A general framework for evaluating and comparing soft clusterings

Andrea Campagner, Davide Ciucci, Thierry Denœux

A General Framework for Evaluating and Comparing Soft Clusterings

Andrea Campagner[a], Davide Ciucci[a], Thierry Denœux[b,c]

[a] *Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca, Milano, Italy*
[b] *Université de technologie de Compiègne, CNRS, UMR 7253 Heudiasyc, Compiègne, France*
[c] *Institut universitaire de France, Paris, France*

Abstract

In this article, we propose a general framework for the development of external evaluation measures for soft clustering. Our proposal is based on the interpretation of soft clustering as representing uncertain information about an underlying, unknown hard clustering. We present a general construction, based on optimal transport theory, by which any evaluation measure can be naturally extended to soft clustering. The proposed "transport-based measure" provides an objective, interval-valued comparison index that represents the range of compatibility between two soft clusterings. We study the metric and complexity properties of the proposed approach, as well as its relationship with other existing proposals. We also propose approximation and bounding algorithms that make the approach practical for large datasets. Finally, we illustrate the application of the proposed method through two computational experiments.

Key words: Clustering analysis, Soft clustering, Evaluation, Validation, Comparison

1. Introduction

Clustering analysis is an important task within machine learning and data analysis. Intuitively, the aim of clustering analysis is to detect, in an unsupervised fashion, a grouping of objects within categories, called *clusters*. Clustering analysis has important applications in different settings, including anomaly detection [1], community detection [26], and biological data analysis [35].

One of the most important steps in clustering analysis regards the evaluation of the obtained results [49]. Two main evaluation approaches can be distinguished. *Internal validation* indices [34, 25] evaluate the quality of a clustering result with reference to intrinsic properties of the result itself (e.g., separation of clusters). By contrast, *external evaluation* methods [32] objectively assess

Email address: a.campagner@campus.unimib.it (Andrea Campagner)

the quality of a clustering result, by means of a comparison between two or more clusterings, one of which is usually assumed to be the *correct one*. In this article, we refer exclusively to external evaluation.

In the case of hard clustering, that is clustering methods in which each object is *unambiguously* assigned to a single cluster (e.g., k-means), several evaluation criteria have been considered [2]. These include the Rand index [40], purity [42], information-theoretic measures [47], or the partition distance [11]. By contrast, how to properly evaluate the results of a clustering analysis is much less clear in the case of soft clustering methods, i.e., techniques that provide an explicit representation of uncertainty [39, 22]. Several soft clustering methods have been proposed in the literature, each of which represents clustering uncertainty in different ways. In *rough* [33] clustering, clusters are represented as pairs of sets containing the objects that *certainly* or only *possibly* belong to the cluster. In *probabilistic* [7] and *fuzzy* [4, 6, 41] clustering, the object-cluster assignment is represented through a probability distribution or fuzzy partition. In *possibilistic* [31, 41] clustering, uncertainty is represented through possibility distributions. Finally, in *evidential* [17, 20, 14, 15] clustering, the uncertain object-cluster assignment is represented by belief functions. Remarkably, evidential clustering has been shown to generalize all the above mentioned methodologies [18].

In the literature, the development of soft clustering evaluation measures has mainly focused on the extension of common measures, notably the Rand index, to the setting of fuzzy clustering [2, 10, 24, 27], while only recently a formulation of this approach has been introduced for the more general case of evidential clustering [18]. Most proposals in the literature, however, have been shown to be severely lacking in terms of satisfied properties [18, 27]. For example, most measures fail to be semi-metrics, hence cannot be used to check the equivalence between two clusterings. Additionally, the question of which appropriate metric properties an evaluation measure should satisfy has scarcely been studied and, consequently, the existing metrics can be hard to interpret or apply.

As a second, and most remarkable limitation, existing measures fail to properly quantify and distinguish different types of uncertainty that can arise in soft clustering [18]. In particular, two relevant types of uncertainty can be distinguished: *ambiguity*, i.e., the inability to uniquely assign an object to a single cluster (typical of rough clustering); and *partial assignment*, i.e., the assignment of objects to multiple clusters, each with a given degree of membership (typical of fuzzy and probabilistic clustering). Existing evaluation measures, however, conflate these two different types of uncertainty. This issue is especially problematic for more general types of soft clustering, such as evidential clustering [18], in which the two forms of uncertainty coexist.

We claim that the above mentioned limitations stem from a lack of understanding concerning the following two natural questions: What should be quantified by an evaluation measure for soft clustering? Which properties should such a measure have? In this article, following previous accounts in the literature [9, 18, 27], we argue that such measures have two possible aims. Since we interpret soft clustering as describing the uncertainty with respect to an underlying but unknown hard clustering, a first aim is to provide a picture of

the uncertainty within the two soft clusterings to be compared. Second, such a measure should provide a way to objectively compare two clusterings or to assess the quality of a clustering with respect to a *ground truth*. By drawing from the previous literature [9, 18], in this article we argue that a natural requirement is to adopt interval-valued measures that satisfy certain reasonable metric properties. Intuitively, the lower bound of the interval should quantify the *compatibility* between two soft clusterings, i.e., whether there exists a hard clustering compatible with both soft clusterings [18]. By contrast, the upper bound should quantify their *similarity*, i.e., to which degree the two clusterings are strictly equivalent, penalizing any possible ambiguity. To provide a formal translation of these principles, we require that the lower bound should be a *consistency*, while the upper bound should be a *metric* (see Section 2.1).

In this article, we propose a general approach to address the above mentioned questions and limitations, by introducing a mathematical construction that can be applied to extend any clustering evaluation measure to the case of soft clustering. The proposed *transport-based distance* (see Section 3.2) relies on the interpretation of soft clustering as representing a distribution over hard clusterings (see Section 3.1). We then use construction methods from optimal transport theory [46] to provide interval-valued comparison indices that can be used to objectively compare two soft clusterings in terms of their mutual consistency and equality, while providing an account of the uncertainty involved in the comparison. We provide an in-depth study of the proposed method, with respect to both its computational complexity and metric properties. Furthermore, we describe approximation techniques that can be used to reduce the computational complexity of the method and we show that some known measures previously proposed in the literature emerge as special cases of our framework (see Section 4). Finally, we illustrate the application of our approach through two simple computational experiments (see Section 5).

2. Background and Related Work

In the following section, we first provide basic background on metric spaces and related structures in Section 2.1. Then, we summarize some important notions related to clustering in Section 2.2. Finally, a brief review of clustering evaluation measures for soft clustering is provided in Section 2.3. Background material on belief function theory is presented in Appendix A.

2.1. Metric Spaces

Let us first recall some basic notions on metric spaces [45]. Let X be a countable set. A *metric* over X is a function $d : X \times X \mapsto \mathbb{R}_+$ s.t.:

(M1) $\forall x \in X,\ d(x, x) = 0$;

(M2) $\forall x, y \in X,\ x \neq y \implies d(x, y) > 0$;

(M3) $\forall x, y \in X,\ d(x, y) = d(y, x)$;

(M4) $\forall x, y, z \in X,\ d(x, z) \leq d(x, y) + d(y, z)$.

Metric d is *normalized* if $\max_{x,y \in X} d(x, y) = 1$. If d is a normalized metric, then its dual $s = 1 - d$ is called a *similarity* over X. Several weakenings of the notion of metric have been considered in the literature. Here we recall the following:

- d is a *pseudo-metric* iff it satisfies (M1), (M3) and (M4);
- d is a *semi-metric* iff if satisfies (M1), (M2) and (M3);
- d is a *meta-metric* iff it satisfies (M2), (M3) and (M4);
- d is a *ρ-relaxed metric* (with $\rho \in \mathbb{R}_+$) iff it satisfies (M1), (M2), (M3) and

$$\forall x, y, z \in X,\ \ d(x, z) \leq \rho \left(d(x, y) + d(y, z) \right).$$

Obviously, combinations of the above concepts can be considered (e.g., a semi-pseudo-metric is a function satisfying only (M1) and (M3)). If d is a normalized pseudo- (resp., semi-, meta- ρ-relaxed) metric, then $s = 1 - d$ is called a pseudo- (resp., semi-, meta- ρ-relaxed) similarity. For simplicity, in the following, we will refer to any semi-pseudo similarity as a *consistency*.

A metric d over X can be extended to a metric over 2^X. The resulting metric d_H is called the *Hausdorff metric* based on d, defined as

$$d_H(A, B) = \max\{\max_{a \in A} d(a, B), \max_{b \in B} d(A, b)\}, \tag{1}$$

where $d(a, B) = \min_{b \in B} d(a, b)$ and $d(A, b) = \min_{a \in A} d(a, b)$. If d is a (pseudo-, meta-, semi-) metric, then d_H satisfies the same properties.

Similarly, a metric d over X can be extended to a metric over the space $\mathcal{P}(X)$ of probability measures over X. The resulting metric, denoted as d_W, is called the *Wasserstein metric* (also known as Kantorovich-Rubinstein metric) [29, 46] based on d. It is formally defined, for any two probability measures Pr_1 and Pr_2 on X, as

$$\begin{aligned} d_W(Pr_1, Pr_2) = \min_{\sigma} & \sum_{(x_1, x_2) \in X^2} \sigma(x_1, x_2) d(x_1, x_2) \\ \text{s.t.} & \sum_{x_2 \in X} \sigma(x_1, x_2) = Pr_1(x_1) \\ & \sum_{x_1 \in X} \sigma(x_1, x_2) = Pr_2(x_2) \\ & \sum_{(x_1, x_2) \in X^2} \sigma(x_1, x_2) = 1 \\ & \forall (x_1, x_2) \in X, \sigma(x_1, x_2) \geq 0. \end{aligned} \tag{2}$$

If d is a (pseudo-, meta-, semi-) metric then d_W satisfies the same properties. The Wasserstein metric is the minimal expected distance between two distributions Pr_1 and Pr_2 for all joint distributions whose marginals are Pr_1 and Pr_2. It can also be seen as the minimal cost of transforming one distribution into the other by moving the probability masses.

2.2. Clustering

Let $X = \{x_1, ..., x_n\}$ be a set of objects. A *hard clustering* is a unique assignment of objects in X to *clusters*. Formally, a hard clustering is a mapping $C : X \mapsto \Omega$, where $\Omega = \{\omega_1, ..., \omega_k\}$ is a set of clusters. This representation is called *object-based*. By an abuse of notation we identify cluster ω_i with the set $\{x \in X : C(x) = \omega_i\}$. We denote by χ_{ω_i} the indicator function corresponding to ω_i. Any two clusterings C_1, C_2 that are equivalent up to a relabeling of the clusters are considered to be identical. We denote the corresponding equivalence relation by $C_1 \sim C_2$.

Example 2.1. *Let $X = \{x_1, \ldots, x_5\}$ and $\Omega = \{\omega_1, \omega_2, \omega_3\}$. Then, the mapping $C : X \mapsto \Omega$ s.t. $C(x_1) = C(x_5) = \omega_1$, $C(x_2) = C(x_3) = \omega_2$ and $C(x_4) = \omega_3$ is a hard clustering. For simplicity, we represent C as the tuple $c = (1, 2, 2, 3, 1)$, where $c_i = j$ iff $C(x_i) = \omega_j$.*

Given a hard clustering C, its *relational representation* is the equivalence relation $[C] \subseteq X \times X$ such that $\forall (x, y) \in X^2$, $(x, y) \in [C] \Leftrightarrow C(x) = C(y)$. Obviously, $C_1 \sim C_2$ iff $[C_1] = [C_2]$.

Example 2.2. *The clustering defined in Example 2.1 can be equivalently described by its relation representation*

$$[C] = \{(x_1, x_5), (x_5, x_1), (x_2, x_3), (x_3, x_2)\}.$$

As mentioned in Section 1, in *soft clustering* the unique assignment assumption is relaxed: the intuition is that we allow uncertainty in the assignment of objects to clusters. With reference to evidential clustering, which represents the most general framework among the ones we consider, the uncertainty about cluster assignment is represented as a Dempster-Shafer mass function (see Appendix A). Formally, using the object-based representation, an *evidential clustering* is a set $M = \{m_x\}_{x \in X}$, where each m_x is a *mass function*, i.e., a mapping $m_x : 2^\Omega \mapsto [0, 1]$ such that $\sum_{A \subseteq \Omega} m_x(A) = 1$.

Example 2.3. *Let X and Ω be as in Example 2.1. Then $M = \{m_{x_i}\}_{x_i \in X}$ defined as: $m_{x_1}(\{\omega_1\}) = 1$; $m_{x_2}(\{\omega_2\}) = 1$; $m_{x_3}(\{\omega_2, \omega_3\}) = m_{x_3}(\Omega) = 0.5$; $m_{x_4}(\{\omega_3\}) = 1$; $m_{x_5}(\Omega) = 0.5, m_{x_5}(\{\omega_1\}) = m_{x_5}(\{\omega_2\}) = m_{x_5}(\{\omega_3\}) = 1/6$ is an evidential clustering.*

If the mass functions m_x are *logical*, i.e., if they are such that $m_x(A) = 1$ for some subset $A \subseteq \Omega$, then, the collection $R = \{m_x\}_{x \in X}$ is said to be a *rough clustering*. For each $x \in X$, we denote the unique $A \subseteq \Omega$ s.t. $m_x(A) = 1$ as $R(x)$. A rough clustering can be equivalently represented by associating with each cluster ω a pair $(l(\omega), u(\omega))$ of subsets of X verifying $l(\omega) \subseteq u(\omega)$. Intuitively, $l(\omega)$ contains the elements that "certainly" belong to the cluster, while $u(\omega)$ contains the elements that "possibly" belong to the cluster. The set $u(\omega) \setminus l(\omega)$, called the *boundary* of the cluster, contains the elements whose assignment to cluster ω is uncertain. Finally, a rough clustering can be seen as a set of hard clusterings. Namely, a hard clustering C is compatible with R if

for all $x \in X$, $C(x) \in R(x)$. Then, we identify R with the set $C(R) = \{C : C \text{ is compatible with } R\}$.

Example 2.4. *Let X and Ω be defined as in Example 2.1. Then R defined by $R(x_1) = \omega_1$, $R(x_2) = \omega_2$, $R(x_3) = \{\omega_2, \omega_3\}$, $R(x_4) = \omega_3$, $R(x_5) = \Omega$ is a rough clustering. Using the tuple-based notation introduced in Example 2.1, R can be equivalently represented by the set*

$$C(R) = \{(1,2,3,3,3), (1,2,3,3,2), (1,2,3,3,1), \\ (1,2,2,3,3), (1,2,2,3,2), (1,2,2,3,1)\}.$$

For simplicity, we denote $C(R)$ as $(1, 2, \{2, 3\}, 3, \{1, 2, 3\})$.

If all m_x are Bayesian, then the collection $F = \{m_x\}_{x \in X}$ is a *fuzzy* or *probabilistic* clustering. Finally, if all m_x are consonant, then the collection $P = \{m_x\}_{x \in X}$ is a *possibilistic clustering*. Both fuzzy and possibilistic clustering can be represented as a collection of cluster membership vectors $F = \{\mu_x\}_{x \in X}$, where $\mu_x(\omega) = \sum_{A \ni \omega} m_x(A)$. In possibilistic clustering it is assumed that, for all $x \in X$, $\max_{\omega \in \Omega} \mu_x(\omega) \leq 1$, while in fuzzy clustering it is usually assumed that for all $x \in X$, $\sum_{\omega \in \Omega} \mu_x(\omega) = 1$.

Example 2.5. *Let X and Ω be defined as in Example 2.1. Then, F defined as $\mu_{x_1}(\omega_1) = 1$, $\mu_{x_2}(\omega_2) = 1$, $\mu_{x_3}(\omega_2) = \mu_{x_3}(\omega_3) = 0.5$, $\mu_{x_4}(\omega_3) = 1$ and $\mu_{x_5}(\omega_1) = \mu_{x_5}(\omega_2) = \mu_{x_5}(\omega_3) = \frac{1}{3}$ is a fuzzy clustering.*

P defined as $\mu_{x_1}(\omega_1) = 1$, $\mu_{x_2}(\omega_2) = 1$, $\mu_{x_3}(\omega_2) = \mu_{x_3}(\omega_3) = 1$, $\mu_{x_4}(\omega_3) = 1$ and $\mu_{x_5}(\omega_1) = \mu_{x_5}(\omega_2) = 1, \mu_{x_5}(\omega_3) = 0.8$ is a possibilistic clustering.

Similarly to the case of hard clustering, a relational representation can be defined also for the case of evidential clustering [18]. In this case, let $\Theta = \{s, \neg s\}$ be the frame where s denotes that two objects are in the same cluster, and $\neg s$ denotes the opposite event. Given an evidential clustering M, the corresponding relational representation can be obtained, for any two distinct objects $(x, y) \in X^2$, by combining m_x and m_y by Dempster's rule [43], and computing the restriction of the resulting mass function to Θ [18]. The resulting mass function $m^{x,y}$ is then defined by

$$m^{x,y}(\{s\}) = \sum_{\omega \in \Omega} m_x(\omega) m_y(\omega) \tag{3a}$$

$$m^{x,y}(\{\neg s\}) = \sum_{A \cap B = \emptyset} m_x(A) m_y(B) - m(\emptyset) \tag{3b}$$

$$m^{x,y}(\Theta) = \sum_{A \cap B \neq \emptyset} m_x(A) m_y(B) - m(s) \tag{3c}$$

$$m^{x,y}(\emptyset) = m_x(\emptyset) + m_y(\emptyset) - m_x(\emptyset) m_y(\emptyset). \tag{3d}$$

Obviously, it holds that $m^{x,x}(\{s\}) = 1$.

Example 2.6. *Let M be the evidential clustering defined in Example 2.3. Then, the relational representation $[M]$ of M is defined as the reflexive (i.e., $m^{x,x}(s) = 1$) and symmetric (i.e., $m^{x,y} = m^{y,x}$) closure of the relation shown in Table 1.*

Table 1: Relational representation of the evidential clustering defined in Example 2.3.

Pair	$m(\emptyset)$	$m(s)$	$m(\neg s)$	$m(\Theta)$
(x_1, x_2)	0	0	1	0
(x_1, x_3)	0	0	0.5	0.5
(x_1, x_4)	0	0	1	0
(x_1, x_5)	0	$\frac{1}{6}$	$\frac{1}{3}$	0.5
(x_2, x_3)	0	0	0	1
(x_2, x_4)	0	0	1	0
(x_2, x_5)	0	$\frac{1}{6}$	$\frac{1}{3}$	0.5
(x_3, x_4)	0	0	0	1
(x_3, x_5)	0	0	$\frac{1}{12}$	$\frac{11}{12}$
(x_4, x_5)	0	$\frac{1}{6}$	$\frac{1}{3}$	0.5

For the case of fuzzy and possibilistic clustering, two alternative approaches to obtain a relational representation have been considered. In the first approach, proposed by Campello et al. [10], $\mu^{x,y}$ is defined, based on a t-norm $\wedge$ and the dual t-conorm $\vee$ as

$$\mu^{x,y}(s) = \bigvee_{\omega \in \Omega} \mu_x(\omega) \wedge \mu_y(\omega) \tag{4a}$$

$$\mu^{x,y}(\neg s) = \bigvee_{\omega_1 \neq \omega_2 \in \Omega} \mu_x(\omega_1) \wedge \mu_y(\omega_2). \tag{4b}$$

When $\wedge = \otimes_P$ (i.e., the product t-norm), $\vee = \oplus_P$ (i.e., the bounded sum t-conorm) and $\{\mu_x\}_{x \in X}$ defines a fuzzy partition, then we obtain the same definition given previously for evidential clustering.

In the second approach, proposed by Hüllermeier et al. [27], the relational representation is defined based on a normalized metric d_F on $[0,1]^\Omega$ (i.e., the set of functions $\Omega \mapsto [0,1]$). Then, we simply define

$$\mu^{x,y}(\neg s; d_F) = d_F(\langle \mu_x(\omega_1), \ldots, \mu_x(\omega_k) \rangle, \langle \mu_y(\omega_1), \ldots, \mu_y(\omega_k) \rangle) \tag{5a}$$

$$\mu^{x,y}(s; d_F) = 1 - \mu^{x,y}(\neg s; d_F). \tag{5b}$$

Finally, by reference to the case of evidential clustering, we can remark that, if we interpret a soft clustering as describing our uncertainty in regard to some underlying (unknown) hard clustering, then two types of uncertainty can be distinguished. First, *partial assignment*, i.e., the existence of conflicting evidence supporting the assignment of an object x to two different clusters ω_1 and ω_2, in which case we have $m_x(\{\omega_1\}) > 0$ and $m_x(\{\omega_2\}) > 0$. Second, *ambiguity*, i.e., the assignment of some mass to non-singleton events (i.e., $m_x(A) > 0$ for some $A \subseteq \Omega$ such that $|A| > 1$), which describes our inability to exactly determine to which cluster an object belongs. It is easy to observe that fuzzy clustering only considers partial assignment, since all the mass is assigned to singletons. By contrast, in the case of rough clustering, only ambiguity is present.

2.3. Clustering Comparison Measures

Several measures have been defined to compare clusterings. We can distinguish between approaches that rely on the relational representation on the

one hand, and on the object-based representation on the other hand. Here, we provide a survey of both families of measures for hard and soft clustering.

2.3.1. Relational-based Comparison Measures

Given two hard clusterings C_1, C_2, a sensible approach to compare them is to evaluate the number of pairs of objects $(x, y) \in X^2$ on which they agree. In particular, the Rand index is defined as the proportion of pairs of distinct objects that are either in the same cluster in C_1 and in the same cluster in C_2, or in different clusters in C_1 and in different clusters in C_2. Formally,

$$\mathsf{Rand}(C_1, C_2) = \frac{2|\{(x_i, x_j) : 1 \leq i < j \leq n \text{ and } [(x_i, x_j) \in (A \cup B^c)]\}|}{n(n-1)}, \quad (6)$$

where $A = [C_1] \cap [C_2]$, $B = [C_1] \cup [C_2]$ and $(\cdot)^c$ is the complement operator. Other evaluation measures can be defined based on the same principle, including the Jaccard and Fowlkes-Mallows [23] indices. It is easy to show that these indices are similarities on hard clusterings, irrespective of their representation.

Several extensions of the above mentioned indices to the soft clustering setting have been considered. In the case of fuzzy and possibilistic clustering, Campello et al. [10] proposed an approach based on extending the computation of Eq (6) to the fuzzy case, based on a t-norm $\wedge$ and a t-conorm $\vee$. Similarly, Frigui et al [24] proposed an approach that is a special case of the formulation given by Campello et al., where $\wedge = \otimes_P$, $\vee = \oplus_P$ and only fuzzy clusterings are considered. This latter approach has also been generalized to the rough clustering setting [21], by means of a transformation from rough to fuzzy clustering. A potential flaw of the formulations proposed in [10, 21, 24] is that the corresponding measures fail to satisfy the properties mentioned in Section 1, hence they cannot be used to provide an objective comparison between two fuzzy or rough clusterings. In particular, since they fail to be pseudo- and semi-similarities, it can happen that $\mathsf{Rand}(F, F) < 1$ for some fuzzy clustering F. Consequently, the generalizations of the Rand index proposed in [10, 21, 24] cannot be directly applied to objectively compare the results of two fuzzy clustering algorithms since, even when the two algorithms report the same clustering, these measures could report a value smaller than 1, denoting a difference between them.

An alternative approach to generalize the Rand and Jaccard indices, for the case of fuzzy and possibilistic clustering, was proposed by Anderson et al. [2]. By adopting a matrix-based representation, the authors show that an equivalent (for the case of hard clustering) formulation of Eq (6) can be generalized to fuzzy and possibilistic clusterings, and can be computed in time $O(n)$. Despite its favorable complexity, it has been shown that the obtained generalization of the Rand index has some non-intuitive properties. In particular, as the indices proposed in [10, 21, 24], it can fail to be a meta-similarity. Furthermore, its value is not restricted to the range $[0, 1]$, and may even be negative: the absence of an absolute comparison scale thus makes the measure proposed in [2] hardly applicable for evaluating the quality of a fuzzy clustering with respect to a known ground truth, or for comparing the results of two different fuzzy clustering algorithms.

To address these limitations, for the case of the Rand index, a different approach was considered by Hüllermeier et al. [27]. This approach is based on an equivalent definition of the Rand index as

$$\mathsf{Rand}(C_1, C_2) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} 1 - |\mathbb{1}_{(x_i,x_j) \in [C_1]} - \mathbb{1}_{(x_i,x_j) \in [C_2]}|. \quad (7)$$

Let d^F be a normalized metric on $[0,1]^\Omega$. Then, formula (7) for the Rand index can be generalized to fuzzy clustering as

$$\mathsf{Rand}'_F(P_1, P_2) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} 1 - |\mu_1^{x_i,x_j}(s; d^F) - \mu_2^{x_i,x_j}(s; d^F)|, \quad (8)$$

where $\mu_\ell^{x_i,x_j}(s; d^F)$ for $\ell = 1, 2$ are defined by (5). In contrast with the previous measures, Rand'_F is a similarity for the relational representation, and a pseudo-similarity for the object-based representation. If d^F is the cosine distance, we obtain the approach proposed in [8]. The approach proposed in [27] has been generalized to the case of evidential clustering by Denoeux et al. [18]. Based on the relational representation of evidential clustering (see Section 2.2), the Rand index is expressed as:

$$\mathsf{Rand}_E(M_1, M_2) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} 1 - d_M(m_1^{x_i,x_j}, m_2^{x_i,x_j}) \quad (9)$$

where d_M is a normalized metric for mass functions. The authors consider, in particular, the Belief distance d_B and Jousselme's distance d_J, showing that the corresponding measures are pseudo-similarities for the object-based representation and similarities for the relational representation. Nonetheless, the authors of [18] note that their approach is not completely satisfactory for the comparison of evidential clusterings, as d_B does not distinguish between *partial assignment* and *ambiguity*, while d_J penalizes ambiguity more than partial assignment.

Example 2.7. *Let M be as in Example 2.3. Consider the pair of objects (x_2, x_3), for which $m^{x_2,x_3}(\Omega) = 1$. Furthermore, let M_1 be an evidential clustering such that $m^{x_2,x_3}(\{s\}) = 1$, and M_2 an evidential clustering such that $m^{x_2,x_3}(\{s\}) = m^{x_2,x_3}(\{\neg s\}) = 0.5$. Then, it holds that*

$$d_B(m_M^{x_2,x_3}, m_{M_1}^{x_2,x_3}) = d_B(m_{M_2}^{x_2,x_3}, m_{M_1}^{x_2,x_3}) = 0.5.$$

Thus, Rand_B does not distinguish between ambiguity and partial assignment. By contrast, it holds that

$$d_J(m_M^{x_2,x_3}, m_{M_1}^{x_2,x_3}) = 0.71 > 0.5 = d_B(m_{M_2}^{x_2,x_3}, m_{M_1}^{x_2,x_3}).$$

Thus, Rand_J penalizes ambiguity more than partial assignment.

To address this issue, the authors of [18] also propose an alternative generalization of the Rand index, based on the *degree of conflict*

$$\mathcal{K}(m_1^{x_i,x_j}, m_2^{x_i,x_j}) = \sum_{A,B:A \cap B = \emptyset} m_1^{x_i,x_j}(A) \cdot m_2^{x_i,x_j}(B).$$

The authors observe that the obtained measure, denoted as Rand_K, is a consistency and suggest that a pair of measures (e.g., Rand_B and Rand_K) could be used to obtain a comprehensive evaluation measure.

2.3.2. Object-based Comparison Measures

While the measures reviewed in Section 2.3.1 rely on the relational representation, different measures based on the object-based representation have also been proposed. This latter family of measures addresses a limitation of the previously mentioned relational representation-based measures, namely: let M_1, M_2 be two evidential clusterings such that $M_1 \neq M_2$ but $[M_1] = [M_2]$ (that is, the two clusterings are different w.r.t. the object-based representation but have identical relational representation); then, if d is a relational-based comparison measure, $d(M_1, M_2) = 0$ (see also Corollary 4.1).

The *partition distance* [11] for two hard clusterings C_1, C_2 is defined as the minimum number of objects to be moved to transform C_1 into C_2 (or, equivalently, C_2 into C_1). Assume, without loss of generality, that $|\Omega_1| = |\Omega_2|$ (indeed, if $|\Omega_1| \neq |\Omega_2|$, we can add empty clusters to the clustering with the smallest number of clusters). Then, the partition distance can be computed as

$$d_\pi(C_1, C_2) = \frac{1}{2(n-1)} \min_\tau \sum_{i=1}^{k} |\omega_i^1 \Delta \omega_{\tau(i)}^2|, \tag{10}$$

where τ is a permutation of $\{1, \ldots, k\}$, ω_i^j is the i-th cluster in clustering C_j, and Δ is the symmetric difference operator. We can easily see that d_π is a normalized metric: indeed, at most $n-1$ objects need to be moved between clusters to make the two clusterings equivalent and each moved object is counted twice; hence, the $2(n-1)$ factor in the denominator normalizes the range of d_π.

An extension of the partition distance to the case of fuzzy clustering was proposed by Zhou [50], based on the fact that the partition distance is a special case of the Wasserstein construction (see Eq. (2) in Section 2.1). The obtained measure is a proper generalization of the partition distance and is a metric. A similar approach was adopted by Anderson et al [3], who proposed a generalization of the partition distance to fuzzy and possibilistic clusterings. While for the case of fuzzy clusterings the obtained measure is a metric, in the case of possibilistic clusterings the obtained measure is not even a meta-metric and may result in values greater than 1. Thus, it is subject to the same limitations as the generalization of the Rand index proposed in [2]. To our knowledge, the extension of the partition distance to rough and evidential clustering has not been considered in the literature.

A second family of object-based comparison measures is based on information theory [47]. It includes the mutual information [47] and the variation of information [37]. These approaches have been recently extended to the case of rough clustering in [9], based on the representation of a rough clustering R as a collection $C(R)$ of compatible hard clusterings. Namely, the authors of [9] propose an interval-valued comparison measure representing the minimum

and maximum values of mutual information among all pairs of hard clusterings compatible with the rough clusterings to be compared. To our knowledge, no extension of these metrics to the more general case of evidential clustering has been proposed so far.

3. A General Framework for Soft Clustering Evaluation Measures

As shown in the previous section, most of the research on comparison measures for soft clustering has focused on the analysis of some specific indices, while a general methodology for obtaining such measures is still missing. Furthermore, as noted in [18, 27], most of the existing methods fail to satisfy the metric properties described in Section 1. Consequently, they can hardly be used for the objective comparison of soft clusterings. Notably, the more principled approaches introduced in [18] can also have some drawbacks, such as the inability to properly distinguish between different types of uncertainty arising in soft clustering.

In this section, we propose a general approach that attempts to address these limitations, based on the representation of a soft clustering as a mass function over hard clusterings. As already discussed in Section 1, this approach aims at addressing two different purposes of an evaluation measure for soft clustering, namely, uncertainty representation and objective comparison. In regard to the first aim, we will focus on interval-valued measures, as a compromise between succinctness and expressivity. Further, since a general soft clustering accounts for two different types of uncertainty (i.e., ambiguity and partial assignment), it seems reasonable to require any suitable measure to represent the full range of compatibility between two soft clusterings, varying with respect to how much we are willing to penalize ambiguity versus partial assignment. If we consider ambiguity to be completely acceptable, then the measure should quantify the compatibility between the two soft clusterings: under this constraint, the measure should arguably be a consistency, as previously discussed in [18, 27]. By contrast, when we consider ambiguity to be equivalent to an error, then the measure should quantify the exact equality between the two soft clusterings, and should hence be a metric (or, dually, a similarity).

In the following, we first introduce the distribution-based representation of soft clustering in Section 3.1. Then, we address the two above-mentioned aims for a soft clustering comparison measure by means of the *transport-based measures* (Section 3.2), which provide an objective interval-valued evaluation index based on the Wasserstein distance between two soft clusterings.

3.1. Distribution-based Representation of Soft Clustering

As mentioned at the beginning of this section, the approaches we propose are based on an alternative representation of soft clusterings as distributions over hard clusterings. Thus, intuitively, we assume that a soft clustering represents our uncertain knowledge about an underlying, unknown hard clustering.

As shown in Section 2.2, this is clearly true for the case of rough clustering: indeed, a rough clustering R can be represented as a set $C(R)$ of hard clusterings

C. Based on this observation, we extend this representation to general soft clusterings. Formally, given an evidential clustering M we define a mass function m_M as

$$m_M(R) = \prod_{x \in X} m_x(R(x)), \tag{11}$$

where R is any rough clustering. That is, an evidential clustering is represented as a mass function over hard clusterings or, equivalently, as a probability distribution over rough clusterings. Given an evidential clustering M and its distribution-based representation m_M, we denote by $\mathcal{F}(M)$ the collection of focal rough clusterings of m_M, that is $\mathcal{F}(M) = \{R : m_M(R) > 0\}$.

The distribution-based representation for rough, fuzzy and possibilistic clustering can then be obtained as a special case of Eq (11). Indeed, in the case of rough clustering, m_M is logical (i.e., $|\mathcal{F}(m_M)| = 1$), while in the case of fuzzy clustering the focal rough clusterings are all singletons (i.e., hard clusterings). Finally, in the case of possibilistic clustering, the possibility distribution $Poss_P$ can equivalently be represented as a consonant mass function, i.e., the focal rough clusterings are nested. More precisely, given a fuzzy clustering $F = \{\mu_x\}_x$, where $\mu_x : \Omega \mapsto [0, 1]$ is a probability distribution, we can represent it as the probability distribution on hard clusterings given by

$$Pr_F(C) = \prod_{x \in X} \mu_x(C(x)). \tag{12}$$

On the other hand, given a possibilistic clustering P and a t-norm $\wedge$, we can view P as a possibility distribution over hard clusterings:

$$Poss_P(C) = \bigwedge_{x \in X} \mu_x(C(x)). \tag{13}$$

If $\wedge$ is the product t-norm, we recover the case of fuzzy clustering.

Example 3.1. *Let F, P, M be the soft clusterings defined in Examples 2.5 and 2.3. Then, Pr_F is defined as*

$$Pr_F((1,2,3,3,3)) = Pr_F((1,2,3,3,2)) = Pr_F((1,2,3,3,1)) =$$
$$Pr_F((1,2,2,3,3)) = Pr_F((1,2,2,3,2)) = Pr_F((1,2,3,3,1)) = \frac{1}{6}.$$

Similarly, $Poss_P$ is defined as

$$Poss_P((1,2,3,3,3)) = Poss_P((1,2,2,3,3)) = 0.8$$
$$Poss_P((1,2,3,3,2)) = Poss_P((1,2,3,3,1)) = 1$$
$$Poss_P((1,2,2,3,2)) = Poss_P((1,2,2,3,1)) = 1.$$

Finally, m_M is defined as

$$m_M((1,2,\{2,3\},3,\Omega)) = m_M((1,2,\Omega,3,\Omega)) = 0.25$$
$$m_M((1,2,\{2,3\},3,1)) = m_M((1,2,\{2,3\},3,2)) = m_M((1,2,\{2,3\},3,3)) = \frac{1}{12}$$
$$m_M((1,2,\Omega,3,1)) = m_M((1,2,\Omega,3,2)) = m_M((1,2,\Omega,3,3)) = \frac{1}{12}.$$

Based on the distribution-based representation, in the following section, we describe a general approach that can be used to extend a comparison measure between hard clusterings to a comparison measure between soft clusterings.

3.2. Transport-based Measures

In this section we introduce an approach to extend any clustering comparison metric to the case of soft clustering. This approach relies on the Wasserstein construction from Optimal Transport theory [46] (see Section 2.1) to compute a distance between the distributional representations of the two soft clusterings to be compared. The metric properties we seek, then, directly follow from the constructions we employ.

To illustrate this idea, we recall that every evidential clustering (hence, every soft clustering) can be represented as a distribution over rough clusterings, as shown in Section 3.1. Hence, a comparison measure for soft clustering could be obtained by computing the *cost* of making the two distributions equivalent, by moving masses from one rough clustering to another, where the cost of such movements is determined by a base distance over rough clusterings. To define such measures, we will proceed in two steps: first, we will define a base distance over rough clusterings; then, we will extend this distance to general soft clustering using an approach based on the Wasserstein Distance.

Comparison between rough clusterings. Let d be a normalized distance over hard clusterings. Since, as mentioned in Section 1, we focus on interval-valued measures, we are interested in the definition of a pair of measures $\langle d_l, d_u \rangle$, where $1 - d_l$ is a consistency and d_u is a metric (equivalently, $1 - d^u$ is a similarity). Intuitively, d_u should measure the equivalence between two rough clusterings by completely discounting ambiguity (i.e., treating ambiguity as if it was equivalent to an error). By contrast, $1-d_l$ should measure the compatibility between two rough clusterings by determining whether they have a common disambiguation (i.e., a common assignment of objects to clusters). Let R_1, R_2 be two rough clusterings and $C(R_1), C(R_2)$ be the corresponding sets of compatible hard clusterings. We consider the following pair of measures:

$$d_0^R(R_1, R_2) = \min_{C_1 \in C(R_1), C_2 \in C(R_2)} d(C_1, C_2) \tag{14a}$$

$$d_1^R(R_1, R_2) = d_H(C(R_1), C(R_2)). \tag{14b}$$

That is, d_0^R is the minimum possible distance obtained by considering the hard clusterings that are compatible with R_1 and R_2, while d_1^R is the corresponding value of the Hausdorff distance[1]. It is easy to observe that $d_0^R(R_1, R_2) = 0$ as long as there exists a hard clustering C that is compatible with both R_1, R_2. In contrast, $d_1^R(R_1, R_2) = 0$ iff $C(R_1) = C(R_2)$. Furthermore, the following result directly follows from the definition of d_0^R, d_1^R:

[1] We remark that, if d_1^R was defined similarly to (14a) as $\max_{C_1 \in C(R_1), C_2 \in C(R_2)} d(C_1, C_2)$, i.e., by replacing the minimum in (14a) by the maximum, then this alternative version of d_1^R would not be a metric [9].

Proposition 3.1. *$1-d_0^R$ is a consistency, while d_1^R is a normalized metric (i.e., $1-d_1^R$ is a similarity).*

Proof. Clearly, d_0^R satisfies (M3). Similarly, d_0^R satisfies also (M1), while it fails to satisfy (M2). For the case of (M4), consider three rough clusterings R_1, R_2, R_3 s.t. $C(R_1) \cap C(R_2) \neq \emptyset$, $C(R_2) \cap C(R_3) \neq \emptyset$, while $C(R_1) \cap C(R_3) = \emptyset$. Then, clearly, d_0^R does not satisfy (M4). For the case of d_1^R, it suffices to note that d is a normalized metric and X is countable. □

Thus, as a consequence of the previous result, d_0^R meets the requirement of not penalizing, but instead allowing and promoting ambiguity: indeed, two rough clusterings are considered equivalent as longs as they have a compatible hard clustering in common. By contrast, d_1^R fully penalizes ambiguity, equating it to an error in clustering assignment, by declaring two rough clusterings to be equivalent if and only if all their compatible hard clusterings coincide.

More generally, if we define $d_\alpha^R = \alpha d_1^R + (1-\alpha) d_0^R$, for any $\alpha \in [0,1]$, then the following result holds:

Theorem 3.1. *Let $\alpha \in [0,1]$. Then:*

- *If $\alpha \geq \frac{1}{2}$, d_α^R is a metric;*
- *$\forall \rho \geq 1$, if $\alpha \geq \frac{1}{2\rho}$, d_α^R is a ρ-relaxed metric.*

Proof. That d_α^R is symmetric (i.e., satisfies (M3)) is evident from the definition. Similarly, evidently for each $\alpha > 0$ d_α^R satisfies (M1) and (M2). The result then follows from Example 2.2 and Lemma 2.3 in [48]. □

Intuitively, d_α^R can be understood as an intermediate measure considering the relative cost associated to ambiguity to be α: that is, α can be interpreted as the ratio of the cost of ambiguity over the cost of error. Indeed, consider the case where we compare a rough clustering R with a hard clustering C. Then, if $\alpha = 0$ it holds that $d_\alpha^R = d_0^R$, which equals 0 as long as $C \in C(R)$. This corresponds to assigning the *cost of ambiguity* to be equal to 0, since we ignore all $C' \in C(R) \setminus \{C\}$. By contrast, if $\alpha = 1$, then $d_\alpha^R = d_1^R$, which equals 0 iff $C(R) = \{C\}$: in particular, if R_1, R_2 are two rough clusterings with $C(R_1) \subset C(R_2)$, then $d_1^R(C, R_1) \leq d_1^R(C, R_2)$. Thus, when $\alpha = 1$, the cost of ambiguity is equated to the cost of an error, as every assignment of objects to the boundary of a cluster is penalized to the same degree as the assignment to an incorrect cluster. More generally, α represents the trade-off between ambiguity and error that the specific user is willing to tolerate for the application at hand, and it should be set accordingly. Thus, a user who is willing to tolerate a certain degree of ambiguity (hence, a potentially larger number of objects assigned to the boundary of some cluster), in order to reduce the risk of clustering errors should set α closer to 0. This could make sense, for example, in medical applications where some degree of ambiguity could be tolerated if it avoids clustering together objects that correspond to patients associated with different clinical characteristics. Conversely, a user who prefers obtaining a

clustering close to a hard one, and is willing to tolerate a potentially larger amount of cluster assignment errors, should set α closer to 1. Obviously, the previous discussion only provides guidelines to set the α parameter; its value should be carefully optimized to match the requirements of the application at hand.

Example 3.2. *Let C, R be the clusterings defined in Examples 2.1 and 2.4, and let $d = 1 - \mathsf{Rand}$. Then $d_0^R = 0$ and $d_1^R = 0.5$. Thus, for every $\alpha \in [0, 1]$ it holds that $d_\alpha^R \in [0, 0.5]$. In particular, for the case of $\alpha = 0.5$, we have $d_{0.5}^R = 0.25$. On the other hand, if d is the partition distance d_π, then $d_0^R = 0$ and $d_1^R = 0.4$. Thus, for every $\alpha \in [0, 1]$ it holds that $d_\alpha^R \in [0, 0.4]$. In particular, for the case of $\alpha = 0.5$, we have $d_{0.5}^R = 0.2$.*

By construction, d_0^R and d_1^R are well suited for the objective comparison between two soft clusterings, as they satisfy the required metric property of being (the dual of) a consistency and a metric. Unfortunately, from the computational point of view, the calculation of d_0^R and d_1^R is likely to be intractable, as shown by the following proposition.

Proposition 3.2. *Let R_1, R_2 be two rough clusterings represented through, either, the object-based or relational representations. Let $k = |\Omega|$ be the number of clusters, and $m = |\{x \in X : |R(X)| \neq 1\}|$. Then, the problem of computing d_0^R and d_1^R is* NP-HARD*: in particular, if k is constant, then both problems are fixed-parameter tractable with respect to the parameter m.*

By contrast, both problems are in P *if R_1, R_2 are represented through the distribution-based representation. In this latter case, the complexity is $o(k^m)$.*

Proof. For the case of d_0^R, this can be obtained by a reduction to integer programming. Computing $d_1^R(R_1, R_2)$ is equivalent to computing the Hausdorff distance between $C(R_1), C(R_2)$, which can be seen as a max-min 0-1 optimization problem. Since $C(R_1), C(R_2)$ are finite, this can be transformed to a min-max 0-1 optimization problem, which is NP-HARD [30]. The second part of the theorem easily follows by noting that d_0^R, d_1^R can easily be computed by enumerating all elements in $C(R_1), C(R_2)$. Finally, fixed parameter tractability of computing d_0^R, d_1^R derives from the two previous results. □

To address the problem of computational hardness, several approximations will be described in Section 4.

Extension to general soft clusterings. The previous approach can be extended to the cases of fuzzy, possibilistic and evidential clustering, by noting that all these three forms of soft clusterings can be expressed as probability distributions over rough clusterings. For the case of evidential clustering, this has already been shown in Eq (11). For the case of fuzzy clustering, it follows from Eq (12) that any fuzzy clustering can be represented as a distribution over rough clusterings whose corresponding sets of compatible clusterings are singletons. Finally, the case for possibilistic clustering directly follows from the fact that a possibilistic clustering is an evidential clustering in which all mass functions are

consonant. As a consequence, we can use an approach based on the Wasserstein Distance to extend d_0^R and d_1^R to evidential clustering. Namely, given two evidential clusterings M_1, M_2 and their corresponding focal sets $\mathcal{F}(M_1), \mathcal{F}(M_2)$, the general definition of the *transport-based measure* is given by

$$
\begin{aligned}
d_\alpha^E(M_1, M_2) = \min_{\sigma} & \sum_{(R_1,R_2)\in\mathcal{F}(M_1)\times\mathcal{F}(M_2)} \sigma(R_1, R_2) d_\alpha^R(R_1, R_2) \qquad (15)\\
\text{s.t.} & \sum_{R_2\in\mathcal{F}(M_2)} \sigma(R_1, R_2) = m_{M_1}(R_1)\\
& \sum_{R_1\in\mathcal{F}(M_1)} \sigma(R_1, R_2) = m_{M_2}(R_2)\\
& \sum_{(R_1,R_2)\in\mathcal{F}(M_1)\times\mathcal{F}(M_2)} \sigma(R_1, R_2) = 1\\
& \forall (R_1, R_2) \in \mathcal{F}(M_1) \times \mathcal{F}(M_2), \sigma(R_1, R_2) \geq 0.
\end{aligned}
$$

Intuitively, the transport-based measure based on d_α^R can be interpreted as the minimal cost of transforming the mass function over hard clusterings described by M_1 into the mass function described by M_2. This cost is computed by finding the joint mass function whose marginals are equal to M_1 and M_2, and which minimizes the expected value of $d_\alpha^R(R_1, R_2)$. Due to the properties of the Wasserstein distance, it is easy to show that the following properties hold:

Theorem 3.2. *Let $\alpha \in [0, 1]$. Then,*

- *If $\alpha \geq \frac{1}{2}$, d_α^E is a metric;*
- *$\forall \rho \geq 1$, if $\alpha \geq \frac{1}{2\rho}$, d_α^E is a ρ-relaxed metric;*
- *$1 - d_0^E$ is a consistency.*

Proof. The result is a direct consequence of Theorem 3.1 and the properties of the Wasserstein distance. □

Corollary 3.1. *Let F_1, F_2 be two fuzzy clusterings. Then, $\forall \alpha_1, \alpha_2 \in [0, 1]$ it holds that $d_{\alpha_1}^E(F_1, F_2) = d_{\alpha_2}^E(F_1, F_2)$.*

Proof. The result directly follows from the observation that, for a fuzzy clustering F the focal sets of m_F are all singletons. □

As d_α^R for rough clusterings, d_α^E can be used as an objective comparison measure for evidential clusterings. In particular, $1 - d_0^E$ provides a measure of consistency between two evidential clusterings, while d_1^E can be understood as an equality index for evidential clustering. Moreover, as previously discussed, d_α^E, with $\alpha \in [0, 1]$, can be interpreted as a measure of compatibility between two evidential clusterings, where the relative cost of ambiguity is exactly α. The transport-based measure generalizes the previous proposals in [18, 27], by providing a general framework to extend any hard clustering comparison measure

to the setting of soft clustering, thus enabling the direct comparison of any two soft clusterings, irrespective of their representation (either relational-based or object-based) and class (based either on rough, fuzzy, possibilistic or evidential clustering). Furthermore, it also allows a more flexible trade-off between ambiguity and partial assignment or clustering errors, by controlling the ambiguity parameter α.

Example 3.3. *Let C, F, P, M be the clusterings defined in Examples 2.1, 2.3, 2.5, and let $d = 1 - \mathsf{Rand}$. Then, it holds that:*

- *For all $\alpha \in [0,1]$ $d_\alpha^E(C,F) = 0.267$;*
- $d_0^E(C,P) = 0$, $d_1^E(C,P) = 0.48$;
- $d_0^E(C,M) = \frac{1}{12}$, $d_1^E(C,M) = 0.442$.

If, instead, we let $d = d_\pi$ we obtain:

- *For all $\alpha \in [0,1]$ $d_\alpha^E(C,F) = 0.23$;*
- $d_0^E(C,P) = 0$, $d_1^E(C,P) = 0.4$;
- $d_0^E(C,M) = 0.07$, $d_1^E(C,M) = 0.37$.

In regard to the computational complexity, since the problem of computing d_0^R, d_1^R is already computationally hard, it is evident that the problem of computing d_α^E is also computationally hard. An interesting problem, however, would be to determine whether, at least for the case of fuzzy clustering, a computationally efficient algorithm exists for solving Eq (15). We leave this problem for future work. In the following section, however, we show that d_0^E and d_1^E can be efficiently bounded by means of a polynomial-time algorithm in the cases of the Rand index and the partition distance.

As a last result in this section, we study the special case where one of the clusterings to be compared is a hard clustering C. This case is particularly interesting, since it frequently arises in applications. Indeed, in many cases, e.g., when one wants to test a novel soft clustering algorithm, one uses an existing ground truth clustering C as a reference for comparison. The following theorem shows that, in this particular case, the interval-valued transport-based measure between C and an evidential clustering M can be interpreted as the lower and upper expectations of the distance between C and the unknown hard clustering partially specified by M:

Theorem 3.3. *Let M, C be, respectively, an evidential clustering and a hard clustering. Let d be a normalized metric over hard clusterings, and let $\underline{E}(d), \overline{E}(d)$ be the lower/upper expectation of d with respect to m_M. Then $d_0^E(M,C) = \underline{E}(d)$, $d_1^E(M,C) = \overline{E}(d)$.*

Proof. First, we note that for all $R \in \mathcal{F}(M)$, it is easy to observe that under the conditions stated in the Theorem it holds that $d_1^R = \max_{C' \in C(R)} d(C, C')$. Since C is a hard clustering, $m_C(\{C\}) = 1$. Therefore, the result follows by noting that Eq (15) reduces to the expectation, with respect to m_M, of d_α^R. □

4. Approximation Methods

In the previous section, we proposed a general approach, the transport-based measures, to extend hard clustering comparison measures to the case of soft clustering. This approach can be used to obtain objective comparison criteria, by applying the Wasserstein construction to the distribution-based representation of the soft clusterings to be compared. Nonetheless, the computation of the transport-based measure is generally intractable. For this reason, in this section, we introduce some approximation methods and algorithms. First, in Section 4.1, we describe a general approach based on sampling, which can be applied to any base distance between hard clusterings. In Sections 4.2 and 4.3, we then discuss generalizations of the Rand index and the partition distance, as representatives, respectively, of relational-based and object-based comparison measures, and we show that they can be used to approximate the transport-based measure.

4.1. Sampling-based Approximation Algorithms

In this section, we provide a general sampling-based approach that can be used to approximate the value of d_α^E. We start with the case of rough clustering, that is with d_0^R, d_1^R. Assume that, given two rough clusterings R_1 and R_2, we draw s samples $(C_1^1, C_2^1), \dots, (C_1^s, C_2^s)$ uniformly from $C(R_1)$ and $C(R_2)$. Then, we can approximate $\hat{d}_0^R = \min_{i\in\{1,\dots,s\}} d(C_1^i, C_2^i)$ and $\hat{d}_1^R = d_H(\{C_1^i\}_{i=1}^s, \{C_2^i\}_{i=1}^s)$. It is easy to show that the following result holds:

Proposition 4.1. *The following bounds hold for any $t > 0$:*

$$Pr(d_1^R - \hat{d}_1^R > \epsilon) \leq F(d_1^R - \epsilon)^s \tag{16a}$$

$$Pr(\hat{d}_0^R - d_0^R > \epsilon) \leq 1 - \left(1 - F(\epsilon - d_0^R)\right)^s, \tag{16b}$$

where F is the cumulative distribution function (CDF) of the probability distribution p_R defined as

$$p_R(t) = \frac{|\{C_1 \in C(R_1), C_2 \in C(R_2) : d(C_1, C_2) = t\}|}{|d_R(R_1, R_2)|}. \tag{17}$$

Noting that $F(d_1^R - t)$ (resp., $F(t - d_1^R)$) is stricly less than 1, it holds that, for each ϵ, $Pr(d_1^R - \hat{d}_1^R > \epsilon)$ (resp. $Pr(\hat{d}_0^R - d_0^R > \epsilon)$) has exponential decay in the size of the sample s.

Proof. The result directly follows from the distribution of the order statistics $\hat{d}_0^R, \hat{d}_1^R$. □

Thus, even though computing $\langle d_0^R, d_1^R \rangle$ is computationally hard, we can obtain good approximations of its value by a simple sampling procedure that can be easily implemented in polynomial time.

Remark. *We note that, despite the previous bounds, the quality of the approximation largely depends on $p_R(t)$. In particular, the convergence in Eq* (16) *is*

correlated with the tailedness *of p_R defined in Eq* (17)*: the heavier the tails of p_R, the lower the approximation error. This can be directly understood by looking at the bounds in Eq.* (16)*: indeed, if the tails of p_R are thin, than $F(d_1^R - \epsilon)$ will be close to 1, thus resulting in a large approximation error.*

The problem mentioned in the previous remark can be illustrated through the following example.

Example 4.1. *Let C be the hard clustering defined in Example 2.1, let R' be the rough clustering s.t. $\forall x \in X, R(x) = \Omega$, and let $d = 1 - \mathsf{Rand}$. Then, clearly, $d_0^R = 0$ and $d_1^R = 1$. However, the probability distribution p_R will be concentrated around the expected value of $1 - \mathsf{Rand}$ under the uniform distribution [44]. Consequently, $\hat{d}_0^R$ and $\hat{d}_1^R$ will also be close to this value with high probability.*

For the case of fuzzy clustering, let

$$d_{F_1,F_2}(v) = \sum_{C_1,C_2:d(C_1,C_2)=v} Pr_{F_1}(C_1) \cdot Pr_{F_2}(C_2).$$

If we use a sampling procedure to estimate $d_\alpha^E(F_1, F_2)$, we can obtain a stronger tail bound by applying Hoeffding's inequality:

Proposition 4.2. *Assume that d is a normalized metric on hard clusterings, and F_1, F_2 are two fuzzy clusterings. Then,*

$$Pr(|\hat{d}_\alpha^E(F_1, F_2) - d_\alpha^E(F_1, F_2)| \geq \epsilon) \leq 2e^{-2s\epsilon^2}. \tag{18}$$

Hence, the deviation has exponential decay in the size of the sample s.

Combining Eqs (16) and (18), an analogous result can be found also for the case of d_α^E. Indeed, we obtain:

Proposition 4.3. *Assume that d is a normalized metric on hard clusterings. Let $\hat{d}_0^E, \hat{d}_0^E$ be the sample estimates of d_0^E, d_1^E. Then:*

$$Pr(|\hat{d}_0^E - d_0^E| \geq \epsilon) \leq 2e^{-2s\epsilon^2} \tag{19a}$$

$$Pr(|\hat{d}_1^E - d_1^E| \geq \epsilon) \leq 2e^{-2s\epsilon^2}. \tag{19b}$$

The previous bound, however, assumes that $\hat{d}_0^E, \hat{d}_1^E$ are computed by sampling pairs R_1, R_2 of rough clusterings from the distributions m_{M_1}, m_{M_2} and then computing the exact values of $d_0^R(R_1, R_2), d_1^R(R_1, R_2)$. As a consequence of Proposition 3.2, this may not be feasible when $|X|$ is large. In such cases, a possible solution would be to compute $\hat{d}_0^E, \hat{d}_1^E$ using a nested sampling procedure (i.e, first we sample a rough clustering R from m_M, then we sample a hard clustering C from $C(R)$). In this case, however, one should expect a larger approximation error as a consequence of our previous remarks.

Finally, we note that all the above mentioned sampling-based approximation methods can easily be implemented in polynomial time, more precisely with time complexity $O(n^2 s + s^3 \log s)$, where $n = |X|$ and s is the sample size. In particular, the cost of sampling is $\Theta(n^2 s)$, while the cost of computing the approximated transport-based distance is $O(s^3 \log s)$ [5].

4.2. Bounds for the Transport-based Rand Index

In this section, we discuss a generalization of the Rand index to evidential clustering that can be computed in polynomial time. Remarkably, we will show that this alternative definition allows us to bound the transport-based measure, when we assume as base distance $d(C_1, C_2) = 1 - \mathsf{Rand}(C_1, C_2)$.

The approach described in this section is based on the following observation: the hardness of computing the transport-based measure derives from it being defined as the Wasserstein distance between two probability distributions whose supports have size that is exponential in the size of the original soft clusterings. Thus, intuitively, the complexity could be reduced if, instead of computing the Wasserstein distance between the distributional representations of the two soft clusterings to be compared, this computation would be *pushed inside* the formula of the Rand index.

This intuition can be formalized by noting that the Rand Index can be seen as an instance of the Wasserstein metric. Indeed, consider two objects x, y and two hard clusterings C_1, C_2. Using the relational representation, which is the only information used to compute the Rand index, we can represent the hard clusterings as point masses $p_1^{x_i,x_j}, p_2^{x_i,x_j}$ on $\Theta = \{s, \neg s\}$, as shown in Section 2.2. Namely, $p_1^{x_i,x_j}$ and $p_2^{x_i,x_j}$ denote whether objects x_i and x_j are in the same cluster (in which case $p^{x_i,x_j}(s) = 1$) or not (in which case $p^{x_i,x_j}(\neg s) = 1$), for hard clusterings C_1, C_2. Then, the Rand index can be expressed as

$$\mathsf{Rand}(C_1, C_2) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} 1 - r^{x_i,x_j}, \tag{20}$$

where

$$\begin{aligned} r^{x_i,x_j} = & \min_{\sigma} \sum_{(a,b)\in\Omega^2} \sigma(a,b) d(a,b) \\ \text{s.t.} \quad & \sum_{b\in\Omega} \sigma(a,b) = p_1^{x_i,x_j}(a) \\ & \sum_{a\in\Omega} \sigma(a,b) = p_2^{x_i,x_j}(b) \\ & \sum_{(a,b)\in\Omega^2} \sigma(a,b) = 1 \\ & \forall (a,b) \in \Omega^2, \sigma(a,b) \geq 0, \end{aligned} \tag{21}$$

with $d(a,b) = \mathbb{1}_{a \neq b}$.

It can be easily seen that $r_{x_i,x_j} = |p_1^{x_i,x_j}(s) - p_2^{x_i,x_j}(s)| = |p_1^{x_i,x_j}(\neg s) - p_2^{x_i,x_j}(\neg s)|$, which coincides with Eq (7). We note that the same approach can be equivalently applied to probabilistic and fuzzy clusterings, by simply relaxing the requirement that $p_1^{x_i,x_j}, p_2^{x_i,x_j}$ should be point masses. The solution of Eq (15) is still $r_{x_i,x_j} = |\mu_1^{x_i,x_j}(s) - \mu_2^{x_i,x_j}(s)| = |\mu_1^{x_i,x_j}(\neg s) - \mu_2^{x_i,x_j}(\neg s)|$. Thus, the approach proposed in [27] can be derived as a special case to our definition.

To extend the above approach to evidential clustering, we must consider a generalization of the metric d to sets on Θ. If we apply Eq (14) to the case where $A, B \subseteq \Theta$ we obtain

$$d_0(A,B) = \mathbb{1}_{A \neq B \wedge A \cap B = \emptyset} \tag{22a}$$
$$d_1(A,B) = \mathbb{1}_{A \neq B} \tag{22b}$$
$$d_\alpha(A,B) = \alpha d_1^R(A,B) + (1-\alpha) d_0^R(A,B). \tag{22c}$$

As shown in Section 3.2, d_1 is a metric, while $1 - d_0$ is a consistency.

Based on d_α we can generalize Eq (20) to evidential clustering, obtaining

$$\begin{aligned} r_\alpha^{x_i,x_j} = \min_m \sum_{(A,B) \in 2^\Theta \times 2^\Theta} & m(A,B) d_\alpha(A,B) \\ \text{s.t.} \quad & \sum_{B \in 2^\Theta} m(A,B) = m_1^{x_i,x_j}(A) \\ & \sum_{A \in 2^\Theta} m(A,B) = m_2^{x_i,x_j}(B) \\ & \sum_{A,B \in 2^\Theta} m(A,B) = 1 \\ & \forall A, B \in 2^\Theta, m(A,B) \geq 0, \end{aligned} \tag{23}$$

where $\alpha \in [0,1]$ represents, as in the previous sections, the relative cost of ambiguity with respect to the cost of error. The *α-Rand index* can then be defined as follows.

Definition 4.1. *Let $\alpha \in [0,1]$, M_1, M_2 be two evidential clusterings, $n = |X|$ the number of objects and $r_\alpha^{x_i,x_j}$ be defined as above. Then,*

$$\mathsf{Rand}_\alpha(M_1, M_2) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} 1 - r_\alpha^{x_i,x_j}. \tag{24}$$

We note that, in general, $\alpha_1 \leq \alpha_2 \implies \mathsf{Rand}_{\alpha_2} \leq \mathsf{Rand}_{\alpha_1}$. Thus, in particular $\mathsf{Rand}_1 \leq \mathsf{Rand}_0$. Furthermore, in regard to the computational complexity, since the size of the frame Θ is constant, $r_\alpha^{x_i,x_j}$ can be computed for each pair $(x_i, x_j) \in X^2$ in time $O(1)$. Therefore, Rand_α can be computed in time $O(n^2)$.

Example 4.2. *Let C, R, F, P, M be the clusterings defined in Examples 2.1, 2.3, 2.4 and 2.5. Then:*

- $\mathsf{Rand}_1(C,R) = 0.52$ *and* $\mathsf{Rand}_0(C,R) = 1$*;*
- *For all* $\alpha \in [0,1]$, $\mathsf{Rand}_\alpha(C,F) = 0.79$*;*
- $\mathsf{Rand}_1(C,P) = 0.54$ *and* $\mathsf{Rand}_0(C,P) = 1$*;*
- $\mathsf{Rand}_1(C,M) = 0.55$ *and* $\mathsf{Rand}_0(C,M) = 0.95$*.*

We note that only for the cases of $\mathsf{Rand}_0(C,R)$ *and* $\mathsf{Rand}_0(C,P)$ *does the value of the approximation coincide with the exact solution of Eq* (15)*, which was computed in Examples 3.2, 3.3.*

As a first result, we characterize the relationship between Rand_α and Eq (15), showing that Rand_α allows to bound the value of the latter when $d = 1 - \mathsf{Rand}$:

Theorem 4.1. *Let* d_0^E, d_1^E *be defined as in Eq* (15)*, by setting* $d = 1 - \mathsf{Rand}$*. Then* $\mathsf{Rand}_1 \leq 1 - d_0^E \leq \mathsf{Rand}_0$ *and* $1 - d_1^E \leq 1 - d_0^E \leq \mathsf{Rand}_0$*.*

Proof. We first prove the inequality $1 - d_0^E \leq \mathsf{Rand}_0$. First, assume that M_1, M_2 are both rough clusterings. Then d_0^R is obtained by choosing two clusterings C_1, C_2 s.t. $C_1 \in C(M_1), C_2 \in C(M_2)$ and $d(C_1, C_2) = 1 - max_{C_1', C_2'} \mathsf{Rand}(C_1, C_2)$. Let x_i, x_j, with $i < j$, be any pair of objects considered in the computation of $\mathsf{Rand}(C_1, C_2)$ and consider $r_{C_1,C_2}^{x_i,x_j}$. Note that $r_{C_1,C_2}^{x_i,x_j} \in \{0,1\}$ and $r_{C_1,C_2}^{x_i,x_j} = 1 \implies r_0^{x_i,x_j}$. Therefore, the inequality follows in the case of rough clustering. The general case for evidential clustering follows by noting that an evidential clustering is a probability distribution over rough clustering and from the definition of the Wasserstein distance. For the inequality $\mathsf{Rand}_1 \leq 1 - d_0^E$ we can apply a similar technique, noting that $r_{C_1,C_2}^{x_i,x_j} = 0 \implies r_1^{x_i,x_j} = 0$. Therefore, the result follows. □

The previous result shows that we can bound d_0^E, d_1^E by computing Rand_α, which can be done in time $O(n^2)$ instead of $O(2^n)$. We leave it as an open problem to characterize the quality of the approximation provided by Rand_α, that is, to determine whether it is possible to find bounds for the approximation error $\epsilon = |d_\alpha^E - \mathsf{Rand}_\alpha|$.

Next, we study the properties of Rand_α. First, we note that when M_1, M_2 are fuzzy clusterings, the following result holds.

Theorem 4.2. *Let* F_1, F_2 *be fuzzy clusterings. Then, for all* $\alpha \in [0,1]$*,* $\mathsf{Rand}_\alpha = \mathsf{Rand}_0 = \mathsf{Rand}_1$*.*

Proof. The result can be easily obtained by noting that d_0, d_α, d_1 are all equivalent when restricted to $\{s, \neg s\}$. □

In terms of metric properties, the following result holds:

Theorem 4.3. *Assume that evidential clusterings are represented through the relational representation, and* $\alpha \in [0,1]$*. Then,*

- *If* $\alpha \geq \frac{1}{2}$*,* Rand_α *is a similarity;*
- Rand_0 *is a consistency;*
- $\forall \rho \geq 1$*, if* $\alpha \geq \frac{1}{2\rho}$*,* $1 - \mathsf{Rand}_\alpha$ *is a* ρ*-relaxed metric.*

Proof. These claims directly derives from the properties of d_α (see Theorem 3.1) and the definition of the Wasserstein distance. □

By contrast, if we consider the object-based representation, then Rand_α is, at most, a pseudo-similarity:

Corollary 4.1. *Assume that evidential clusterings are represented through the object-based representation, and $\alpha \in [0,1]$. Then,*

- *If $\alpha \geq \frac{1}{2}$, Rand_α is a pseudo-similarity;*
- *$\forall \rho \geq 1$, if $\alpha \geq \frac{1}{2\rho}$, $1 - \mathsf{Rand}_\alpha$ is a ρ-relaxed pseudo-metric.*

Proof. To show that the claim holds, it is sufficient to observe that the map from the object-based representation to the relational representation is not injective. Indeed, as long as $n < 2^k - 1$, to any possible m^{x_i,x_j} may correspond infinitely many evidential clusterings M, since the associated multi-linear system (given by Eq (3)) is under-determined. Then clearly, for any two such M_1, M_2 it holds that $1 - \mathsf{Rand}_\alpha(M_1, M_2) = 0$. Thus, combining this result with Theorem 4.3, the claim follows. □

Remark. *The proof of Corollary 4.1 suggests the conjecture that Rand_α may be a similarity also in terms of the relational representation, if we require $n \geq 2^k - 1$. While it can be easily shown by direct algebraic manipulation that this property indeed holds when $k = 2$, we leave the proof of this conjecture in the general case as future work.*

We have seen that the cost of computing Rand_α is $O(n^2)$. While in general, computing the Wasserstein distance requires solving a linear programming problem, when m_1, m_2 are normalized, for the special cases of $r_0^{x_i,x_j}, r_{0.5}^{x_i,x_j}$ and $r_1^{x_i,x_j}$ we can find a closed-form solution. In particular, we have the following theorem.

Theorem 4.4. *Let $r_\alpha^{x_i,x_j}$ be defined as in Eq (23) and assume $m_1^{x_i,x_j}, m_2^{x_i,x_j}$ are normalized (i.e., $m_1^{x_i,x_j}(\emptyset) = m_2^{x_i,x_j}(\emptyset) = 0$). Then:*

$$
\begin{aligned}
r_0^{x_i,x_j} &= \frac{1}{2}\left[|m_1^{x_i,x_j}(s) + m_2^{x_i,x_j}(\neg s) - 1| + |m_1^{x_i,x_j}(\neg s) + m_2^{x_i,x_j}(s) - 1|\right. \\
&\qquad \left. - m_1(\Theta) - m_2(\Theta)\right] \\
r_1^{x_i,x_j} &= \frac{1}{2}\sum_{A \subseteq \Theta} |m_1^{x_i,x_j}(A) - m_2^{x_i,x_j}(A)| \\
r_{0.5}^{x_i,x_j} &= \frac{1}{2}\left[|Bel_1^{x_i,x_j}(s) - Bel_2^{x_i,x_j}(s)| + |Bel_1^{x_i,x_j}(\neg s) - Bel_2^{x_i,x_j}(\neg s)|\right].
\end{aligned}
$$

Proof. Consider first $r_1^{x_i,x_j}$. By the definition of the Wasserstein distance and Θ, the optimal assignment is given by setting for each $A \subseteq \Theta$ $m(A,A) = \min\{m_1^{x_i,x_j}(A), m_2^{x_i,x_j}(A)\}$ and then arbitrarily allocating the remaining mass. Therefore, $r_\alpha^{x_i,x_j} = \frac{1}{2}\sum_{A \subset \Theta} \max\{m_1^{x_i,x_j}(A), m_2^{x_i,x_j}(A)\} - m(A,A)$, from which the formula in the theorem directly follows. For the case of $r_{0.5}^{x_i,x_j}$, we note that the optimal assignment is given by any joint mass function that maximizes $\sum_{A \subseteq \Theta}\left[m(A,A) + \frac{m(A,\Theta)+m(\Theta,A)}{2}\right]$. In particular, the assignment obtained by

setting $m(A, A) = \min\{m_1^{x_i,x_j}(A), m_2^{x_i,x_j}(A)\}$, then allocating masses to sets of the form $m(A, \Theta), m(\Theta, A)$, and then arbitrarily allocating the remaining mass satisfies the above mentioned maximization problem. The result then easily follows from the same arguments used for $r_1^{x_i,x_j}$. Finally, for the case of $r_0^{x_i,x_j}$ we note that optimal assignment is given by any joint mass function m that maximizes $\sum_{A\subseteq\Theta} m(A, A) + m(A, \Theta) + m(\Theta, A)$. In particular, for any such mass function, $r_0^{x_i,x_j} = 1$ iff $m_1^{x_i,x_j}(A) = m_2^{x_i,x_j}(A^c) = 1$, while $r_0^{x_i,x_j} = 0$ iff $m_1^{x_i,x_j}(A)+m_1^{x_i,x_j}(\Theta) = m_2^{x_i,x_j}(A)+m_2^{x_i,x_j}(\Theta) = 1$, for some $A \subseteq \Theta$. The same assignment described for the case of $r_{0.5}^{x_i,x_j}$ satisfies the above mentioned properties. The result then follows from algebraic manipulations and the definition of the Wasserstein distance. □

Finally, as a consequence of the previous result, we obtain the following corollary.

Corollary 4.2. *Let M_1, M_2 be two evidential clusterings such that, for each pair of objects $x_i, x_j \in X$, $m_1^{x_i,x_j}$ and $m_2^{x_i,x_j}$ are normalized. Then, $\mathsf{Rand}_{0.5} = \mathsf{Rand}_B$, where Rand_B is obtained by using the belief distance [18] in Eq* (9).

This result explains why Rand_B, defined in [18], was found to be unable to distinguish ambiguity from partial assignment and provides an intuitive interpretation of Rand_α. Indeed, as previously mentioned in Section 3.2, α can be interpreted as the (relative) *cost of ambiguity*. Therefore, as a consequence of Theorem 4.3, Rand_B is equivalent to the case where the cost of ambiguity is $\alpha = 0.5$, i.e., the same cost assigned to a uniformly randomized assignment of objects to clusters. Indeed, let m_1 be a logical mass function, and let $m_2(\Omega) = 1$ and $m_2'(s) = m_2'(\neg s) = \frac{1}{2}$. Then, in both cases, $\mathsf{Rand}_{0.5} = \frac{1}{2}$. By contrast, it is easy to observe that Rand_0 conflates ambiguity and correctness: indeed, for the previous example we have $\mathsf{Rand}_0 = 1$ for m_1 and m_2. Finally, Rand_1 conflates ambiguity and error: for m_1 and m_2, we have $\mathsf{Rand}_1 = 0$. Thus, in conclusion: if we let α vary in $(0, 0.5)$ we obtain a measure in which ambiguity interpolates between correctness and uncertainty, while if we let α vary in $(0.5, 1)$ we obtain a measure in which ambiguity interpolates between partial assignment and error.

4.3. Bounds for the Transport-based Partition Distance

In this section, we discuss a generalized version of the partition distance which satisfies the properties we required for a comparison measure between soft clusterings. As for the Rand index, this generalization of the partition distance can be computed in polynomial time and can be used to bound the value of the transport-based measure defined in Section 3.2, when the base distance is d_π (that is, the partition distance between hard clusterings).

As in the previous section, the main idea underlying the bounding approach is to push the computation of the Wasserstein distance inside the definition of the partition distance. For this purpose, we note that the partition distance can

be formulated equivalently as

$$\delta_C(C_1, C_2) = \min_w \frac{1}{2(|X|-1)} \sum_{i=1}^{k} \sum_{x \in X} \Delta(\chi_{\omega_1^i}(x), \chi_{\omega_2^{w(i)}}(x)), \tag{25}$$

where $\chi_\omega(\cdot)$ is the indicator function corresponding to cluster ω, and

$$\Delta(\chi_{\omega_1^i}(x), \chi_{\omega_2^{w(i)}}(x)) = |\chi_{\omega_1^i}(x) - \chi_{\omega_2^{w(i)}}(x)|$$

is the symmetric difference operator on indicator functions. We note that this approach cannot be directly extended to the case of evidential clustering, since an evidential clustering M is usually represented as a collection of functions $\{m_x\}_{x \in X}$, where $m_x : 2^\Omega \mapsto [0,1]$, while the formulation in Eq (25) would require a collection of functions $\{m_\omega\}_{\omega \in \Omega}$, with $m_\omega : X \mapsto [0,1]$. Nonetheless, the desired representation can be obtained by means of a change of frame. Indeed, for each cluster $\omega \in \Omega$, we consider the frame $\Theta_\omega = \{\omega, \neg\omega\}$ and the restriction of m_x to Θ_ω given by

$$m_x^{C\downarrow\Theta_\omega}(\{\omega\}) = m_x(\omega) \tag{26}$$

$$m_x^{C\downarrow\Theta_\omega}(\{\neg\omega\}) = \sum_{\emptyset \neq A \subseteq \Omega : \omega \notin A} m_x(A) \tag{27}$$

$$m_x^{C\downarrow\Theta_\omega}(\Theta_\omega) = \sum_{A \subseteq \Omega : \{\omega\} \subset A} m_x(A) \tag{28}$$

$$m_x^{C\downarrow\Theta_\omega}(\emptyset) = m_x(\emptyset). \tag{29}$$

If we define $m_\omega(x) = m_x^{C\downarrow\Theta_\omega}(\{\omega\})$, then a generalization of the operator Δ to the case of evidential clustering can easily be obtained as a solution to the Wasserstein problem, by assuming a base distance on Θ_ω. If we assume the same base distances on Θ_ω as given in Eqs (22), then Δ can be computed as a solution to the optimal transport problem, obtaining the same formulation as in Theorem 4.4. Based on the definition of Δ_α, we can generalize Eq (25) to the case of evidential clustering as

$$\delta_\alpha^E(M_1, M_2) = \min_\tau \frac{1}{2(n-1)} \sum_{i=1}^{k} \sum_{x \in X} \Delta_\alpha \left(m_x^{C\downarrow\Theta_{\omega_1^i}}, m_x^{C\downarrow\Theta_{\omega_2^{\tau(i)}}} \right), \tag{30}$$

where $\Delta_\alpha(m_1, m_2)$ is the distance function between mass functions that we previously defined, and τ is a permutation of $\{1, \ldots, k\}$. We can remark that, when M_1 and M_2 are hard clusterings, then Eq (30) is equivalent to Eq (25), hence the former is a generalization of the partition distance to evidential clustering.

Then, we characterize the relation between δ_α^E and transport-based measure when $d = d_\pi$. As for the Rand index, we show that the measure defined in this section can be used to provide bounds for the transport-based measure:

Theorem 4.5. *Let d_0^E, d_1^E be defined as in Eq (15), by setting d to be the partition distance d_π. Then $\delta_0^E \leq d_0^E \leq \delta_1^E$ and $\delta_0^E \leq d_0^E \leq d_1^E$.*

Proof. The proof is similar to the one for Theorem 4.1 and is therefore omitted. □

As for the case of the Rand index, we leave as open problem to find bounds for the approximation error $\epsilon = |d_\alpha^E - \delta_\alpha^E|$.

Then, we study the properties of δ_α^E. It can be easily proved that δ_α^E satisfies the following properties:

Theorem 4.6. *Let $\alpha \in [0, 1]$. Then,*

- *If $\alpha \geq \frac{1}{2}$, δ_α^E is a metric;*
- *$1 - \delta_0^E$ is a consistency;*
- *$\forall \rho \geq 1$, if $\alpha \geq \frac{1}{2\rho}$, δ_α^E is a ρ-relaxed metric.*

Proof. The claims derive easily from the properties of $\Delta_\alpha = r_\alpha^{x_i, x_j}$, the partition distance and the Wasserstein distance. □

Example 4.3. *Let C, R, F, P, M be the clusterings defined in Examples 2.1, 2.3, 2.4 and 2.5. Then,*

- *$\delta_0^E(C, R) = 0$ and $\delta_1^E(C, R) = 0.5$;*
- *For all $\alpha \in [0, 1]$, $\delta_1^E(C, F) = 0.23$;*
- *$\delta_0^E(C, P) = 0$ and $\delta_1^E(C, P) = 0.48$;*
- *$\delta_0^E(C, M) = 0.07$ and $\delta_1^E(C, M) = 0.47$.*

We note that for the cases of $\delta_0^E(C, R), \delta_0^E(C, R), \delta_0^E(C, M)$ and $\delta_\alpha^E(C, F)$, the values of the approximation coincide with the exact solution of Eq (15).

Also, δ_α^E can be computed in polynomial time:

Proposition 4.4. *The complexity of computing δ_α^E is $O(n2^k + k^3)$.*

Proof. Computing δ_α^E first requires transforming the evidential clusterings M_1 and M_2 in the representation described in Eqs (26)-(29). This transformation can be performed in time $O(n2^k)$. As shown in Eq (23), the value of Δ_α can be computed in $O(1)$ time. Since the value of Δ_α must be computed for each pair of clusters $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$, this requires a total time of $O(k^2)$, where $k = \max\{|\Omega_1|, |\Omega_2|\}$. Thus, the inner sum of Eq (30) has total complexity $O(nk^2)$. Finally, Eq (30), once the values of the inner loop have been computed, can be solved through any standard algorithm for the weighted balanced assignment problem (e.g., the Hungarian method) in time $O(k^3)$. □

We note that, even though the term $n2^k$ is exponential in the number of clusters k, the computation of δ_α^E is still polynomial in the size of the evidential clusterings M_1, M_2, since in general these latter have size which is exponential in k. Furthermore, we can observe that, in terms of computational complexity,

the method discussed in this section and the one introduced in Section 4.2 are quite different. Indeed, the method for the Rand index discussed in Section 4.2 has complexity which is quadratic in the number of objects, i.e., $O(n^2)$. By contrast, the generalization of the partition distance introduced in this section has complexity which scales as $O(n2^k + k^3)$. This means that, since in general $k \ll n$, computing the (approximated) partition distance can be expected to be more computationally efficient than computing the (approximated) Rand index.

Interestingly, it is easy to show that the approach proposed in [3, 50] arises as a special case of our definition:

Proposition 4.5. *Let F_1, F_2 two fuzzy clusterings. Then, for each $\alpha \in [0, 1]$ $\delta_\alpha^E(F_1, F_2) = d_\pi^F(F_1, F_2)$.*

Proof. In the case of fuzzy clustering we have that, for each pair of clusters ω_1, ω_2, $\Delta_\alpha = \left(m_x^{C\downarrow\Theta_{\omega_1}}, m_x^{C\downarrow\Theta_{\omega_2}}\right) = |\omega_1(x) - \omega_2(x)|$. The result then follows from the definition of the fuzzy partition distance given in [3, 50]. □

5. Illustrative Experiments

In this section, we discuss two simple experiments, with the aim of illustrating the application of the proposed approach and the corresponding approximation methods. In particular, in the first experiment reported in Section 5.1, we compare five different clustering algorithms (based on the k-means clustering procedure) on a small-dimensional benchmark dataset, and we illustrate the computation of the metrics defined in Section 3, as well as its approximations introduced in Section 4. In Section 5.2, we then show through a second experiment how even for a moderately large dataset the approximations defined in Section 4 can be computed in reasonable time, and how this information can be used to bound model performance, even when the transport-based measure cannot be computed.

5.1. Iris Data

In the first experiment, we provide a simple illustration of the proposed metrics using the Iris dataset, a small-scale benchmark problem with 150 objects, four numerical features and three classes, each of which containing 50 instances. We selected this dataset as it is widely known that the three above mentioned classes are approximately linearly separable (see Figure 1). Thus, we expect any soft clustering algorithm to be able to find a clustering of the data in which most objects are precisely assigned to a single cluster. We note that, as a consequence of Theorem 3.2, this is a necessary condition for the exact versions of our proposed metrics to be computable in a reasonable time. The Iris dataset was selected specifically to allow direct computation of the exact metrics proposed in Section 3.2, without incurring running time-related bottlenecks.

For the purpose of our experiment, we considered five different clustering algorithms, all in the k-means family of algorithms. Namely, we considered: k-means (KM), rough k-means (RKM) [38], fuzzy c-means (FCM) [6], possibilistic

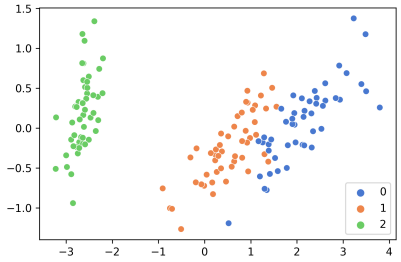

Figure 1: Graphical visualization of the Iris dataset in PCA space, with respect to the first two principal components. As shown by the figure, the classes are almost linearly separable, with a small overlap between classes 1 and 2.

c-means (PCM) [31] and evidential c-means (ECM) [36]. For all algorithms, in order to reduce the computational complexity of computing the exact value of the measures, we set the hyper-parameters so as to obtain a soft clustering as close as possible to a hard one. In particular, for RKM we set $\epsilon = 1.1$, for FCM and PCM we set $m = 5$, and for ECM we set $\delta = 10, \beta = 5, \alpha_{ECM} = 5$. We compared the output of each algorithm with the ground truth labeling of the Iris dataset. We considered, in particular, the following metrics:

- The transport-based Rand index (T-RI);
- The transport-based partition distance (T-PD);
- The sampling-based approximations (S-RI, S-PD) of the two previous measures (see Section 4.1). The sampling-based approximations were computed based on drawing 1000 samples with replacement;
- The bounds on the Rand index (A-RI) (see Section 4.2);
- The bounds on the partition distance (A-PD) (see Section 4.3).

For all metrics, the ambiguity parameter α was set to 0.5, to obtain a balanced trade-off between ambiguity and error, with the former having half the weight of the latter. All code was implemented in Python (v. 3.8.8), using the scikit-learn (v. 0.24.1), numpy (v. 1.20.1) and scipy (v. 1.6.2) libraries[2]. In particular, in the implementation of T-RI and S-RI we used as sub-routine the scikit-learn implementation of the Rand index (which has complexity $O(n)$, where n is the number of objects), while for T-PD, S-PD, A-RI and A-PD we used custom implementations based on the algorithms described in the previous sections (in particular, we note that the complexity of A-RI is $O(n^2)$). As mentioned previously, the exact transport-based measures T-RI and T-PD were computed by

[2]The code is freely accessible at `https://github.com/AndreaCampagner/scikit-cautious`.

Table 2: Results of the first experiment. For the Rand index higher is better, while for the partition distance lower is better.

Metric	KM	RKM	FCM	PCM	ECM
T-RI	0.877 (0.034s)	(0.874, 0.886) (0.802s)	0.876 (784.388s)	(0.839, 0.941) (979.053s)	(0.781, 0.944) (1394.69s)
S-RI	-	(0.874, 0.886) (0.429s)	0.876 (11.266s)	(0.860, 0.927) (19.848s)	(0.681, 0.819) (19.845s)
A-RI	-	(0.871, 0.889) (6.949s)	0.876 (6.924s)	(0.826, 0.966) (7.146s)	(0.747, 0.887) (7.024s)
T-PD	0.111 (0.031s)	(0.099, 0.113) (0.803s)	0.112 (184.707s)	(0.033, 0.122) (224.31s)	(0.041, 0.229) (431.57s)
S-PD	-	(0.100, 0.113) (0.202s)	0.112 (11.391s)	(0.072, 0.103) (13.739s)	(0.154, 0.209) (16.424s)
A-PD	-	(0.099, 0.113) (1.807s)	0.112 (1.665s)	(0.033, 0.136) (2.926s)	(0.039, 0.221) (3.386s)

direct application of their definitions given in Section 3.2. This direct computation is made possible by the fact that the classes of the Iris datasets are approximately linearly separable. Indeed, even though, as claimed in Theorem 3.2, the complexity of doing so is in general exponential in the number of clusters and the number of ambiguous objects, computing T-RI and T-PD is feasible when these quantities are small, as the corresponding problems are fixed-parameter tractable.

The results of the experiment are reported in Table 2, in terms of the metrics values as well as the running time (in seconds). As shown in the table, the approximation algorithms (both the sampling-based algorithms, as well as the ad hoc algorithms for the Rand index and the partition distance) were much more efficient than the exact versions of the metrics, for all algorithms except RKM, in which the ad hoc approximation algorithms reported worse running time than the exact versions. This follows from the observation that the result of RKM was very close to a hard clustering, with only two objects not assigned to a precise cluster. The difference in performance then follows by noting that the scikit-learn implementation of the Rand index has time complexity $O(n)$, while A-RI has time complexity $O(n^2)$.

In terms of running time, we can observe that the cost of computing the exact versions of the proposed measures sharply increases when considering more general soft clustering algorithms. Indeed, the running time of T-RI and T-PD for ECM were approximately twice the respective running times for either FCM and PCM. On the other hand, the differences in running times for the approximation algorithms (S-RI, S-PD, A-RI, A-PD) were much smaller, and indeed the running times for FCM, PCM and ECM were similar.

In terms of approximation quality, even though for RKM and FCM there were no differences between the sampling-based (i.e., S-RI, S-PD) and the ad hoc algorithms (i.e., A-RI, A-PD), this was not the case for PCM and ECM. In these latter cases, the ad hoc algorithms reported lower approximation error than the sampling-based ones. This observation can be understood as a consequence of the remark in Section 4.1, in which we discussed the quality of

approximation of the sampling-based algorithms. Thus, although the sampling-based methods have explicit bounds on the approximation error, the ad hoc algorithms introduced in Section 4.2 and 4.3 may yield a lower empirical approximation error in practical scenarios, as shown in this illustrative example. In particular, we note that the sampling-based approach systematically underestimated the uncertainty in clustering comparison results, by producing intervals that were narrower than those obtained by means of the ad hoc approximation algorithms. Nonetheless, both approximation methods provided consistent results, in the sense that smaller values according to one method were associated with smaller values according to the other one.

In regard to the performance of the applied clustering algorithms, we note that the proposed metrics allow a comparison of these results. For example, it could be noted that, according to all metrics, RKM reported soft clusterings that were associated with a much smaller amount of uncertainty than both PCM and ECM, which instead reported comparable results in this sense. Notably, while both RKM and FCM yielded results very similar to those obtained with the hard clustering algorithm KM, by contrast both PCM and ECM yielded slightly higher values for the upper bound values of the metrics. This observation shows that the additional amount of ambiguity and uncertainty introduced by these algorithms allowed to retrieve compatible hard clusterings closer to the ground truth class assignment than those compatible with either RKM and FCM. Nonetheless, we note that, in general, the lower bound values reported by RKM were comparable with, or better than, those reported by PCM and ECM. We can deduce from this observation that, in general, RKM may achieve results comparable to PCM and ECM in terms of accuracy, while producing, at the same time, soft clusterings having a much smaller degree of uncertainty.

5.2. Simulated data

In the second experiment, our aim is to illustrate how, even on moderately large datasets (on which the exact transport-based measure cannot be computed feasibly, as a consequence of Theorem 3.2), the approximate measures proposed in Section 4 can still be applied and used to obtain indications about the performance of different clustering algorithms, as well as to perform a comparison between their results. To this aim, we considered a synthetically-generated dataset composed of 10,000 objects and two features. Each of the objects was classified into three different classes, generated by drawing from a mixture of Gaussian distributions with a large probability of overlap, as illustrated in Figure 2.

As in the previous experiment, we applied the KM, RKM, FCM, PCM and ECM algorithms to the dataset, and we compared the results obtained by each of these algorithms with the ground-truth partition. Furthermore, we also compared the clusterings obtained by each pair of algorithms, to evaluate their results. We considered, in particular the S-RI, A-RI, S-PD and A-PD measures (see Section 5.1). As in the first experiment, for all metrics, the value of the ambiguity weight α was set to 0.5. Similarly, the sampling-based approximation measures S-RI and S-PD were computed based on the drawing of 1000 samples with replacement. In order to visualize the pairwise measure values graphically

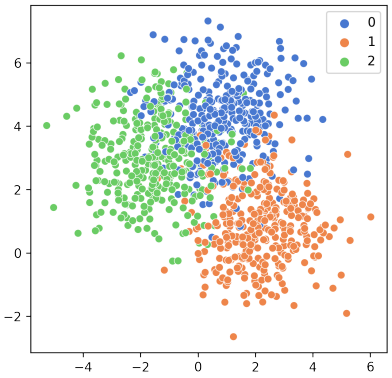

Figure 2: Graphical representation of the synthetic dataset used in the second experiment.

(for increased readability), we applied the multi-dimensional scaling algorithm for interval-valued data described in [19].

The results of the experiment are reported in Figures 3 and 4. As in the first experiment, we note that the sampling-based approximation methods severely under-estimated the uncertainty as compared to the ad hoc algorithms. Indeed, as shown in Figures 3 and 4, the size of the circle-based representations associated with the sampling-based approaches were much smaller. This result highlights how the sampling process that underlies the computation of S-RI and S-PD might lead to an underestimation of the degree of ambiguity as compared to A-RI and A-PD, likely due to the fact that a limited amount of samples does not allow us to extensively explore the space of compatible hard clusterings. Nonetheless, despite these differences, the results of the two approximation methods were aligned, for both the Rand index and the partition distance. Indeed, in all cases, the RKM and KM algorithms were more similar to the ground truth and to each other than the remaining clustering algorithms. According to all the considered metrics, the circle-based representation of RKM was always very close to that corresponding to KM, and was also the closest one to the ground truth.

This result, together with the results of the first experiment, highlights the efficiency of the RKM algorithm, which, by allowing a limited degree of ambiguity in the assignment of objects to clusters, makes it possible to retrieve the ground truth separation of objects into classes with an higher accuracy as compared to other methods, reporting results similar to, but slightly better than, the hard clustering algorithm KM. Similarly, the ECM and PCM algorithms were found to be relatively similar according to all computed metrics, and similarly associated with a larger amount of ambiguity compared to all the other clustering algorithms. In particular, according to all the considered metrics, PCM was the algorithm that generated the clusters associated with the great-

est amount of uncertainty. This result suggests that, at least for this simulated dataset, ECM might be preferable to PCM as it yields similar results in terms of accuracy (i.e., closeness to the ground truth) with, at the same time, a smaller amount of ambiguity. Nonetheless, this finding needs to be confirmed by further experiments, as it was not observed in the first experiment on the Iris dataset. Interestingly, the FCM algorithm was found to be the most dissimilar from all other clustering algorithms in the comparison. Finally, we highlight how the proposed approximation methods (both sampling-based and ad hoc) are able to provide bounds for the exact values of the comparison measures and can furthermore be used to compare two or more clustering algorithms and evaluate their quality.

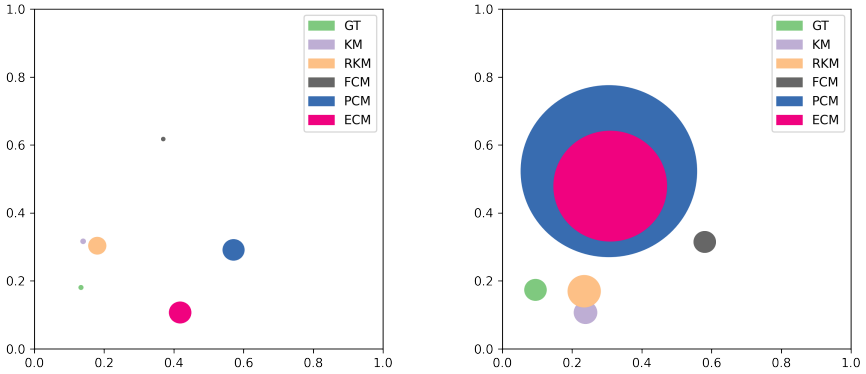

Figure 3: Multi-dimensional scaling representation of the pairwise S-RI (left) and A-RI (right) indices. The relative position of the circles corresponds to their similarity, while the size of the circles represents the width of the corresponding intervals.

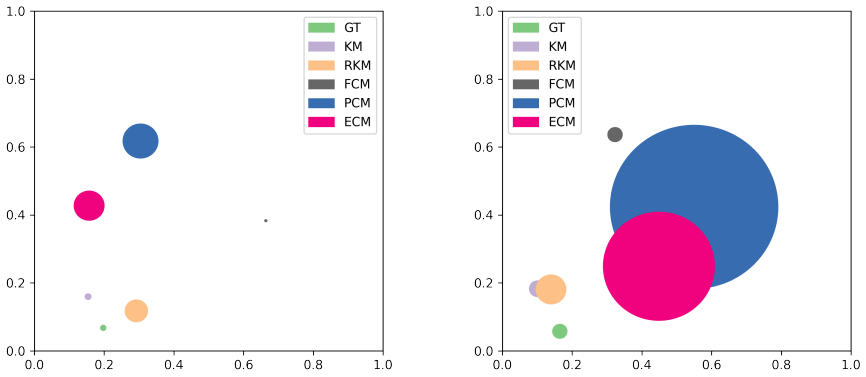

Figure 4: Multi-dimensional scaling representation of the pairwise S-PD (left) and A-PD (right) indices. The relative position of the circles corresponds to their similarity, while the size of the circles represents the width of the corresponding intervals.

6. Conclusion

In this article, we proposed a general framework for extending clustering comparison measures from hard clustering to evidential clustering (hence, as special cases, also to rough, fuzzy and possibilistic clustering). Our approach is based on the interpretation of soft clusterings as distributions over hard clustering, and uses optimal transport theory to provide a general construction method for evaluation and comparison measures of soft clusterings using what we called *transport-based measures*. We have studied the theoretical properties of this approach, in terms of metric properties and computational complexity. Furthermore, since a major limitation of the proposed approach lies in its high computational complexity, we also proposed some strategies for approximation, based on either a sampling approach or the design of alternative comparison measures (computable in polynomial time), which were shown to provide bounds for the transport-based measure.

Our contributions are summarized in Table 3, which describes, for every proposed measure, the corresponding metric properties and computational complexity. In regard to the metric properties we can observe that, compared to the existing measures previously proposed in the literature and discussed in Section 2.1, both the exact transport-based measures as well as the approximations for the Rand index and partition distance satisfy the desirable properties that were described in the introduction. Our proposal, then, extends and formalizes the previous proposals in [18, 27] by providing a theoretically grounded and general framework to enable the comparison between and across different classes of soft clustering methods. We believe that our contribution is particularly significant in this regard since, as discussed in the introduction and in Section 2.1, these properties enable the use of the proposed measures both as objective criteria

Table 3: Summary of the proposed comparison measures and their approximations and bounds

Measure	Section	Metric Properties	Computational Complexity
Transport-based Measure	3.2	$1 - d_0^E$ consistency d_1^E metric	d_0^E NP-HARD d_1^E NP-HARD
Sampling-based Approximations	4.1	-	$O(n^2 s + s^3 \log s)$
Approximation for Rand index	4.2	Rand_0 consistency Rand_1 similarity	$O(n^2)$
Approximation for partition distance	4.3	$1 - \delta_0^E$ consistency δ_1^E metric	$O(n2^k + k^3)$

to compare the results of any soft clustering algorithm with a known ground truth, thus allowing their external validation, and as objective criteria to compare the results of multiple soft clustering methods, possibly of different types (such as rough, fuzzy, possibilistic or evidential). Furthermore, since each of the proposed measures is based on a pair of functions, satisfying respectively the properties of being a consistency and a metric (or, dually, a similarity), the proposed measures also provide flexibility in modeling the trade-off between ambiguity and error. Consequently, they allow a greater degree of personalization to the users' needs than previous proposals. In regard to computational complexity, the proposed measures offer a trade-off between exactness of the results and computational feasibility. Indeed, whereas the transport-based measures allow the exact computation of the above mentioned evaluation and comparison criteria but are, in general, NP-HARD and thus infeasible to apply in large-scale problems, the ad hoc and sampling-based approximations yield reasonable estimates of the above mentioned quantities at a reduced computational cost.

Finally, to illustrate and discuss the above mentioned characteristics of the proposed measures, we have demonstrated their application through two simple experiments, in which we have described both the relationships between the transport-based measures and their approximations, as well as how these approximations can be used even in larger-scale problems.

We believe that this article could be a first step toward the development of approaches for the comparison of soft clustering algorithms. As further steps, we deem the following problems to be worthy of further investigation:

- In Section 3.2, we have shown that, in general, computing the transport-based distance is computationally hard, for both the cases of rough and evidential clustering. An important open problem would be to understand whether (and for which base distances) $d_\alpha^E(F_1, F_2)$ can be computed in polynomial time, when F_1, F_2 are fuzzy clusterings.

- In Section 4, we described algorithms for two commonly used clustering comparison measures, namely the Rand index and the partition distance. Then, we showed that these algorithms can be used to bound the value of the transport-based measure with the respective base distances. Since in specific settings other measures may be more appropriate, it would be

interesting to develop approximation methods also for other common comparison measures, such as the mutual information or the Jaccard index.

- In the experiments reported in Section 5.1, we have shown that the bounding algorithms introduced in Sections 4.2 and 4.3 have lower empirical approximation error than the sampling-based procedures introduced in Section 4.1. An interesting problem would be to find a theoretical characterization of the approximation error for the bounding algorithms.

References

[1] Ahmed, M., Mahmood, A.N., Hu, J., 2016. A survey of network anomaly detection techniques. Journal of Network and Computer Applications 60, 19–31.

[2] Anderson, D.T., Bezdek, J.C., Popescu, M., Keller, J.M., 2010. Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Transactions on Fuzzy Systems 18, 906–918.

[3] Anderson, D.T., Zare, A., Price, S., 2012. Comparing fuzzy, probabilistic, and possibilistic partitions using the earth mover's distance. IEEE Transactions on Fuzzy Systems 21, 766–775.

[4] Ashtari, P., Haredasht, F.N., Beigy, H., 2020. Supervised fuzzy partitioning. Pattern Recognition 97, 107013.

[5] Bassetti, F., Gualandi, S., Veneroni, M., 2020. On the computation of Kantorovich–Wasserstein distances between two-dimensional histograms by uncapacitated minimum cost flows. SIAM Journal on Optimization 30, 2441–2469.

[6] Bezdek, J.C., 1981. Pattern recognition with fuzzy objective function algorithms. Springer Science & Business Media.

[7] Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E., 2019. Model-based clustering and classification for data science: with applications in R. Cambridge University Press.

[8] Brouwer, R.K., 2009. Extending the Rand, adjusted Rand and Jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems 32, 213–235.

[9] Campagner, A., Ciucci, D., 2019. Orthopartitions and soft clustering: soft mutual information measures for clustering validation. Knowledge-Based Systems 180, 51–61.

[10] Campello, R.J., 2007. A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters 28, 833–841.

[11] Day, W.H., 1981. The complexity of computing metric distances between partitions. Mathematical Social Sciences 1, 269–287.

[12] Dempster, A., et al., 1967. Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics 38, 325–339.

[13] Denœux, T., 2001. Inner and outer approximation of belief structures using a hierarchical clustering approach. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 9, 437–460.

[14] Denœux, T., 2020. Calibrated model-based evidential clustering using bootstrapping. Information Sciences 528, 17–45.

[15] Denoeux, T., 2021. Nn-evclus: Neural network-based evidential clustering. Information Sciences 572, 297–330.

[16] Denœux, T., Dubois, D., Prade, H., 2020. Representations of uncertainty in ai: beyond probability and possibility, in: A Guided Tour of Artificial Intelligence Research. Springer, pp. 119–150.

[17] Denoeux, T., Kanjanatarakul, O., 2016. Evidential clustering: a review, in: International symposium on integrated uncertainty in knowledge modelling and decision making, Springer. pp. 24–35.

[18] Denœux, T., Li, S., Sriboonchitta, S., 2017. Evaluating and comparing soft partitions: An approach based on Dempster–Shafer theory. IEEE Transactions on Fuzzy Systems 26, 1231–1244.

[19] Denoeux, T., Masson, M., 2000. Multidimensional scaling of interval-valued dissimilarity data. Pattern Recognition Letters 21, 83–92.

[20] Denœux, T., Masson, M.H., 2004. EVCLUS: evidential clustering of proximity data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 34, 95–109.

[21] Depaolini, M.R., Ciucci, D., Calegari, S., Dominoni, M., 2018. External indices for rough clustering, in: International Joint Conference on Rough Sets, Springer. pp. 378–391.

[22] D'Urso, P., 2017. Informational paradigm, management of uncertainty and theoretical formalisms in the clustering framework: A review. Information Sciences 400—401, 30–62.

[23] Fowlkes, E.B., Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. Journal of the American Statistical Association 78, 553–569.

[24] Frigui, H., Hwang, C., Rhee, F.C.H., 2007. Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition 40, 3053–3068.

[25] Gagolewski, M., Bartoszuk, M., Cena, A., 2021. Are cluster validity measures (in) valid? Information Sciences 581, 620–636.

[26] Harenberg, S., Bello, G., Gjeltema, L., Ranshous, S., Harlalka, J., Seay, R., Padmanabhan, K., Samatova, N., 2014. Community detection in large-scale networks: a survey and empirical evaluation. Wiley Interdisciplinary Reviews: Computational Statistics 6, 426–439.

[27] Hullermeier, E., Rifqi, M., Henzgen, S., Senge, R., 2011. Comparing fuzzy partitions: A generalization of the Rand index and related measures. IEEE Transactions on Fuzzy Systems 20, 546–556.

[28] Jousselme, A.L., Maupin, P., 2012. Distances in evidence theory: Comprehensive survey and generalizations. International Journal of Approximate Reasoning 53, 118–145.

[29] Kantorovich, L.V., 1960. Mathematical methods of organizing and planning production. Management science 6, 366–422.

[30] Ko, K.I., Lin, C.L., 1995. On the complexity of min-max optimization problems and their approximation, in: Minimax and Applications. Springer, pp. 219–239.

[31] Krishnapuram, R., Keller, J.M., 1993. A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems 1, 98–110.

[32] Lei, Y., Bezdek, J.C., Romano, S., Vinh, N.X., Chan, J., Bailey, J., 2017. Ground truth bias in external cluster validity indices. Pattern Recognition 65, 58–70.

[33] Lingras, P., West, C., 2004. Interval set clustering of web users with rough k-means. Journal of Intelligent Information Systems 23, 5–16.

[34] Lipor, J., Balzano, L., 2020. Clustering quality metrics for subspace clustering. Pattern Recognition 104, 107328.

[35] Liu, X., Song, W., Wong, B.Y., Zhang, T., Yu, S., Lin, G.N., Ding, X., 2019. A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biology 20, 1–18.

[36] Masson, M.H., Denoeux, T., 2008. ECM: an evidential version of the fuzzy c-means algorithm. Pattern Recognition 41, 1384–1397.

[37] Meilă, M., 2007. Comparing clusterings—an information based distance. Journal of Multivariate Analysis 98, 873–895.

[38] Peters, G., 2014. Rough clustering utilizing the principle of indifference. Information Sciences 277, 358–374.

[39] Peters, G., Crespo, F., Lingras, P., Weber, R., 2013. Soft clustering: Fuzzy and rough approaches and their extensions and derivatives. International Journal of Approximate Reasoning 54, 307–322.

[40] Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66, 846–850.

[41] Ruspini, E.H., Bezdek, J.C., Keller, J.M., 2019. Fuzzy clustering: A historical perspective. IEEE Computational Intelligence Magazine 14, 45–55.

[42] Schütze, H., Manning, C.D., Raghavan, P., 2008. Introduction to information retrieval. volume 39. Cambridge University Press Cambridge.

[43] Shafer, G., 1976. A mathematical theory of evidence. Princeton University Press.

[44] Steinley, D., Brusco, M.J., 2018. A note on the expected value of the rand index. British Journal of Mathematical and Statistical Psychology 71, 287–299.

[45] Sutherland, W.A., 2009. Introduction to metric and topological spaces. Oxford University Press.

[46] Villani, C., 2021. Topics in optimal transportation. volume 58. American Mathematical Society.

[47] Vinh, N.X., Epps, J., Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. The Journal of Machine Learning Research 11, 2837–2854.

[48] Xia, Q., 2009. The geodesic problem in quasimetric spaces. Journal of Geometric Analysis 19, 452–479.

[49] Xiong, H., Li, Z., 2018. Clustering validation measures, in: Data Clustering. Chapman and Hall/CRC, pp. 571–606.

[50] Zhou, D., Li, J., Zha, H., 2005. A new Mallows distance based metric for comparing clusterings, in: Proceedings of the 22nd International Conference on Machine Learning, pp. 1028–1035.

A. Background on Belief Functions

In this section, we recall the basic notions on belief function theory [12, 16, 43]. Let X be a finite set and 2^X the corresponding power set. A *mass function* is a function $m : 2^X \mapsto [0, 1]$ s.t. $\sum_{A \in 2^X} m(A) = 1$. If $m(\emptyset) \neq 0$, then m is *unnormalized*. Given a mass function m, we define the *belief* and *plausibility* functions as:

$$Bel(A) = \sum_{B:\emptyset \neq B \subseteq A} m(B); \quad Pl(A) = \sum_{B:B \cap A \neq \emptyset} m(B). \tag{31}$$

It is easy to observe that Bel and Pl are dual of each other, that is $Bel(A) = 1 - m(\emptyset) - Pl(A^c)$ and $Pl(A) = 1 - m(\emptyset) - Bel(A^c)$. Given two mass functions, m_1 and m_2, we can define their combination as:

$$m_1 \oplus m_2(A) = \frac{1}{1 - \mathcal{K}(m_1, m_2)} \sum_{B,C:B \cap C = A} m_1(B) \cdot m_2(C), \tag{32}$$

where $\mathcal{K}(m_1, m_2) = \sum_{A,B:A \cap B = \emptyset} m_1(A) \cdot m_2(B)$ is the *conflict* between m_1 and m_2. If $\mathcal{K}(m_1, m_2) = 1$, then $m_1 \oplus m_2$ is undefined.

We define the *focal sets* of m as $\mathcal{F}(m) = \{A \in 2^X : m(A) > 0\}$. If $|\mathcal{F}(m)| = 1$, then m is said to be *logical*. If the focal sets are all singletons, then m is said to be *Bayesian*: in this case, m is a probability distribution and, $\forall A \subseteq X$, it holds that $Bel(A) = Pl(A)$. If, on the other hand, the focal sets are nested (i.e., $\forall A, B \in \mathcal{F}_m$, either $A \subseteq B$ or $B \subseteq A$) then m is said to be *consonant*, and it can be shown that Bel is a *necessity measure* and Pl is a *possibility measure* [16].

Let $f : X \mapsto \mathbb{R}$ be a function. Then, we can extend the notion of expected value to the setting of belief function theory by defining the lower and upper expected value as follows:

$$\underline{E}(f) = \sum_{A \subseteq X} m(A) \cdot \min_{x \in A} f(x) \tag{33}$$

$$\overline{E}(f) = \sum_{A \subseteq X} m(A) \cdot \max_{x \in A} f(x) \tag{34}$$

Another useful notion regards the definition of distance functions between belief functions. In particular, we recall the definitions of the belief distance [13] and the Jousselme distance [28], while we refer the reader to [28] for a more comprehensive review on the topic. Let m_1, m_2 be two mass functions, both defined on 2^X, then:

$$d_B(m_1, m_2) = \frac{1}{2} \sum_{A \subseteq X} |Bel_1(A) - Bel_2(A)|, \tag{35}$$

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2}||m_1||^2 + ||m_2||^2 - 2\langle m_1, m_2 \rangle}, \tag{36}$$

where $\langle m_1, m_2 \rangle = \sum_{A \subseteq X} \sum_{B \subseteq X} m_1(A) m_2(A) \frac{|A \cap B|}{|A \cup B|}$ and $||m||^2 = \langle m, m \rangle$.

Finally, we recall the notions of *extension* and *restriction* of a mass function. Let X, Y be two sets. We say that X is a *refinement* of Y (equivalently, Y is a *coarsening* of X) if exists a function $\rho : 2^Y \mapsto 2^X$ such that: 1) $\{\rho(\{y\}) : y \in Y\}$ is a partition of X; 2) $\forall A \subseteq Y$, $\rho(A) = \bigcup_{y \in A} \rho(\{y\})$. Given a mass function m defined on Y, the *vacuous extension* of m to X is defined by:

$$\forall A \subseteq Y, m^{Y \uparrow X}(\rho(A)) = m(A). \tag{37}$$

On the other hand, given a mass function m defined on X, the *restriction* of m to Y is defined by:

$$m^{X \downarrow Y}(A) = \sum_{B \subseteq X : \overline{\rho}^{-1}(B) = A} m(B), \tag{38}$$

where $\overline{\rho}^{-1}(B) = \{y \in Y : \rho(\{y\}) \cap B \neq \emptyset\}$ is the *outer reduction* of B.