A Fuzzy Variant of the Rand Index for Comparing Clustering Structures

Eyke Hullermeier, Maria Rifqi

A Fuzzy Variant of the Rand Index for Comparing Clustering Structures

Eyke Hüllermeier[1] and Maria Rifqi[2]

1. FB Mathematik/Informatik, Philipps-Universität Marburg
D-35032 Marburg, Germany
2. University Pierre et Marie Curie–Paris 6, CNRS UMR 7606, LIP6,
Paris, F-75016, France
Email: eyke@mathematik.uni-marburg.de, maria.rifqi@lip6.fr

Abstract— *In this paper, we introduce a fuzzy extension of the Rand index, a well-known measure for comparing two clustering structures. In contrast to an existing proposal, which is restricted to the comparison of a fuzzy partition with a non-fuzzy reference partition, our extension is able to compare two proper fuzzy partitions with each other. Elaborating on the formal properties of our fuzzy Rand index, we show that it exhibits desirable metrical properties.*

Keywords— Clustering, Distance, Fuzzy Partition, Metric, Rand Index, Similarity

1 Introduction

The problem to compare two partitions of a set of objects occurs quite naturally in various domains, notably in data analysis and clustering. For example, one way to evaluate the result of a clustering algorithm is to compare the clustering structure produced by the algorithm with a correct partition of the data (which of course presumes that this information is available). In cluster analysis, so called *external evaluation measures* have been developed for this purpose [1, 2]. However, measures of that kind are not only of interest as evaluation criteria, i.e., for comparing a hypothetical partition with a true one. Instead, distance measures for partitions are interesting in their own right and can be used for different purposes.

In [3], for example, the authors consider the problem of clustering data in a very high-dimensional space. To increase efficiency, they propose to map the data into a low-dimensional space first and to cluster the transformed data thus obtained afterward. In this context, a distance measure for clustering structures (partitions) is useful to measure the loss of information incurred by the data transformation: If the transformation is (almost) lossless, the clustering structures in the two spaces should be highly similar, i.e., their distance should be small. On the other hand, a significant difference between the two partitions would indicate that the transformation does have a strong effect in the sense of distorting the structure of the data set.

Even though a large number of evaluation criteria and similarity indexes for clustering structures have been proposed in the literature, their extension to the case of fuzzy partitions has received much less attention so far. This is especially true for external evaluation criteria and measures comparing two clustering structures, whereas *internal criteria* for evaluating a single partition[1] have been studied more thoroughly (see, e.g., [4] and [5] for early proposals).

In a recent paper by Campello [6], the author has proposed an extension of the *Rand index* [7], a well-known measure of similarity between two partitions of a data set. Even though Campello's proposal is quite interesting, it also exhibits a number of disadvantages. Most notably, it is properly defined only for the comparison of a fuzzy partition with a non-fuzzy reference partition. It is true that this restriction can be tolerated if the index is used as an external evaluation criterion since, as correctly argued by the author, a reference partition provided by an external source is typically non-fuzzy. Yet, our example above has clearly shown that there is also a need for measures comparing two fuzzy partitions.

In this paper, we propose an alternative extension of the Rand index (which is, in principle, also applicable to related similarity measures for clustering structures). As opposed to Campello's proposal, our variant is able to compare two proper fuzzy partitions with each other. Moreover, we study our fuzzy Rand index from a formal point of view and show that it satisfies the desirable properties of a metric (when being used as a distance function).

The remainder of the paper is organized as follows. In the next section, we briefly review the proposal of Campello and discuss some of its properties in a critical way. In Section 3, we introduce our new measure and elaborate on its formal properties. The paper concludes with a short summary and an outlook on future work in Section 4.

2 Review of Campello's Proposal

Before reviewing Campello's proposal and discussing some of its properties, we briefly recall the original definition of the Rand index.

2.1 The Rand Index

Let $\mathbf{P} = \{P_1, \ldots, P_k\} \subset 2^X$ and $\mathbf{Q} = \{Q_1, \ldots, Q_\ell\} \subset 2^X$ be two (crisp) partitions of a finite set $X = \{x_1, x_2 \ldots x_n\}$ with n elements, which means that $P_i \neq \emptyset$, $P_i \cap P_j = \emptyset$ for all $1 \leq i \neq j \leq k$, and $P_1 \cup P_2 \cup \ldots \cup P_k = X$ (and analogously for $\mathbf{Q}$). Let

$$C = \{ (x_i, x_j) \in X \times X \mid 1 \leq i < j \leq n \}$$

denote the set of all tuples of elements in X.[2] We say that two elements $(x, x') \in C$ are *paired* in $\mathbf{P}$ if they belong to the same

[1] Typically, such criteria compare the intra-cluster variability, i.e., the variability among objects within the same cluster (which should be small) with the inter-cluster variability, i.e., the variability among objects from different clusters (which should be high).

[2] Since we consider unordered tuples, we should more correctly write $\{x_i, x_j\}$ instead of (x_i, x_j).

cluster, i.e., if there is a cluster $P_i \in \mathbf{P}$ such that $x \in P_i$ and $x' \in P_i$. Moreover, we distinguish the following subsets of C:

- $C_1 \equiv$ the set of tuples $(x,x') \in C$ that are paired in $\mathbf{P}$ and paired in $\mathbf{Q}$;
- $C_2 \equiv$ the set of tuples $(x,x') \in C$ that are paired in $\mathbf{P}$ but not paired in $\mathbf{Q}$;
- $C_3 \equiv$ the set of tuples $(x,x') \in C$ that are not paired in $\mathbf{P}$ but paired in $\mathbf{Q}$;
- $C_4 \equiv$ the set of tuples $(x,x') \in C$ that are neither paired in $\mathbf{P}$ nor in $\mathbf{Q}$.

Obviously, $\{C_1,C_2,C_3,C_4\}$ is a partition of C, and $a+b+c+d = |C| = n(n-1)/2$, where

$$a = |C_1|,\ b = |C_2|,\ c = |C_3|,\ d = |C_4|. \tag{1}$$

The tuples $(x,x') \in C_1 \cup C_4$ are the *concordant* pairs, i.e., the pairs for which there is agreement between $\mathbf{P}$ and $\mathbf{Q}$, while the tuples $(x,x') \in C_2 \cup C_3$ are the *discordant* pairs for which the two partitions disagree. The Rand index is defined by the number of concordant pairs divided by the total number of pairs:

$$\mathrm{R}(\mathbf{P},\mathbf{Q}) = \frac{a+d}{a+b+c+d} \tag{2}$$

Thus defined, the Rand index is a similarity measure which assumes values between 0 and 1. It can easily be turned into a distance function by defining

$$\mathrm{D}_R(\mathbf{P},\mathbf{Q}) = 1 - \mathrm{R}(\mathbf{P},\mathbf{Q}) = \frac{b+c}{a+b+c+d}\ .$$

It is worth mentioning that D_R satisfies the classical properties of a distance (reflexivity, separation, symmetry, and triangular inequality).

2.2 Campello's Fuzzy Rand Index

The aim of Campello's paper is to extend the Rand index to the case of fuzzy partitions. To this end, he first reformulates it within a set-theoretic framework. An extension to the fuzzy case can then be accomplished in a straightforward way by using generalized set-theoretical operators. Recall that $k = |\mathbf{P}|$ and $\ell = |\mathbf{Q}|$, and consider the following sets:

- $V \equiv$ the set of pairs $(x,x') \in C$ that belong to the same cluster in $\mathbf{P}$; it can be expressed as $V = \bigcup_{i=1\ldots k} V_i$, where V_i is the set of pairs that both belong to the i-th cluster $P_i \in \mathbf{P}$.
- $W \equiv$ the set of pairs $(x,x') \in C$ that belong to different clusters in $\mathbf{P}$; it can be expressed as $W = \bigcup_{1\le i\ne j\le k} W_{ij}$, where W_{ij} is the set of pairs such that $x \in P_i$ and $x' \in P_j$.
- $Y \equiv$ the set of pairs $(x,x') \in C$ that belong to the same cluster in $\mathbf{Q}$; it can be expressed as $Y = \bigcup_{i=1\ldots\ell} Y_i$, where Y_i is the set of pairs that both belong to the i-th cluster $Q_i \in \mathbf{Q}$.
- $Z \equiv$ the set of pairs $(x,x') \in C$ that belong to different clusters in $\mathbf{Q}$; it can be expressed as $Z = \bigcup_{1\le i\ne j\le \ell} Z_{ij}$, where Z_{ij} is the set of pairs such that $x \in Q_i$ and $x' \in Q_j$.

The Rand index can directly be written in terms of the cardinalities of these sets, since the four quantities (1) are obviously given by

$$a = |V \cap W|, \qquad b = |V \cap Z|,$$
$$c = |W \cap Y|, \qquad d = |W \cap Z|.$$

In the fuzzy case, the above sets become fuzzy sets. Let $P_i(x) \in [0,1]$ denote the degree of membership of element $x \in X$ in the cluster $P_i \in \mathbf{P}$. The sets V, W, Y, and Z can then be defined through fuzzy-logical expressions involving a t-norm $\top$ and t-conorm $\bot$:

$$\begin{aligned} V(x,x') &= \bot_{i=1}^{k} \top(P_i(x),P_i(x')) \\ W(x,x') &= \bot_{1\le i\ne j\le k} \top(P_i(x),P_j(x')) \\ Y(x,x') &= \bot_{i=1}^{\ell} \top(Q_i(x),Q_i(x')) \\ Z(x,x') &= \bot_{1\le i\ne j\le \ell} \top(Q_i(x),Q_j(x')) \end{aligned} \tag{3}$$

Moreover, defining the intersection of sets by the t-norm combination of membership degrees and resorting to the commonly used sigma-count principle [8] for defining set cardinality, one obtains

$$\begin{array}{rcccl} a & = & |V \cap Y| & = & \sum_{(x,x')\in C} \top(V(x,x'),Y(x,x')) \\ b & = & |V \cap Z| & = & \sum_{(x,x')\in C} \top(V(x,x'),Z(x,x')) \\ c & = & |W \cap Y| & = & \sum_{(x,x')\in C} \top(W(x,x'),Y(x,x')) \\ d & = & |W \cap Z| & = & \sum_{(x,x')\in C} \top(W(x,x'),Z(x,x')) \end{array} \tag{4}$$

As before, the Rand index can then be defined as in (2), namely as the fraction

$$\frac{a+d}{a+b+c+d}\ .$$

2.3 Properties of Campello's Fuzzy Rand Index

Having defined a similarity or, equivalently, a distance function, it is natural to ask for desirable metrical properties of that function. In the case of the above fuzzy Rand index, however, this question has to be considered with caution, since Campello is actually only interested in comparing a fuzzy partition $\mathbf{P}$ with a non-fuzzy partition $\mathbf{Q}$. And indeed, formal properties of the measure are not investigated in his paper.

On the other hand, it is noted by Campello himself that, formally, the measure can in principle be applied to compare two fuzzy partitions. When doing so, however, it turns out quickly that it fails to be a proper metric. In fact, it does not even satisfy reflexivity, the perhaps most basic axiom: Even for two identical partitions $\mathbf{P}$ and $\mathbf{Q}$, the quantities b and c in (4) will generally not vanish, a necessary condition for having $\mathrm{R}(\mathbf{P},\mathbf{Q}) = 1$.

Consider, for example, the simple fuzzy partition $\mathbf{P}$ illustrated in Fig. 1, which consists of two clusters P_1 and P_2. Instead of a hard boundary, there is a "soft" transition between P_1 and P_2; the elements x_1, x_2, x_3, and x_4 partially belong to both clusters and have membership degrees, respectively, of $3/4$, $1/2$, $1/2$, $1/4$ in P_1 and $1/4$, $1/2$, $1/2$, $3/4$ in P_2. Comparing $\mathbf{P}$ to itself in terms of the fuzzy Rand index, we obtain $\mathrm{R}(\mathbf{P},\mathbf{P}) < 1$.

Upon closer examination, it seems that the core principle of Campello's extension is not suitable for comparing partitions in a fuzzy sense. In fact, despite being defined in terms

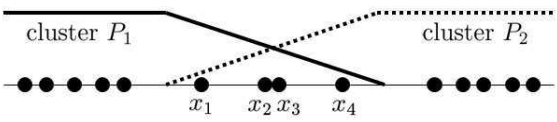

Figure 1: Illustration of a simple fuzzy partition of a subset of the reals (indicated by circles). The partition consists of two clusters, P_1 (left) and P_2 (right). While some elements definitely belong to only one of the clusters, some "critical" points in the middle have partial membership in both clusters.

of fuzzy logical formulas, the very idea of the approach has arguably a more "probabilistic flavor". This becomes especially obvious when using the product as a t-norm and the (bounded) sum as t-conorm. Then, if $P_i(x)$ is interpreted as the probability that x belongs to the i-th cluster, $V(x,x')$ is nothing else than the probability that x and x' are put in the same cluster, given that the two corresponding clusters are chosen *independently of each other* according to the distributions $(P_1(x),P_2(x)\dots P_k(x))$ and $(P_1(x'),P_2(x')\dots P_k(x'))$, respectively. Likewise, $W(x,x')$ is the probability that x and x' are put into different clusters.

Even if one accepts the probabilistic interpretation of a single membership degree, the additional assumption of independence is clearly not tenable. In fact, this property is obviously violated when comparing a partition with itself, since for each element $x \in X$, a cluster can then only be chosen once and not two times independently of each other. But even if $\mathbf{P}$ and $\mathbf{Q}$ are not identical, independence of cluster membership is in conflict with the topological relationships between the elements and clusters. In the example in Fig. 1, for instance, it is not reasonable to put x_1 and x_4 into cluster P_2 and x_2 and x_3 into cluster P_1. When putting elements independently of each other into clusters, however, this is a possible scenario. And indeed, this scenario contributes to Campello's fuzzy Rand index according to (3).

Seen from this point of view, one may even question the usefulness of the approach for its original purpose, namely the comparison of a fuzzy with a non-fuzzy partition. What the fuzzy partition in our example truly suggests is that we are uncertain about the boundary between the two clusters. More concretely, the fuzzy partition suggests four possible non-fuzzy partitions:

- $\mathbf{P}_1$ which puts the boundary left to x_1;
- $\mathbf{P}_2$ with boundary between x_1 and x_2;
- $\mathbf{P}_3$ with boundary between x_3 and x_4;
- $\mathbf{P}_4$ which puts the boundary right to x_4.

Thus, it seems reasonable to define an extension of the Rand index as an aggregation (e.g., weighted average) of the results of the non-fuzzy comparisons, namely

$$\mathrm{R}(\mathbf{P}_1,\mathbf{Q}),\mathrm{R}(\mathbf{P}_2,\mathbf{Q}),\mathrm{R}(\mathbf{P}_3,\mathbf{Q}),\mathrm{R}(\mathbf{P}_4,\mathbf{Q}).$$

In Campello's approach, there are not 4 but 16 scenarios which have an influence on the result, since each of the four cluster memberships is determined independently of each other. In general, the result will therefore be different. In fact, differences already occur for single pairs of elements. For example, since x_2 and x_3 are always in the same cluster in $\mathbf{P}_1,\dots,\mathbf{P}_4$, it is natural to say that they are paired with degree 1. According to Campello's approach, however, the degree to which x_2 and x_3 are in the same cluster in $\mathbf{P}$ is given by

$$V(x_2,x_3) = \bot(\top(1/2,1/2),\top(1/2,1/2)),$$

which corresponds to the truth degree of the proposition that "x_1 is put into P_1 AND x_2 is put into P_1 OR x_1 is put into P_2 AND x_2 is put into P_2". In general, this degree will be < 1 (except for special $(\top,\bot)$-combinations such as $\top = \min$ and $\bot =$ bounded sum).

3 A New Fuzzy Rand Index

In this section, we propose a new fuzzy variant of the Rand index which is able to compare any pair of fuzzy partitions and, moreover, has desirable metric properties. In the following, we focus on the view of the Rand index as a distance function. Thanks to the affine transformation $\mathrm{D}_R = 1-\mathrm{R}$, all results can directly be transferred to the original conception as a measure of similarity.

3.1 Definition

Given a fuzzy partition $\mathbf{P} = \{P_1,P_2\dots P_k\}$ of X, each element $x \in X$ can be characterized by its membership vector

$$\mathbf{P}(x) = (P_1(x),P_2(x)\dots P_k(x)) \in [0,1]^k, \tag{5}$$

where $P_i(x)$ is the degree of membership of x in the i-th cluster P_i. We define a fuzzy equivalence relation on X in terms of a similarity measure on the associated membership vectors (5). Generally, this relation is of the form

$$E_{\mathbf{P}}(x,x') = 1 - \|\mathbf{P}(x) - \mathbf{P}(x')\|, \tag{6}$$

where $\|\cdot\|$ is a proper distance on $[0,1]^k$. The basic requirement on this distance is that it yields values in $[0,1]$. The relation (6) generalizes the equivalence relation induced by a conventional partition (where each cluster forms an equivalence class). In passing, we note that this definition is invariant toward a permutation (renumbering) of the clusters in $\mathbf{P}$, which is clearly a desirable property.

Now, given two fuzzy partitions $\mathbf{P}$ and $\mathbf{Q}$, the idea is to generalize the concept of concordance as follows. We consider a pair (x,x') as being concordant in so far as $\mathbf{P}$ and $\mathbf{Q}$ agree on their degree of equivalence. This suggest to define the *degree of concordance* as

$$1 - |E_{\mathbf{P}}(x,x') - E_{\mathbf{Q}}(x,x')| \in [0,1]. \tag{7}$$

Analogously, the *degree of discordance* is

$$|E_{\mathbf{P}}(x,x') - E_{\mathbf{Q}}(x,x')| \ .$$

Our distance measure on fuzzy partitions is then defined by the normalized sum of degrees of discordance:

$$d(\mathbf{P},\mathbf{Q}) = \frac{\sum_{(x,x')\in C} |E_{\mathbf{P}}(x,x') - E_{\mathbf{Q}}(x,x')|}{n(n-1)/2} \tag{8}$$

Likewise,

$$1 - d(\mathbf{P},\mathbf{Q}) \tag{9}$$

corresponds to the normalized degree of concordance and, therefore, is a direct generalization of the original Rand index.

3.2 Formal Properties

In this section, we first show that our proposal is indeed a proper generalization of the Rand index. Afterward, we study the metrical properties of the measure.

Proposition: *In the case where* $\mathbf{P}$ *and* $\mathbf{Q}$ *are non-fuzzy partitions, the measure (9) reduces to the original Rand index.*

Proof: In the non-fuzzy case, the membership vectors (5) are 0/1-vectors. More specifically, each vector has a single entry $P_i(x) = 1$, while all other entries are 0. Consequently, the fuzzy equivalence (6) reduces to the conventional equivalence, that is, $E_{\mathbf{P}}(x,x') = 1$ if x and x' are in the same cluster and $E_{\mathbf{P}}(x,x') = 0$ otherwise. Likewise, (7) yields 1 if (x,x') is a concordant pair and 0 otherwise. Consequently, the measure (9) is the (normalized) sum of concordant pairs and, therefore, equals the original Rand index. □

Recall that a non-negative $U^2 \to \mathbb{R}$ mapping $d(\cdot)$ is called a metric on U if is satisfies the following properties for all $u,v,w \in U$:

- Reflexivity: $d(u,u) = 0$
- Separation: $d(u,v) = 0$ implies $u = v$
- Symmetry: $d(u,v) = d(v,u)$
- Triangle inequality: $d(u,w) \leq d(u,v) + d(v,w)$

The properties of reflexivity and symmetry are quite obviously valid for our measure (8). To show the triangle inequality, consider three fuzzy partitions $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{R}$ and fix a single tuple $(x,x') \in C$. Let

$$a = E_{\mathbf{P}}(x,x'),\ b = E_{\mathbf{Q}}(x,x'),\ c = E_{\mathbf{R}}(x,x').$$

Since a, b, and c are real numbers (from the unit interval), and the simple difference on the reals satisfies the triangle inequality, we have $|a-c| \leq |a-b| + |b-c|$. Now, since this inequality holds for each pair $(x,x') \in C$, it remains valid when summing over all these pairs. In other words, it is also satisfied by (8), which means that

$$d(\mathbf{P},\mathbf{R}) \leq d(\mathbf{P},\mathbf{Q}) + d(\mathbf{Q},\mathbf{R}).$$

The separation property is not immediately valid for (8). Roughly speaking, this is due to the fact that, by mapping elements to their membership vectors (5), some information about the partition itself is lost. In particular, it is possible that two partitions, even though they are not identical, cannot be distinguished in terms of the distances between these vectors.

Nevertheless, we can guarantee the separation property by restricting to a reasonable subclass of fuzzy partitions. We call a fuzzy partition $\mathbf{P} = \{P_1, P_2 \dots P_k\}$ *normal*, if it satisfies the following:

N1 For each $x \in X$: $P_1(x) + \ldots + P_k(x) = 1$.

N2 For each $P_i \in \mathbf{P}$, there exists an $x \in X$ such that $P_i(x) = 1$.

In other words, we consider Ruspini partitions [9] and assume that each cluster has a prototypical element. Moreover, we assume the following equivalence relation on X:

$$E_{\mathbf{P}}(x,x') = 1 - \frac{1}{2}\sum_{i=1}^{k} |P_i(x) - P_i(x')|. \tag{10}$$

Note that $0 \leq E_{\mathbf{P}}(x,x') \leq 1$ for all $(x,x') \in X^2$ under assumption N1.

Now, consider two normal fuzzy partitions $\mathbf{P}$ and $\mathbf{Q}$, and suppose that $d(\mathbf{P},\mathbf{Q}) = 0$. According to our definition of $d(\cdot)$, this obviously means that

$$E_{\mathbf{P}}(x,x') = E_{\mathbf{Q}}(x,x') \tag{11}$$

for all $(x,x') \in C$. We call a set $\{p_1, p_2 \dots p_k\} \subset X$ a prototype set for $\mathbf{P}$, if $P_i(p_i) = 1$ for all $i = 1,\dots,k$ (note that a prototype set is not necessarily unique). We distinguish two cases.

(a) There are no identical prototype sets for $\mathbf{P}$ and $\mathbf{Q}$ (note that this is necessarily the case if $\mathbf{P}$ and $\mathbf{Q}$ have a different number of clusters). Then, we can find elements $x,x' \in X$ such that x and x' are prototypes for $\mathbf{P}$ but not for $\mathbf{Q}$. Note that N1 and N2 jointly imply that a prototype is represented by a 0/1 membership vector, and that $|\mathbf{P}(x) - \mathbf{P}(x')| = 1$ for two different prototypes x and x'. Moreover, these properties imply that the extreme distance of 1 can *only* be assumed for prototypes, whereas $|\mathbf{P}(x) - \mathbf{P}(x')| < 1$ if either x or x' is not a prototype. Thus, it follows that $E_{\mathbf{P}}(x,x') = 0$ and $E_{\mathbf{Q}}(x,x') > 0$, which means that condition (11) is violated. Hence, we have constructed a contradiction with the assumption that $d(\mathbf{P},\mathbf{Q}) = 0$.

(b) There are identical prototype sets $\{p_1,\dots,p_k\} = \{q_1,\dots,q_\ell\}$, respectively, for $\mathbf{P}$ and $\mathbf{Q}$ (which means that $k = \ell$, i.e., $\mathbf{P}$ and $\mathbf{Q}$ do have the same number of clusters). We can then establish a one-to-one correspondence between prototypes such that, without loss of generality, $p_i = q_i$ for $i = 1,\dots,k$. From properties N1 and N2, it follows that the membership degree of any element x in the cluster P_i is a function of $E_{\mathbf{P}}(x,p_i)$. In fact, noting that $\mathbf{P}(p_i)$ is a 0/1 vector with a single 1 on position i, we get

$$\begin{aligned} E_{\mathbf{P}}(x,p_i) &= 1 - \frac{1}{2}\sum_{j=1}^{k} |P_j(x) - P_j(p_i)| \\ &= 1 - \frac{1}{2}\left((1-P_i(x)) - \sum_{j\neq i} P_j(x)\right) \\ &= 1 - \frac{1}{2}\left((1-P_i(x)) - (1-P_i(x))\right) \\ &= P_i(x). \end{aligned}$$

From (11), it thus follows that $P_i(x) = Q_i(x)$ for all $x \in X$, i.e., the i-th cluster in $\mathbf{P}$ and the i-th cluster in $\mathbf{Q}$ are identical. Since this holds for all $i \in \{1,2\dots k\}$, we have shown that $\mathbf{P} = \mathbf{Q}$.

The above results can be summarized as follows.

Theorem: *The distance function (8) on fuzzy partitions is a pseudometric, i.e., it is reflexive, symmetric, and subadditive. Moreover, on the restricted class of normal fuzzy partitions or, more specifically, under the assumptions N1, N2, and (10), it also satisfies the separation property and, therefore, is a metric.*

4 Summary and Outlook

We have introduced a generalization of the Rand index for comparing two fuzzy clustering structures. Elaborating on the formal properties of our measure, we have shown that it is a

pseudo-metric and, on a subclass of fuzzy partitions obeying certain normality assumptions, even a metric.

In future work, we plan to extent our approach to other similarity measures for (non-fuzzy) clustering structures which are related to the Rand index in the sense of being defined in terms of the same basic quantities, namely the numbers a, b, c, and d of concordant and discordant object pairs. An example of such a measure is the Jaccard coefficient, which is defined as $a/(a+b+c)$.

References

[1] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering validity checking methods: Part I. *ACM SIGMOD Record*, 31(2):40–45, 2002.

[2] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. Clustering validity checking methods: Part II. *ACM SIGMOD Record*, 31(3):19–27, 2002.

[3] J. Beringer and E. Hüllermeier. Fuzzy clustering of parallel data streams. In J. Valente de Oliveira and W. Pedrycz, editors, *Advances in Fuzzy Clustering and Its Application*, pages 333–352. John Wiley and Sons, 2007.

[4] MP. Windham. Cluster validity for the fuzzy c-means algorithm. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 4(4):357–363, 1982.

[5] XL. Xie and GA. Beni. Validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(8):841–846, 1991.

[6] R. J. G. B. Campello. A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. *Pattern Recognition Letters*, 28(7):833–841, 2007.

[7] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[8] A. DeLuca and S. Termini. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. *Information and Control*, 20:301–312, 1972.

[9] E.H. Ruspini. A new approach to clustering. *Information Control*, 15:22–32, 1969.