index. Indeed, it is known that this term may dominate the other three terms (a, b, and c), thus causing the Rand index to become unable to properly distinguish between good and bad partitions (Sharan and Shamir, 2000; Jiang et al., 2004). This situation gets particularly critical as the number of classes/clusters increases, since the value of the index tends to increase too (Fowlkes and Mallows, 1983; Wallace, 1983; Everitt et al., 2001). An interesting example is described in (Pantel, 2003, p. 67).

The ordinary removal of term d from the original Rand index in Eq. (4) results in the so-called Jaccard coefficient, given by (Jain and Dubes, 1988; Halkidi et al., 2001):

$$\omega_J = \frac{a}{a+b+c} \tag{17}$$

Clearly, the rationale behind the Jaccard coefficient in (17) is essentially the same as that for the Rand index, except for the absence of term d (which does not affect normality, i.e., $\omega_J \in [0,1]$). An interesting interpretation of the differences between these two indexes arises when d is viewed as a "neutral" term – counting pairs of objects that are not clearly indicative either of similarity or of inconsistency – in contrast to the others, viewed as counts of "good pairs" (term a) and "bad pairs" (terms b and c) (Wallace, 1983). From this viewpoint, the Jaccard coefficient can be seen as a proportion of good pairs with respect to the sum of non-neutral (good plus bad) pairs, whereas the Rand index is just the proportion of pairs not definitely bad with respect to the total number of pairs.

The absence of term d is also observed in the following closely related indexes:

$$\omega_F = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{18}$$

$$\omega_M = \sqrt{\frac{b+c}{b+a}} \tag{19}$$

named Fowlkes–Mallows index (Fowlkes and Mallows, 1983; Jain and Dubes, 1988) and Minkowski measure (Sharan and Shamir, 2000; Jiang et al., 2004), respectively. The Fowlkes–Mallows index in (18) can be seen as a non-linear modification of the Jaccard coefficient that also keeps normality (i.e., $\omega_F \in [0,1]$). The Minkowski measure given by Eq. (19), in its turn, is such that the lower its value the better the matching between the partitions ($\omega_M = 0$ in case of a perfect match). This index essentially measures the proportion of disagreements between the partitions to the total number of pairs of objects in the same class of the reference partition.

Another related index is the Γ statistics, given by (Jain and Dubes, 1988):

$$\omega_\Gamma = \frac{Ma - (a+b)(a+c)}{\sqrt{(a+b)(a+c)(M-(a+b))(M-(a+c))}} \tag{20}$$

[2] This is usually the case, e.g., in function approximation applications (Babuška, 1998).

where $M = N(N-1)/2$ is the total number of non-ordered pairs of objects in the data set. Note that the Γ statistics in (20) is a correlation ($\omega_\Gamma \in [-1,1]$) in which term d would appear only implicitly if M was equivalently rewritten as $M = a + b + c + d$.

Some of the above-mentioned indexes, namely, the Adjusted Rand Index, the Jaccard coefficient, and the Fowlkes–Mallows index, have been experimentally shown in (Denoeud et al., 2005),[3] from a statistical perspective, to be more effective than the original Rand index in measuring the closeness between two partitions. As a referential measure of closeness, the authors adopted the minimum number of transfers of objects between groups needed to turn one of the partitions into the other one (called *transfer distance*), which can be computed by solving a weighted matching problem in a complete bipartite graph. In that particular study, the Adjusted Rand Index, the Jaccard coefficient, and the Fowlkes–Mallows index exhibited approximately the same average performance, with the Jaccard coefficient outperforming the others in terms of variance.

Remark 5. The external indexes in Eqs. (16)–(20) can be straightforwardly extended to the fuzzy domain by rewriting terms a, b, c, and d as in (7), with fuzzy sets V, X, Y, and Z as defined in Section 3.2.

3.4. Algorithm

A simple three-step algorithm for computing the fuzzy indexes discussed above is described in the sequel. Given a data set with N objects, a reference partition of these data into k known classes, $R = [r_{ij}]_{k \times N}$, and a fuzzy partition of the same data into v clusters/classes to be evaluated, $Q = [q_{ij}]_{v \times N}$, then:

1. For every $j_1, j_2 = 1, \cdots N$ with $j_1 < j_2$ compute $V(j_1,j_2)$, $X(j_1,j_2)$, $Y(j_1,j_2)$, and $Z(j_1,j_2)$ using Eqs. (11)–(14), respectively.
2. Compute terms a, b, c, and d, as defined in (7), using (15).
3. Compute the desired fuzzy index using the corresponding equation, i.e.: (4) (Rand index); or (16) (Adjusted Rand Index); or (17) (Jaccard coefficient); or (18) (Fowlkes–Mallows index); or (19) (Minkowski measure); or (20) (Γ statistics).

3.5. Example

The same data set displayed in Fig. 1 is used here again for the purpose of illustration. The data classes are considered to be the same as those shown in the figure (circles and squares), which refer to the partition matrix R in (5). However, it is assumed now that the data has been partitioned into the following two fuzzy partition matrices Q_1 and Q_2 to be assessed:

$$Q_1 = \begin{bmatrix} 0.99 & 0.94 & 0.90 & 0.08 & 0.91 & 0.01 & 0.02 & 0 \\ 0.09 & 0.08 & 0.03 & 0.90 & 0.02 & 0.96 & 0.91 & 0.97 \end{bmatrix} \quad (21)$$

$$Q_2 = \begin{bmatrix} 0.59 & 0.54 & 0.50 & 0.48 & 0.51 & 0.41 & 0.42 & 0.40 \\ 0.49 & 0.48 & 0.43 & 0.50 & 0.42 & 0.56 & 0.51 & 0.57 \end{bmatrix} \quad (22)$$

These two particular partitions were intentionally selected because they have some interesting properties: (i) partition Q_1 is close to the reference partition R, whereas Q_2 is not (except for the number of clusters/classes); (ii) inspite of that, Q_1 and Q_2 share R as the hard partition that results from assigning 1 to the highest membership value of each data object and 0 to the others; and (iii) unlike Q_2, partition Q_1 indicates that there is an unequivocal geometrical structure contained in the data, which is similar to that represented by the reference partition. The first and third properties suggest that the Fuzzy Rand Index for Q_1 should be close to 1 and significantly greater than that for Q_2. The second property implies that, despite their considerable differences, Q_1 and Q_2 would be equally scored if the original (hard) Rand index were used.

The Fuzzy Rand Indexes for partitions Q_1 and Q_2 are then computed using the algorithm described in Section 3.4 as $\omega(Q_1, R) = 0.9364$ and $\omega(Q_2, R) = 0.5267$, with the t and s norms implemented as t = min and s = max, respectively. Keeping the max operator yet adopting the algebraic product in place of the min operator slightly changes these values to $\omega(Q_1, R) = 0.9379$ and $\omega(Q_2, R) = 0.5337$. These results are in conformity with the expectations discussed above.

Nevertheless, one could argue that Q_2 is a better representation of the data set in Fig. 1 if the class labels are ignored, i.e., if only the (nearly uniform) spatial distribution of the data is taken into account. For this reason, it is important to stress that the fuzzy versions of the Rand and other related indexes discussed in this paper are external validity criteria and, as such, they are particularly useful when a classification perspective is convenient, namely: (i) when assessing classifiers with fuzzy-like outputs; (ii) when assessing fuzzy clustering partitions obtained as the first step for the further design of a classifier, i.e., when classification is the final goal of clustering; and (iii) when assessing fuzzy clustering partitions using classification data sets under the assumption that the known classes are good representatives of the geometrical structures sought in the data.

4. Conclusions and future work

A fuzzy extension of the Rand index for clustering and classification assessment has been derived. The original

[3] The Fowlkes–Mallows index is referred to as Wallace index in (Denoeud et al., 2005), as a particular case of the more general (asymmetrical) formulation suggested by Wallace (Wallace, 1983).

Rand index has been extended to the fuzzy domain by making it able to evaluate a fuzzy partition of a data set into v categories (clusters or classes) against a reference partition encoding a set of k known data classes. The extended index has been obtained by first rewriting the formulation of the original index in a fully equivalent form by using basic concepts from the set theory. This equivalent formulation has then been extended to the fuzzy domain by using analogous concepts from the fuzzy set theory. The fuzzy counterparts of five related indexes, namely, the Adjusted Rand Index of Hubert and Arabie, the Jaccard coefficient, the Minkowski measure, the Fowlkes–Mallows index, and the Γ statistics, have also been derived from the same methodology. It is worth noticing that the need to a fuzzy extension of the original Rand index had already been mentioned in the literature (e.g., see (Corney, 2002)). However, to the best of the author's knowledge, no formulation had so far been formally presented and analyzed.

The fuzzy extensions of the Rand index and related indexes mentioned above have been shown to be able to get around the information loss that is unavoidable when using their original formulations to evaluate fuzzy partitions. So, the real-world applicability of these fuzzy extensions is kept essentially the same as that for the originals, but further improved since the fuzzy versions can be applied to fuzzy partitions without any loss of information. Particularly, the fuzzy indexes are favorable for evaluating fuzzy partitions from a classification perspective, that is: (i) when assessing classifiers with fuzzy-like outputs; (ii) when assessing fuzzy clustering partitions obtained as the first step for the further design of a classifier, i.e., when classification is the final goal of clustering; and (iii) when assessing fuzzy clustering partitions using classification data sets under the assumption that the known classes are good representatives of the geometrical structures sought in the data. In addition, these indexes allow the use of classification benchmarks for fuzzy clustering assessment and, as such, they become eligible to be adopted as absolute criteria for comparing the performances of relative fuzzy cluster validity measures.

In principle, the methodology presented in this paper can be used to extend – to the fuzzy domain – any index that can be written according to a set-theoretic formulation. However, there are some indexes that cannot be represented this way, such as those based on the concept of number of transfers of objects between groups needed to turn a partition into another. This is the case, for instance, of the *editing distance* proposed in (Pantel, 2003). Another example is the *transfer distance*, which was previously mentioned in Section 3.3 for it has been used in (Denoeud et al., 2005) as a reference for a statistical comparison of external indexes. The extension of these "transfer-based" indexes to the fuzzy domain is still an open problem to be worked out. In particular, a fuzzy version of the transfer distance would be useful to generalize the statistical procedure adopted in (Denoeud et al., 2005) so that it can also be utilized for comparisons of the fuzzy indexes derived here. In future work, the author intends to follow this direction and perform comparisons of the fuzzy indexes proposed, including comparisons in real-world scenarios.

Acknowledgments

This work was supported in part by the Brazilian National Research Council – CNPq (under Grant no. #307554/2003-1) and also by the Research Foundation of the State of São Paulo – Fapesp (under Grant no. #06/50231-5).

Appendix

The subsets V_i, $X_{i_1 i_2}$, Y_l, and $Z_{l_1 l_2}$ for the example displayed in Fig. 1, as defined in Section 3.1, are given by:

$$V_1 = \{(1,2),(1,3),(1,5),(2,3),(2,5),(3,5)\}$$
$$V_2 = \{(4,6),(4,7),(4,8),(6,7),(6,8),(7,8)\}$$
$$X_{12} = \{(1,4),(1,6),(1,7),(1,8),(2,4),(2,6),(2,7),(2,8),(3,4),(3,6),(3,7),(3,8),(4,5),(5,6),(5,7),(5,8)\}$$
$$Y_1 = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4)\}$$
$$Y_2 = \{(5,6)\}$$
$$Y_3 = \{(7,8)\}$$
$$Z_{12} = \{(1,5),(1,6),(2,5),(2,6),(3,5),(3,6),(4,5),(4,6)\}$$
$$Z_{13} = \{(1,7),(1,8),(2,7),(2,8),(3,7),(3,8),(4,7),(4,8)\}$$
$$Z_{23} = \{(5,7),(5,8),(6,7),(6,8)\}$$

Then, using (8) yields:

$$V = \{(1,2),(1,3),(1,5),(2,3),(2,5),(3,5),(4,6),(4,7),(4,8),(6,7),(6,8),(7,8)\}$$
$$X = \{(1,4),(1,6),(1,7),(1,8),(2,4),(2,6),(2,7),(2,8),(3,4),(3,6),(3,7),(3,8),(4,5),(5,6),(5,7),(5,8)\}$$
$$Y = \{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4),(5,6),(7,8)\}$$
$$Z = \{(1,5),(1,6),(2,5),(2,6),(3,5),(3,6),(4,5),(4,6),(1,7),(1,8),(2,7),(2,8),(3,7),(3,8),(4,7),(4,8),(5,7),(5,8),(6,7),(6,8)\}$$

and, accordingly:

$$V \cap Y = \{(1,2),(1,3),(2,3),(7,8)\}$$
$$V \cap Z = \{(1,5),(2,5),(3,5),(4,6),(4,7),(4,8),(6,7),(6,8)\}$$
$$X \cap Y = \{(1,4),(2,4),(3,4),(5,6)\}$$
$$X \cap Z = \{(1,6),(1,7),(1,8),(2,6),(2,7),(2,8),(3,6),(3,7),(3,8),(4,5),(5,7),(5,8)\}$$

which results in $a = |V \cap Y| = 4$, $b = |V \cap Z| = 8$, $c = |X \cap Y| = 4$, $d = |X \cap Z| = 12$, and a Rand index of $\omega = 0.5714$, as previously mentioned in Section 2.

References

Babuška, R., 1998. Fuzzy Modeling for Control. Kluwer Academic Publishers.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithm. Plenum Press.

Bezdek, J.C., Pal, N.R., 1998. Some new indexes of cluster validity. IEEE Trans. Systems Man Cybernet-B 28 (3), 301–315.

Bigus, J.P., 1996. Data Mining with Neural Networks: Solving Business Problems from Application Development to Decision Support. McGraw-Hill.

Campello, R.J.G.B., Hruschka, E.R., 2006. A fuzzy extension of the silhouette width criterion for cluster analysis. Fuzzy Sets Syst. 157, 2858–2875.

Corney, D.P.A., 2002. Intelligent Analysis of Small Data Sets for Food Design, PhD thesis, Department of Computer Science, University College, London, UK.

Denoeud, L., Garreta, H., Guénoche, A., 2005. Comparison of distance indices between partitions, in: Proceedings of the 11th Conference of the Applied Stochastic Models and Data Analysis (ASMDA), Brest/France.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. Wiley.

Dumitrescu, D., Lazzerini, B., Jain, L.C., 2000. Fuzzy Sets and their Application to Clustering and Training. CRC Press.

Everitt, B.S., Landau, S., Leese, M., 2001. Cluster Analysis, fourth ed. Arnold.

Fowlkes, E.B., Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. J. Amer. Statist. Assoc. 78, 553–569.

Gustafson, E.E., Kessel, W.C., 1979. Fuzzy clustering with a fuzzy covariance matrix. in: IEEE Conference on Decision and Control, San Diego/USA, pp. 761–766.

Halkidi, M., Batistakis, Y., Vazirgiannis, M., 2001. On clustering validation techniques. J. Intell. Inform. Syst. 17, 107–145.

Höppner, F., Klawonn, F., Kruse, R., Runkler, T., 1999. Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition. John Wiley & Sons.

Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall.

Jiang, D., Tang, C., Zhang, A., 2004. Cluster analysis for gene-expression data: a survey. IEEE Trans. Knowledge Data Eng. 16, 1370–1386.

Kaufman, L., Rousseeuw, P., 1990. Finding Groups in Data. Wiley.

Kosko, B., 1997. Fuzzy Engineering. Prentice Hall.

Mitchell, T.M., 1997. Machine Learning. McGraw Hill.

Newman, D.J., Hettich, C.B.S., Merz, C., 1998. UCI repository of machine learning databases. University of California, Irvine, Department of Information and Computer Sciences – http://www.ics.uci.edu/~mlearn/MLRepository.html.

Pal, N.R., Bezdek, J.C., 1995. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst. 3 (3), 370–379.

Pantel, P.A., 2003. Clustering by Committee, PhD thesis, Faculty of Graduate Studies and Research – University of Alberta.

Pedrycz, W., Gomide, F.A.C., 1998. An Introduction to Fuzzy sets. Analysis and Design. MIT Press.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Amer. Statist. Assoc., 846–850.

Saporta, G., Youness, G., 2002. Comparing two partitions: some proposals and experiments. In: Härdle, W. (Ed.), Proc. Comput. Statist.. Physica-Verlag.

Sharan, R., Shamir, R., 2000. CLICK: a clustering algorithm with applications to gene expression analysis, in: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, San Diego/USA, pp. 307–316.

Smith, M., 1993. Neural Networks for Statistical Modeling. International Thomson Computer Press.

Wallace, D.L., 1983. Comment on a method for comparing two hierarchical clusterings. J. Amer. Statist. Assoc., 569–576.

Youness, G., Saporta, G., 2004. Une méthodologie pour la comparaison de partitions. Revue de Statistique Appliquée, 97–120.

Zimmermann, H.J., 2001. Fuzzy Set Theory and its Applications, fourth ed. Kluwer Academic Publishers.