A Fuzzy Variant of the Rand Index for Comparing Clustering Structures

Eyke Hüllermeier
FB Mathematik/Informatik
Philipps-Universität Marburg
eyke@mathematik.uni-marburg.de

Maria Rifqi
University Pierre et Marie Curie–Paris 6
Paris, F-75016, France
maria.rifqi@lip6.fr

Draft of a paper with the same title presented at IFSA/EUSFLAT–2009, Lisbon, 2009

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.

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 [3, 4]. 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.

Just to give a motivating example, consider the problem to compare two different representations of the same set of objects. More concretely, the authors in [1] 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., [6] for an early proposal).

In a recent paper by Campello [2], the author has proposed an extension of the *Rand index* [5], 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).

[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 A New Fuzzy Rand Index

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.

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

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

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 (1). Generally, this relation is of the form

$$E_{\mathbf{P}}(x, x') \;=\; 1 - \|\mathbf{P}(x) - \mathbf{P}(x')\|, \tag{2}$$

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 (2) 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

$$1 - |E_{\mathbf{P}}(x, x') - E_{\mathbf{Q}}(x, x')| \;\in\; [0,1]. \tag{3}$$

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:

$$d(\mathbf{P}, \mathbf{Q}) \;=\; \frac{\sum_{(x,x') \in C} |E_{\mathbf{P}}(x, x') - E_{\mathbf{Q}}(x, x')|}{n(n-1)/2} \tag{4}$$

Likewise,

$$1 - d(\mathbf{P}, \mathbf{Q}) \tag{5}$$

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

Proposition: *In the case where* $\mathbf{P}$ *and* $\mathbf{Q}$ *are non-fuzzy partitions, the measure (5) reduces to the original Rand index.*

Definition: 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 $P_i(x) = 1$.

Theorem: *The distance function (4) on fuzzy partitions is a pseudometric, i.e., it is reflexive, symmetric, and subadditive. Moreover, if the fuzzy partitions are normal (i.e., satisfy the assumptions N1 and N2) and, moreover, the equivalence relation on* X *is of the form*

$$E_{\mathbf{P}}(x, x') \;=\; 1 - \frac{1}{2} \sum_{i=1}^{k} |P_i(x) - P_i(x')|, \tag{6}$$

then the distance function also satisfies the separation property and, therefore, is a metric.

References

[1] 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.

[2] 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.

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

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

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

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