Comparing Fuzzy, Probabilistic and Possibilistic Partitions Using the Earth Mover's Distance

Derek T. Anderson, *Member, IEEE*, Alina Zare, *Member, IEEE*, Stanton Price, *Student Member, IEEE*

***Abstract*—A number of noteworthy techniques have been put forth recently in different research fields for comparing clusterings. Herein, we introduce a new method for comparing soft (fuzzy, probabilistic and possibilistic) partitions based on the *earth mover's distance* (EMD) and the *ordered weighted average* (OWA). The proposed method is a metric, depending on the ground distance, for all but possibilistic partitions. It is extremely flexible due to its EMD formulation, OWA aggregation and abstract concept of ground distance. In theory, our method is agnostic to the *type* (uncertainty) of soft partition, clustering algorithm, distance measure used in the clustering algorithm(s) and it is applicable to the clustering of both object and relational data. Validation is performed theoretically, experimentally and also in terms of computational complexity. Emphasis is placed on the set of possibilistic partitions, specifically noise and co-incident clusters, important cases that have received little-to-no attention to date in the comparing clusterings literature. Improvements are reported in terms of metric properties and computational complexity over existing extended concordance / discordance (e.g., soft Rand and Jaccard) approaches and improved design and robustness in comparison to existing transportation problem based approaches.**

***Index Terms*—earth movers distance, comparing partitions, cluster validity, ordered weighted average, possibilistic clustering.**

I. Introduction

LET $O = \{o_1, ..., o_n\}$ denote n objects (e.g., image pixels). When each object in O is represented by a (column) vector x in $\Re^p$, the set $X = \{x_1, ..., x_n\} \subset \Re^p$ is called an *object-data representation* of O (i.e., feature vectors, pattern vectors, etc.). The kth component of the ith feature vector (x_{ki}) is the kth feature measurement or attribute (e.g., pixel intensity, fractal dimension, etc.) of the ith object. Alternatively, when each pair of objects in O is represented by a relationship between them, we have *relational data*. Let $R = [r_{ij}]$ be the matrix of relational values on $O \times O$, r_{ij} being the relation between o_i and o_j. The most common case of relational data is dissimilarity data, say $D = [d_{ij}]$, where d_{ij} is the pair-wise dissimilarity (usually a distance) $d(o_i, o_j)$. D can also be a matrix of similarities or a relation specified by a person observing a process involving pairs of objects (e.g., the Netflix database, where r_{ij} might correspond to the rating of movie i by reviewer j).

D. T. Anderson is with the Electrical and Computer Engineering (ECE) Department, Mississippi State University (MSU), Mississippi State, MS 39762 USA email: anderson@ece.msstate.edu

A. Zare is with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, 65211 USA e-mail: zarea@missouri.edu

S. Price is with the ECE Department, MSU, Mississippi State, MS 39762 USA e-mail: sp438@msstate.edu

Manuscript received May 4th, 2012, revised September 11th, 2012, accepted November 13th, 2012.

Clustering of unlabeled data is the assignment of *labels* to the objects in O. In general, there are four class label types: *crisp, fuzzy, probabilistic,* and *possibilistic* [1]. Let integer c be the number of classes, with $1 < c < n$, and define three sets of *label vectors* in $\Re^c$ as follows:

$$N_{pc} = \{p \in \Re^c : p_i \in [0,1] \quad \forall i, p_i > 0 \quad \exists i\}, \tag{1}$$

$$N_{fc} = \{p \in N_{pc} : \sum_{i=1}^{c} p_i = 1\}, \tag{2}$$

$$N_{hc} = \{p \in N_{fc} : p_i \in \{0,1\} \quad \forall i\}, \tag{3}$$

$$N_{hc} \subset N_{fc} \subset N_{pc}. \tag{4}$$

Specifically, N_{pc} is the set of possibilisitc label vectors, N_{fc} is the set of fuzzy or probabilistic label vectors, and N_{hc} is the set of hard (or crisp) label vectors. Example clustering algorithms that produce label vectors in these sets include: the *possibilistic c-means* (PCM) algorithm [2, 3] (e.g., N_{pc}), the *fuzzy c-means* (FCM) algorithm [4] (e.g., N_{fc}), the *expectation-maximization* (EM) algorithm for Gaussian-mixture decomposition [5] (e.g., N_{fc}), and the *hard c-means* (HCM) algorithm [5] (e.g., N_{hc}). With respect to relational data, a number of clustering procedures exist, e.g., *relational FCM* (RFCM) [6], *non-euclidean relational fuzzy clustering* (NERF) [7], etc. This list is not meant to be comprehensive. A number of other clustering approaches exist, e.g., normalized graph cut, kernel clustering, etc [8][9].

The label vectors in Equations 1, 2 and 3 are used as the columns of three types of *c-partitions* of O, which are sets of (*cn*) values $\{u_{ik}\}$ that can be arrayed as ($c \times n$) matrices, say $U = [u_{ik}]$. Let $\hat{U}_k$ denote the kth column of U (which is a label vector in $\Re^c$). Three important sets are

$$M_{pcn} = \{U \in \Re^{cn} : \hat{U}_k \in N_{pc} \quad \forall k, \quad 0 < \sum_{k=1}^{n} u_{ik} \quad \forall i\}, \tag{5}$$

$$M_{fcn} = \{U \in M_{pcn} : \hat{U}_k \in N_{fc} \quad \forall k\}, \tag{6}$$

$$M_{hcn} = \{U \in M_{fcn} : \hat{U}_k \in N_{hc} \quad \forall k\}. \tag{7}$$

Equations 5, 6, and 7 define, respectively, the familiar sets of possibilistic, fuzzy or probabilistic and crisp *c-partitions* of O. Since the goal of this article is to compare the results of different clusterings, let $CP = \{U_i : 1 \leq i \leq N\}$ denote N different *candidate partitions* of a fixed dataset O that may arise as a result of clustering with one algorithm under different parameter values, initializations, or, more generally, by different clustering algorithms.

Clustering is a classical subject in pattern analysis that has been addressed in a number of ways, i.e., pre-clustering (cluster tendency [10]), cluster analysis (*MANY* algorithms exist [4, 11, 12, 13, 14, 15, 5, 16, 17, 18, 19, 20, 21, 22]), and post-clustering (cluster validation [1, 23, 24], ensemble clustering [25, 26, 27, 28], etc.). The focus of this work is post-clustering. Related questions in post-clustering include: once U is found, do we believe it is the best explanation of substructure in O? Is this U useful? Are there other candidate partitions to which U is similar? Is there a better clustering that we did not find? Can clustering robustness and/or stability be improved by combining the results of multiple clusterings? etc.

A number of indices have been put forth to evaluate the performance of a single clustering. Many approaches operate solely on U, while others utilize U in conjunction with geometric information from X. Typically, these methods evaluate the intra-cluster variation (variation between objects in the same cluster, which should be low) relative to inter-cluster variation (variation between objects in different clusters, which should be high). Examples of this type include the partition coefficient, partition entropy, Fukuyama and Sugeno, and generalized Xie Beni (see [24] for a recent review).

The goal of our work is to design a new flexible index to compare pairs of soft partitions to assist with the analysis, selection and/or combination of results from one or more clusterings. For the reader unfamiliar to comparing clusterings, comparison measures have a number of uses [1], e.g., comparison of a candidate partition to a *reference partition* that purports to represent the *true cluster structure* in O (which is really only of use in designing and exploring the utility of an index as one does not encounter this situation in the clustering of unlabeled data), measurement of the sensitivity of a clustering algorithm to perturbations of the data, measurement of sensitivity with respect to missing data, measurement of the similarity between different soft partitions acquired by algorithmic means, comparison on clustering results at consecutive iterations of an algorithm, etc. A final area worth highlighting is ensemble clustering [25, 26, 27, 28]. Ensemble clustering, which is similar in concept to supervised boosting algorithms, is based on the idea that exploratory analysis of data can be benefited by the combination of multiple clusterings. A number of ensemble clustering works make use of a comparing clusterings index to help identify a diverse set of clusterings to combine (and often a strategy for how to combine them).

The remainder of this article is organized as such. First, we review related comparing clusterings works in the fields of computational intelligence and machine learning. Next, we provide a brief review of the *earth mover's distance* (EMD) [29]. We then discuss how the EMD can be used to compare soft partitions in the context of the *ordered weighted average* (OWA) [30]. To this end, two different EMD ground distances are explored, each of which posses different metric benefits and behaviors. Theoretical comparisons and computational complexity is then analyzed. Experiments are also provided and comparisons to prior work are made. Specifically, we place emphasis on the set of possibilistic partitions, a topic that is under represented in the comparing clustering literature.

II. Related Work

A. Soft Extensions of the Rand and Jaccard

In [1, 31], Anderson et al. put forth a similarity index that generalizes many of the classical indices (e.g., Rand, Jaccard, etc.) used with outputs of crisp clustering algorithms so that they are applicable for candidate partitions of any type. In that work, we compared our results and computational complexity to the work of Campello [32], Frigui [33], Hullermeier and Rifqi [34] and Brouwer [35]. However, our prior approach is only a measure, not a metric. It's advantages include flexibility and computational complexity. Another related work is Hullermeier et al.'s fuzzy Rand [34, 23] and fuzzy Jaccard [23]. Their similarity index is a pseudo-metric on $M_{fcn} \times M_{frn}$ (where c is the number of clusters in clustering 1 and r is the number of clusters in clustering 2) and a full metric on partitions that have at least one 1 in each row of the partition matrix (called *normal* partitions of O), which is a very sparse subset of $M_{fcn} \times M_{frn}$. Note, it is very unlikely that an algorithm such as the FCM or PCM will ever produce a membership of 1 for at least one data point in each cluster. A plausible case in which this will occur is when a prototype resides exactly on top of a data point. However, we note that this property would be guaranteed in medoid clustering (which is a relatively small focused subfield of clustering). A comprehensive mathematical review and comparison of these two methods can be found in [23][1]. Herein, we focus on transportation problem-based methods, as they are the closest to the proposed effort.

B. Transportation Problem-Based Approaches

In the field of machine learning, Coen et al. recently introduced a related approach for comparing clusterings *in space* [36]. Coen's goal is to combine partition information with geometric information from X (e.g., Euclidean distance between pairs of objects in two clusters). Coen defines the similarity

$$d_S(A, B; p, q, d_\Omega) = \frac{d_{OT}(A, B; p, q, d_\Omega)}{d_{NT}(A, B; p, q, d_\Omega)}, \tag{8}$$

$$d_{OT}(A, B; p, q, d_\Omega) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} f_{ij}^* d_\Omega(a_i, b_j), \tag{9}$$

$$d_{NT}(A, B; p, q, d_\Omega) = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} p_i q_j d_\Omega(a_i, b_j), \tag{10}$$

where (A, p) and (B, q) are weighted point sets, $A = \{a_1, ..., a_{|A|}\}$ is a finite subset of the metric space (Ω, d_Ω), and $p = (p_1, ..., p_{|A|})$ is a vector of non-negative weights that sum to one (similar definitions hold for B and q respectively). Specifically, Equation 9 is the *optimal transportation distance* between (A, p) and (B, q), where f^* is the optimal flow, and Equation 10 is the *naive transportation distance*. The intent of Equation 8 is to measure the *degree to which cooperation*

reduces the cost of moving the source A onto the sink B. Coen defines a distance between two clusterings as follows. The distance between clusterings C and D (sets of crisp clusters) of data sets X_1 and X_2 (which may be the same set) is

$$d(C, D) = d_S(C, D; \pi, p, d'_{OT}), \tag{11}$$

where $\pi = (|E|/|X_1| : E \in C)$ and $p = (|F|/|X_2| : F \in D)$ are weight assignments proportional to the number of data points in the clusters and $|\cdot|$ is crisp set cardinality. Coen defines the distance between two clusters (note the specific use of the terms cluster and clustering), $E \in C$ and $F \in D$, as the optimal transportation distance,

$$d'_{OT}(E, F) = d_{OT}(E, F; p, q, d_\Omega), \tag{12}$$

where $p = (1/|E|)$ and $q = (1/|F|)$ (a uniform weight assignment scheme for a single cluster). Thus, both the clustering and cluster distance is formulated as a transportation problem.

Coen's method requires one to *harden* each clustering, i.e., the (destructive) reduction of a soft to crisp partition. Specifically, for x_k and $j = \arg\max_{i \in \{1,\ldots,c\}} u_{ik}$, u_{jk} is assigned a value of 1 and u_{ik}, $\forall i \neq j$, is set to 0. After hardening, Coen's weights are assigned as the percentage of data points in each cluster. While arguably valid for crisp partitions, the hardening of a soft partition destroys uncertainty information that is important to comparing clusterings. For example, in possibilistic clustering, a *noisy* data point is expected to have a low (near zero) membership, e.g., $\sum_{i=1}^{c} u_{ik} \approx 0.00001$. Consider a noisy data point with membership 0.00001 in cluster 1 and 0.00002 in cluster 2. Also, assume a second clustering of the same data set using the same algorithm but with different initialization. Now, imagine a scenario in which the terminal clustering algorithm membership assignment changes slightly, 0.00002 and 0.00001 respectively, in clustering 2. The result is a change in cluster assignment after hardening. Hence, this tiny change of 0.00001 is treated the same as if a data point in the center of a cluster has been moved to another cluster. Fluctuations occur for both noise as well as boundary points, e.g., two clusters and the membership degrees 0.505 and 0.495. Note, noise relates to possibilistic clustering while the topic of boundary points relates to fuzzy, probabilistic and possibilistic. In addition, we argue that a data point with low membership, e.g., 0.00001, should not be able to contribute, so Coen's cardinality calculation, the same as a data point that has a high membership, e.g., 1.

Second, Coen's main idea is to include Euclidean distance information between pairs of points in different clusters in X. However, if not done properly, inclusion of geometric information in X results in methods that are not general enough to address a wide range of clustering settings. For example, consider the FCM algorithm with Euclidean (prototype) or GK (prototype and covariance matrix) distance, or even a spectral or kernel clustering algorithm. Each approach the topic of proximity in extremely different ways. Computing the Euclidean distance between data points in X may not be indicative of the true underlying distance between points, or ultimately clusters at that. Even if one extends this idea in definition to be valid for any distance, e.g., distance measured in the kernel space, the method is now reliant on object data and cannot address relational data.

A final related work in machine learning is the closest to the effort proposed here. In [37], Zhou et al. propose a comparing partitions method based on the Mallows distance. In concept, Zhou's approach is very similar in theme to the initial EMD work of Rubner [29] (the well-known *signature* form). They, like us, achieve metric properties for the cases of crisp, fuzzy and probabilistic partitions. It is also worth noting that they, like most, only study the cases of crisp and probablistic (occasionally fuzzy) clustering, not possibilistic. Zhou's approach is as follows. Let

$$s_1 = ((\tilde{U}_1, \alpha_1), ..., (\tilde{U}_c, \alpha_c)), \tag{13}$$

$$s_2 = ((\tilde{V}_1, \beta_1), ..., (\tilde{V}_r, \beta_r)), \tag{14}$$

be two *signatures*, where $\tilde{U}_i$ is the i^{th} partition matrix row (aka *cluster profile*) in clustering 1 ($1 \times n$) and $\tilde{V}_j$ is the j^{th} partition matrix row (also $1 \times n$) in clustering 2. Zhou assigns α_i to either uniform ($1/c$) or the percentage of data points in a cluster (respectively for the set of β_i values and clustering 2). The distance between two signatures is found using the Mallow's distance.

A big difference between Coen and Zhou's work is the distance calculation between two clusters. Coen takes a transportation problem approach based on the formulation of two point sets. Zhou uses only the partition matrices and a L_P-norm (specifically $P = 1$),

$$d_{L_P}(\tilde{U}_i, \tilde{V}_j) = (\sum_{k=1}^{n} |u_{ik} - v_{jk}|^p)^{\frac{1}{P}}. \tag{15}$$

Zhou's method also discards valuable uncertainty information when hardening the partition matrix and calculating their signature weights. As a result, they are sensitive in a similar respect to what we just discussed for Coen. Also, Zhou uses the L_1-norm as their ground distance. However, the L_P-norm is sensitive in part to the number of data points in a cluster. Meaning, one obtains larger distances between clusters that have more numbers of elements. While Zhou did not explicitly cross out the possibility of using other ground distances, they did not consider and explore such a path and ultimately its impact.

III. Earth Mover's Distance

The EMD has been explored in a number of works in machine learning and computer vision. The EMD aims to measure the perceptual equivalence between two potentially variable size descriptions of distributions. Herein, without loss of generality, we refer to these descriptors as histograms. However, the underlying distributions could be probabilistic, possibilistic, etc. The EMD is based on a solution to the well-known transportation problem, aka Monge-Kantorovich

problem. In [29], Rubner introduced the EMD, and the well-known *signature* form, in the context of *content based image retrieval* (CBIR). In [38], Levina and Bickel proved that the EMD is the Mallows distance for two probability distributions. However, as Levina and Bickel observed, the Mallows and EMD behave differently for the case of two histograms (or signatures) with different masses. Specifically, the EMD has the advantage that it allows for partial matching. This feature is important in the case of CBIR. As we show below, it is also important with respect to comparing possibilistic partitions.

Let h be a (one dimensional) histogram (distribution) of length L_1, $h_i \in \Re^+$ and $1 \leq i \leq L_1$, and let g be a second histogram of length L_2 and $g_i \in \Re^+$. The EMD is defined as the distance between h and g, $0 \leq EMD(h,g)$. The goal of the EMD is to find a flow $F = [f_{ij}]$, where f_{ij} is the *flow* between h_i and g_j, that minimizes the overall cost

$$WORK(h,g,F) = \sum_{i=1}^{L_1}\sum_{j=1}^{L_2} d_{ij}f_{ij}, \tag{16}$$

subject to the constraints

$$f_{ij} \geq 0 \qquad 1 \leq i \leq L_1, 1 \leq j \leq L_2, \tag{17}$$

$$\sum_{j=1}^{L_2} f_{ij} \leq h_i \qquad 1 \leq i \leq L_1, \tag{18}$$

$$\sum_{i=1}^{L_1} f_{ij} \leq g_j \qquad 1 \leq j \leq L_2, \tag{19}$$

$$\sum_{i=1}^{L_2}\sum_{j=1}^{L_1} f_{ij} = \min\left(\sum_{i=1}^{L_2} h_i, \sum_{j=1}^{L_1} g_j\right), \tag{20}$$

where $D = [d_{ij}]$ is the *ground distance* matrix. Intuitively, the EMD can be thought of as follows. Imagine that the two histograms are piles of sand or earth sitting on the *ground*. Then the *distance* between the two piles can be thought of as how far the grains of sand have to be moved to make one pile be transformed into the other. That is, the EMD is the minimal total ground distance traveled weighted by the amount of sand moved. Once the transportation problem is solved [29], and the optimal F^* is found, the EMD is calculated as follows

$$EMD(h,g) = \frac{\sum_{i=1}^{L1}\sum_{j=1}^{L2} f_{ij}^* d_{ij}}{\sum_{i=1}^{L1}\sum_{j=1}^{L2} f_{ij}^*}. \tag{21}$$

The normalization in Equation 21 is the total weight of the smaller histogram (Constraint 20). This is required when two histograms have different total weight (avoids favoring the smaller histogram). While the ground distance can be, in general, any distance, different selections result in different properties (e.g., measure or metric properties which we discuss later in this article).

IV. Comparing Soft Partitions Using the EMD

Once again, we stress that our approach is based solely on partition matrices. That is, additional clustering algorithm specifics, e.g., prototypes, covariance matrix, kernels, etc., and/or the data space, X, is not desired here because ultimately it results in an overly restrictive index. Namely, inclusion of additional information, while attractive at first, does not ultimately lead to flexible cluster comparison methods that are capable of operating across different clustering algorithms, distance measures, uncertainty types (crisp, fuzzy, probabilistic and possibilistic), and object as well as relational data.

Our contribution is the following. Data point x_i is associated with $\hat{U}_i$ $(c \times 1)$ in clustering 1 and $\hat{V}_i$ $(r \times 1)$ in clustering 2 (i.e., two label vectors). Think of each label vector as a histogram and the EMD as the way to compare these two clustering assignments. We interpret this as how much *work* does it take, with respect to a single data point, to transfer one set of membership assignments in clustering 1 into another set in clustering 2. As we show below, this way of using the EMD is more flexible, versus the work of Zhou, as it also allows us to take into account all of the uncertainty information at each step in comparing partitions. However, since our EMD formulation is defined with respect to a single data point, how does one compare two partitions, U and V? Even if one selects the same number of clusters, cluster i in clustering 1 does not necessarily correspond to cluster i in clustering 2. That is, one must solve the (extremely difficult) correspondence problem. However, even if one is able to sufficiently address the correspondence problem for the case of equal number of clusters ($c_1 = c_2$), what is the solution in the case of two clusterings with different numbers of clusters ($c_1 \neq c_2$) or the case of co-incident clusters in possibilistic clustering (a many-to-one mapping)? Instead of seeking an *outright* solution, we take a soft approach. To this end, we leverage the way in which the EMD compares histograms of different lengths and possibly different masses, i.e., two label vectors with different numbers of clusters, and we specify a ground distance (D) for soft cluster correspondence. There are *MANY* possible ways (measures) to address this task. We introduce two different approaches herein that ultimately result in different overall metric properties and behaviors. A ground distance, $D = [d_{ij}]$, between clusters is defined here as

$$d_{ij} = d(\tilde{U}_i, \tilde{V}_j), \tag{22}$$

Note, $d(\tilde{U}_i, \tilde{V}_j)$ is not a *physical distance* in the respect that it does not take into account the location of the clusters or data elements in X (for the reasons already discussed). Instead, it is a distance or measure of overlap, depending on the ground distance employed, between clusters according to their membership assignments. Note, $\tilde{U}_i$ and $\tilde{V}_j$ are the same dimensionality $(1 \times n)$, which means a large number of techniques already exist to measure dissimilarity. We investigate the strengths and weaknesses of two different choices below.

The first ground distance explored is the familiar L_P-norm, Equation 15. It's main advantage is that it results in a comparing partitions metric, with respect to our procedure defined below, for crisp, fuzzy and probabilistic partitions. However, as already discussed, the L_P-norm is sensitive in part to the number of elements in each cluster. Also, we note that the resulting comparing partitions value may be difficult to interpret. However, we assert this does not ultimately make

it any less useful, e.g., computationally, but manual analysis may be more difficult.

The second ground distance explored is based on Hullermeier's extended Jaccard measure [23]. While Hullermeier used it to compare two fuzzy partitions, we investigate its inclusion as a ground distance with respect to two cluster profiles, $\tilde{U}_i$ and $\tilde{V}_j$. Let

$$E_{\tilde{U}_i}(x_k, x_z) = 1 - \frac{1}{2}\left|u_{ik} - u_{iz}\right|, \tag{23}$$

be an equivalence relation on X [23] with respect to cluster $\tilde{U}_i$ (similar definition $E_{\tilde{V}_j}(x_k, x_z)$ with respect to cluster $\tilde{V}_j$). Next, with respect to clusters $\tilde{U}_i$ and $\tilde{V}_j$, let

$$a = \sum_{k=1}^{n-1}\sum_{z=k+1}^{n}(1 - \left|E_{\tilde{U}_i}(x_k, x_z) - E_{\tilde{V}_j}(x_k, x_z)\right|) E_{\tilde{U}_i}(x_k, x_z)E_{\tilde{V}_j}(x_k, x_z), \tag{24}$$

$$b = \sum_{k=1}^{n-1}\sum_{z=k+1}^{n} max(E_{\tilde{U}_i}(x_k, x_z) - E_{\tilde{V}_j}(x_k, x_z), 0), \tag{25}$$

$$c = \sum_{k=1}^{n-1}\sum_{z=k+1}^{n} max(E_{\tilde{V}_j}(x_k, x_z) - E_{\tilde{U}_i}(x_k, x_z), 0), \tag{26}$$

be measures of concordance (a) and discordance (b and c). The familiar (similarity) Jaccard measure (also known as the Tanimoto coefficient) is

$$\frac{a}{a+b+c}. \tag{27}$$

Herein, we measure cluster distance according to

$$d_{Jaccard}(\tilde{U}_i, \tilde{V}_j) = 1 - \frac{a}{a+b+c}. \tag{28}$$

Thus, for both the L_P-norm and the Jaccard, the more similar two clusters, the less *work* it takes to move *earth* in the EMD. The ground distance is therefore either $D = [d_{ij}] = d_{Jaccard}(\tilde{U}_i, \tilde{V}_j)$ or $D = [d_{ij}] = d_{L_P}(\tilde{U}_i, \tilde{V}_j)$. Note, if D is a metric and h and g have equal *mass*, the EMD is a metric [29], thus

- (symmetry) $EMD(h, g) = EMD(g, h)$,
- (separation) $EMD(h, g) = 0$ iff $h = g$,
- (reflexivity) $EMD(h, h) = 0$.
- (triangular inequality) $EMD(h, g) \leq EMD(h, z) + EMD(z, g)$,

The EMD, with respect to x_k, clusterings U and V and ground distance D is

$$EMD(h = \hat{U}_k, g = \hat{V}_k). \tag{29}$$

As already noted, Equation 29 is with respect to a single data point (two label vectors), not two partition matrices (the arraying of multiple label vectors). We aggregate the EMD scores in order to acquire a distance between partitions. While our approach is not in the *theme* of prior signature form work, we argue that our approach is mathematically and conceptually valid, provides a greater degree of flexibility and it includes all of the uncertainty at each calculation. The OWA-based [30] distance between two partitions (clusterings) is defined as

$$d_{CPEMD}(U, V) = \sum_{k=1}^{n} w_k EMD(\hat{U}_{(k)}, \hat{V}_{(k)}), \tag{30}$$

where w_k is the k^{th} OWA weight, $\sum_{k=1}^{n} w_k = 1$, and X has been sorted so that

$$EMD(\hat{U}_{(1)}, \hat{V}_{(1)}) > ... > EMD(\hat{U}_{(n)}, \hat{V}_{(n)}). \tag{31}$$

However, additional information is available in our setting, particularly the partition matrix. We use an OWA re-ordering based on the *significance* (membership) of each data point,

$$max(\bar{U}_{(1)}, \bar{V}_{(1)}) > ... > max(\bar{U}_{(n)}, \bar{V}_{(n)}), \tag{32}$$

$$\bar{U}_k = max_{i=1,...,c} u_{ik}, \tag{33}$$

$$\bar{V}_k = max_{j=1,...,r} v_{jk}. \tag{34}$$

This OWA formulation is extremely flexible as it can model a wide range of different concepts. The OWA can produce many familiar order statistics, e.g., mean ($\vec{w} = [1/n, ..., 1/n]^T$), median ($\vec{w} = [0, ..., 1, ..., 0]^T$), max ($\vec{w} = [1, 0, ..., 0]^T$), min ($\vec{w} = [0, ..., 1]^T$), etc. The OWA has been used a number of times to represent linguistic concepts as well, such as *soft max*, e.g., $\vec{w} = [0.5, 0.3, 0.1, 0.1, 0, ..., 0]^T$. The point is, an OWA offers a great deal of freedom in calculating the final distance. For example, a *soft max* OWA weighting scheme would place emphasis on data points with strong support in one or both of the clusterings. This allows one to call two clusterings similar if *most* points that cluster well are similar. One can base this calculation on percentiles, an exponential decay function, etc. While many attempt to explicitly identify and remove noise from data or lessen their effect within a clustering algorithm, an OWA weight selection such as *soft max* is a graceful way of addressing robustness at the comparing partitions level.

Note, $\hat{U}_k$ and $\hat{V}_k$ have the same *mass* for the sets of crisp, fuzzy and probabilistic. That is, the membership of an element to all clusters is required, by definition, to sum to 1 (Equation 2). In the case of possibilistic partitions, this is not guaranteed. Thus, for crisp, fuzzy and probabilistic, d_{CPEMD} is a metric between partitions, sum of individual $EMD(h = \hat{U}_k, g = \hat{V}_k)$ metric calculations, for the case of L_P-norm and also for the Jaccard as long as the stated *normal* condition is met. For the case of possibilistic, d_{CPEMD} is not guaranteed to (and is expected to most often not be) be metric regardless of the ground distance, it depends on the terminal partition matrix.

V. Computational Complexity

This section is a computational complexity analysis of the proposed work in comparison to [1, 37, 36, 23]. The optimal transportation distance between two *point sets* of cardinality at most L can be computed in worst case time $O(L^3 log(L))$ time (Orlin's algorithm) [39]. Rubner [29] used the transportation simplex, which has complexity $O(L^3)$-$O(L^4)$. In [39], Sameer and Jacobs showed an approximate technique, based on wavelets, that runs in linear time, $O(L)$. Our following analysis is based on Sameer's approximation. Note, in our procedure, L is with respect to the number of clusters. In the work of Coen, L is with respect to both the number of clusters and also the (varying) number of data points in a cluster (so on the order of n). In the work of Zhou, L is with respect to the number of clusters.

Let $|X| = n$, c is the number of clusters in clustering 1, U, and r is the number of clusters in clustering 2, V. Below, we assume $r \leq c$. However, if $c < r$, U and V can simply be swapped without loss of generality. Our approach involves n EMD calculations. For an L_P-norm, e.g., $P = 2$ (Euclidean), on a vector of length n, the cost is $3n$. With respect to Hullermeier's extended Jaccard, one must first calculate the pair-wise $E_{\tilde{U}_i}(x_k, x_z)$ and $E_{\tilde{V}_j}(x_k, x_z)$ terms, a cost, per cluster pair, of $4n(n-1)$. The concordant and discordant term cost, i.e., a, b, c, per cluster pair is $9\frac{n(n-1)}{2}$. The ground distance is a $(r \times c)$ matrix, requiring $rc - \frac{r(r-1)}{2}$ distance calculations. For the L_P-norm, the ground distance matrix cost is $3n(rc - \frac{r(r-1)}{2})$ and the cost of the soft Jaccard is $(9\frac{n(n-1)}{2} + 4n(n-1))(rc - \frac{r(r-1)}{2})$.

Our method uses an OWA, which requires sorting. A common sorting method is the mergesort algorithm, which has a worst cost case of $O(nlog(n))$. The OWA also requires n multiplications and $(n-1)$ additions. In total, with respect to the linear time Sameer wavelet EMD approach, the L_P-norm has a cost of $O(3n(rc - \frac{r(r-1)}{2}) + nL + nlog(n) + 2n - 1)$ and a cost of $O((9\frac{n(n-1)}{2} + 4n(n-1))(rc - \frac{r(r-1)}{2}) + nL + nlog(n) + 2n - 1)$ for the soft Jaccard. Note, the n EMD calculations can be calculated in parallel as they are independent of each other.

In [36], Coen et al. put forth a cluster distance and clustering distance, each of which are based on solutions for the transportation problem. It is not possible to put forth a single computational complexity for their procedure since their distance between clusters depends on how many data points are in each cluster. In general, their cost is on the order $O(3d\frac{n(n-1)}{2} + (rc - \frac{r(r-1)}{2})L_n + L_c)$, where d is the feature space dimensionality, L_n (dependent on n) is the linear time complexity associated with their cluster cost and L_c (dependent on c) is the linear time complexity with respect to the clustering distance. Specifically, the term L_n is expected to be the largest and dominate their calculation. In [37], Zhou et al. use an L_P-norm (specifically $P = 1$) and the Mallow's distance. The cost of the L_1-norm is $3n - 1$ and their overall complexity is on the order of $O((rc - \frac{r(r-1)}{2})(3n-1) + L_c)$. In addition, the complexity of our prior soft Rand approach is $O(2rcn + 3rc)$ [1] and the complexity of Hullermeier's fuzzy Rand [34] is $O(\frac{3rn^2 + 3cn^2 - 3rn - 3cn}{2} + 5)$ [1]. In closing, our computational complexity, with respect to the L_P-norm, is similar to that of Coen, Zhou and our prior soft Rand. However, the use of the extended Jaccard moves our computational complexity above that of Hullermier's fuzzy Rand.

VI. Experimental Validation

In this section, we switch from mathematical analysis to experimental validation. As noted by Hullermeier [23], experimental comparison of comparing clusterings methods is not trivial. We begin by studying a simple, but important, fuzzy case put forth by Hullermeier in [34]. Second, we investigate how each method performs with respect to a sequence of fuzzy partitions with a natural linear order. Last, we explore the impact of co-incident clusters and noise in possibilistic clustering on a comparing clusterings approach.

A. Case 1: (Simple) fuzzy case put forth by Hullermeier [34]

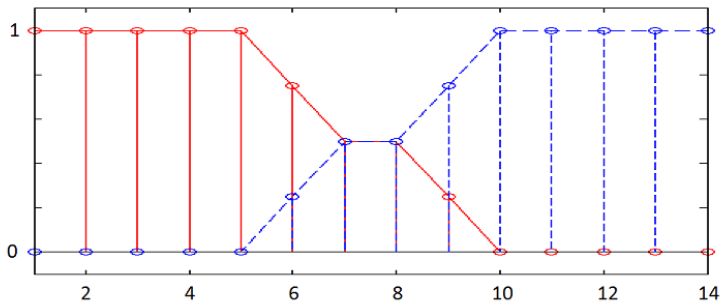
In this first example, we consider a specific (simple) numeric fuzzy case, Figure 1, put forth by Hullermeier [34]. Specifically, let $U1$ be a 2×14 partition with a soft transition between its two clusters. The question is, what is the distance (or similarity) of this partition to itself? In Campello's [32], Frigui's [33], as well as our previous soft Rand approach [1], a similarity value of $s(U1, U1) < 1$ is produced. However, Hullermeier's fuzzy Rand assigns a value of 1 (note, $U1$ is a *normal* partition). With respect to the method proposed here, we obtain a (correct) value of $d_{CPEMD}(U1, U1) = 0$ for both the L_P-norm and soft Jaccard ground distance.

B. Case 2: Sequence of fuzzy partitions [23]

In this second experiment, we perform a similar analysis to another experiment put forth by Hullermeier in [23]. We generate a *sequence* of fuzzy partitions, $(P_1, ..., P_M)$, with a so-called *natural linear order*. The assumption is that the closer index i is to index j, $1 \leq i, j \leq M$, the more similar the partitions should be. As in [23], we generate a sequence of partitions using the FCM algorithm (Matlab's implementation) for the cases of $\{c = 2, c = 3, ..., c = 8\}$ (7 clusterings). We synthetically generated 4 Gaussian clouds with 100 samples per cloud. The means of the Gaussians are $[16, 80]^T$, $[10, 20]^T$, $[70, 20]^T$, and $[64, 70]^T$, and the standard deviation is 10.

Figure 2(a)-(d) is four scatter plots of this data set and the location of the FCM algorithm terminal prototypes for $c = \{5, 6, 7, 8\}$. Figure 3 is an illustration of this experiments results. Graphical results are provided, versus a numeric table, because it is easier as such to visually observe the trends. In Figure 3, each matrix has been normalized between its minimum and maximum for visualization. It is the trends of the relative distances that we analyze herein. That is, we are interested in how these distances respond relative to each other and we want to make sure that $d(P_i, P_j) \leq d(P_i, P_k)$ for $\forall |i - j| \leq |i - k|$. We used an L_P-norm and a *soft max* OWA weighting scheme. Specifically, the weights are generated by an exponential decay function, $w_i = y_i/\tau$, where $y_i = e^{(-i/(\frac{1}{4}n))}$ and $\tau = \sum_i^n y_i$.

In [23], more significant differences were observed between the Hullermeier's fuzzy Rand, our prior soft Rand, Campello's



$U1$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}
c_1	1	1	1	1	1	0.75	0.5	0.5	0.25	0	0	0	0	0
c_2	0	0	0	0	0	0.25	0.5	0.5	0.75	1	1	1	1	1

Fig. 1. (Simple) numeric fuzzy 2×14 partition with a soft transition put forth by Hullermeier in [34]. Note, the x-axis is the data point index, i is the index of x_i, and y-axis is the membership degree.

index and Runkler's index. However, Figure 3 shows that the three methods explored herein behave in very similar respects for a sequence of fuzzy partitions. Each index was designed for the family of fuzzy partitions. Note, Zhou's method was designed for probabilistic clustering, but the sets of fuzzy and probabilistic share the same label vector and partition matrix definitions (Equations 2 and 6). Overall, no method performs extraordinarily poor or exhibit behavior that warrants concern. The big difference is the numeric scale at which these indices operate (which is range compressed in Figure 3).

C. Case 3: Impact of noise and co-incident clusters in possibilistic partitions

In this sub-section, we focus on possibilistic partitions. We investigate two phenomenon, noisy data and co-incident clusters. Noise is a common and difficult problem that plagues clustering algorithms and potentially comparing clusterings methods. As discussed above, noise affects Coen and Zhou's method's as a result of hardening and their crisp (harsh) weight assignments. Ideally, noisy data points are assigned little-to-no membership. In crisp, fuzzy and probabilistic partitions, noise is extremely problematic due to the constraint $\sum_{i=1}^{c} u_{ik} = 1$. The family of possibilistic partitions, and subsequently algorithms, provide a way to represent and identify such caveats in data.

Before proceeding, one must define what is meant by co-incident clusters. For example, consider the FCM algorithm. If the data actually contains c clusters and a user specifies $c+\alpha$ ($\alpha > 1$), the FCM will do as told and partition the data accordingly. However, while two or more prototypes might *reside in* the same cluster, they do not *represent* the same cluster. That is, algorithms such as the FCM split clusters. Our working definition of co-incident clusters is based on the partition matrix. If two clusters are co-incident, we require that they have near identical row vectors.

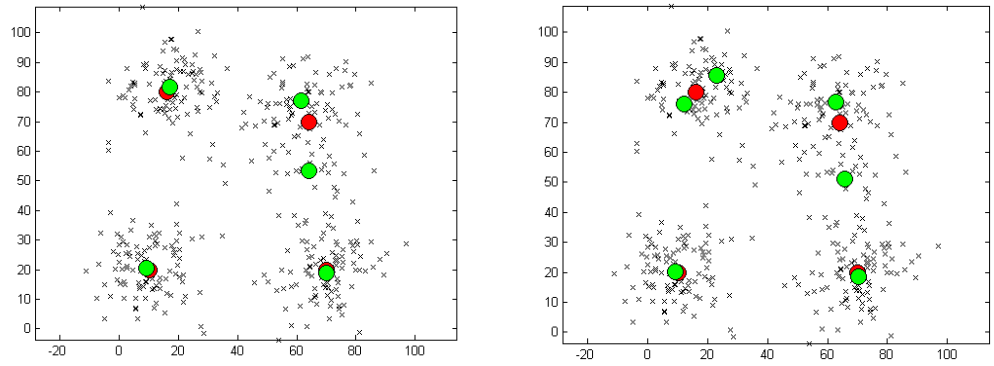
The experiments in this section are based on partitions $U2$-$U7$ (Figure 4). Numeric cases are investigated versus larger synthetic data sets in which it is hard to isolate the problematic numeric nuances. The problems discussed in this section only compound for larger data set sizes. Figure 5 is an illustration of the resulting dissimilarity matrices. In the case of Hullermeier (similarity index s), we use the common transformation $1 - s$ to produce a dissimilarity value to simplify the comparison process. Also, as before, we scale the resulting matrices between minimum and maximum.

Case 3(a): Self-similarity, noise and matrix hardening

First, as expected, the diagonals are all 0's. Second, consider partitions $U2$ and $U3$. The only difference between these partitions is an insignificant value of 0.00001 with respect to x_1. However, while these membership values are tiny, as is their difference, the cluster assignment changes after hardening. As a result, Zhou's weight assignment changes from $[w_{c_1} = 4/6, w_{c_2} = 3/6]^T$ to $[w_{c_1} = 3/6, w_{c_2} = 4/6]^T$. Also, x_1 with $u2_{1,1} = 0.00002$ contributes to the same degree as x_2, which has a much larger maximum membership value of $u2_{1,2} = 1$. Figure 5(c) shows that Zhou's method is the most sensitive of the set explored. Note, with respect to our *soft max* OWA weighting, x_1 has the lowest membership. Data point x_1 is therefore associated with a (tiny) OWA weight of $w_6 = 0.0144$. This makes it's contribution little-to-nothing to the final clustering distance. While this problem compounds for Zhou with respect to more data points, it is mitigated in part using our OWA strategy.

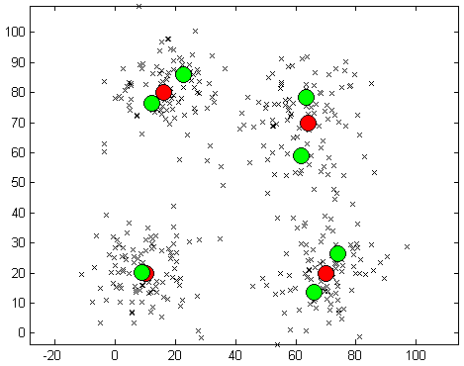
Case 3(b): Co-incident clusters

Consider partitions $U4$ (c_1 and c_3 are co-incident in actual cluster 1) and $U5$. Note, for sake of argument, we consider $U5$ to be the reference partition since it contains no co-incident clusters and it has just one (algorithmically) identified cluster in each actual cluster. Our approach, for both ground distances, reports no difference between these clusterings. Zhou's method agrees with our findings. While the matrices $U4$ and $U5$ are different, even after re-permutation, one can argue that these are very similar, if not the same, clusterings. However, Hullermeier's fuzzy Rand reports that these are dissimilar

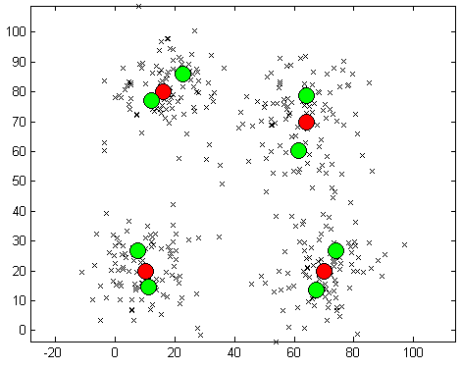


(a) Prototype locations for FCM with $c = 5$

(b) Prototype locations for FCM with $c = 6$



(c) Prototype locations for FCM with $c = 7$



(d) Prototype locations for FCM with $c = 8$

Fig. 2. Scatter plot of our case 2 data set (four Gaussian's) clustered at different values of c using the FCM algorithm. Red circles are the Gaussian means and green circles are the FCM prototype locations.

clusterings.

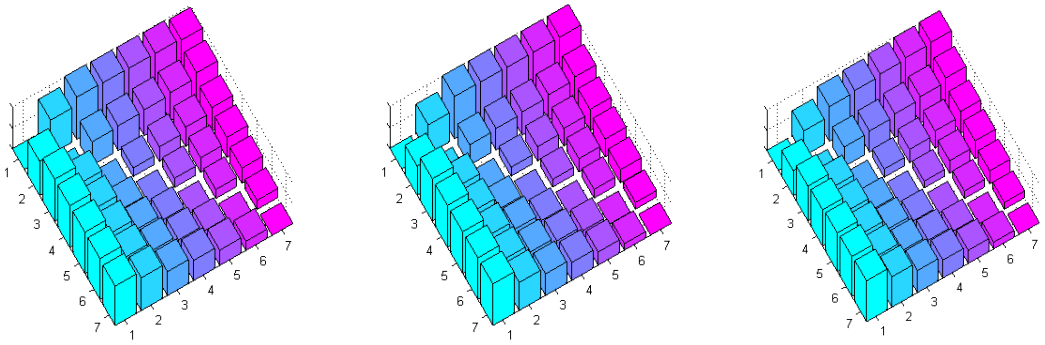
Second, consider partition $U7$, in which $\{c_1, c_3, c_5\}$ are co-incident, as are $\{c_2, c_4, c_6\}$. Our method, again with respect to both ground distances, calls clustering $U7$ the same as clusterings $U4$ and $U5$. Zhou agrees with our decision once again while Hullermeier's index calls these clusterings different.

Third, consider $U6$. It is the most interesting case because each index assigns a different value and it is the most revealing in terms of how these methods really differ. As expected, Hullermeier's method considers each clustering pair dissimilar. Zhou and our index perform similar in many of the cases, however we disagree with respect to clusterings $U6$ and $U7$. This case emphasizes the difference between our approach and Zhou's in terms of EMD formulation. While our approach provides a flexible way to aggregate the comparison of different clustering sub-structure and it is better equipped to address noise, it is ultimately not as well-suited as Zhou's method to address a wide range of co-incident clusters. We are robust to a number of co-incident cluster scenarios, but it depends on the number and type of co-incident clusters identified. This is a result of the fact that we compare per-data point (column vectors), while Zhou compares per-cluster (per row vectors). In the case of clusterings $U6$ and $U7$, we find a value of 0 while Zhou finds a value greater than 0. We assert that most people would (should) say that $U6$ and $U7$ are different because $U6$ identified the first cluster but not the second (again, assuming $U5$ is the reference partition). Zhou measures a distance greater than 0 because their comparison is done per-cluster, while our per-data point comparison is able to move all the mass (membership). However, it is important to remark that a number of works exist to lessen the effect of co-incident clusters in mode seeking algorithms, such as the PCM algorithm, e.g., Timm et al. [40], Caceres et al. [41], Yang et al. [42].

Case 3(c): Extended Jaccard as ground distance

Last, we explore the impact of using the extended Jaccard as the ground distance. Note, the sub-matrix for $U4$-$U7$ is effectively the same distance. While the extended Jaccard, a measure of overlap, may perform arguably well for the sets of crisp, probabilistic and fuzzy, for the case of possibilistic it is very optimistic in terms of what types of clusterings it considers similar. However, we do note that the extended Jaccard is not as sensitive, in terms of the number of data points in each cluster, as the L_P-norm. Also, as discussed in [23], the extended Jaccard is a metric for the case of *normal* partitions.

In closing, we find it important to remark on the following. We elected, for the reasons discussed above, to not make use of information from X. However, we understand and acknowledge Coen's desire to include such information. If it makes sense to include such information for a given application, one can use it in our framework. For example, one can use Coen's



(a) Anderson et al. (this article) (b) Hullermeier et al. fuzzy Rand (reported as $1-s$) (c) Zhou et al. approach

Fig. 3. Graphical illustration of the dissimilarity matrices for the comparison of a sequence of fuzzy partitions with a so-called *natural linear order*. Each matrix has been normalized between its minimum and maximum to assist visual analysis. Index 1 is partition $U2$, 2 is $U3$, etc. Hullermeier's index is converted into a distance using $1-s$.

transportation problem-based approach for cluster distance as our ground distance. This is a further demonstration of the flexibility of our formulation.

VII. Conclusion and Future Work

In summary, we put forth a new flexible framework for comparing soft partitions using the EMD and OWA. The proposed method is a metric, depending on the ground distance, for all but possibilistic partitions. It is extremely flexible due to its EMD formulation, OWA aggregation and abstract concept of ground distance. In theory, our method is agnostic to the *type* (uncertainty) of soft partition, clustering algorithm, distance measure used in the clustering algorithm(s) and it is applicable to the clustering of both object and relational data. Validation is performed theoretically, experimentally, and also in terms of computational complexity. Emphasis is placed on the set of possibilistic partitions, specifically noise and co-incident clusters, important cases that have received little-to-no attention to date in the comparing clusterings literature. We show improvement in terms of metric properties and computational complexity over existing concordance / discordance approaches (e.g., soft Rand and Jaccard) and improved design and robustness to existing transportation problem based approaches (i.e., Coen and Zhou).

Future work includes the exploration of different ground distances. In addition, we will also explore how to remove some of the observed sensitivity to the specific number and type of co-incident clusters. Investigation into a new way to formulate a ground distance based on the combined knowledge of U and X will also be conducted.

References

[1] D. T. Anderson, J. C. Bezdek, M. Popescu, and J. M. Keller, "Comparing fuzzy, probabilistic, and possibilistic partitions," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 5, pp. 906–918, 2010.

[2] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. on Fuzzy Systems*, vol. 1, no. 2, pp. 98–110, 1993.

[3] ——, "The possibilistic c-means algorithm: insights and recommendations," *IEEE Trans. on Fuzzy Systems*, vol. 4, no. 3, pp. 385–393, 1996.

[4] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum Press, 1981.

[5] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. NY: Wiley Interscience, 1973.

[6] R. J. Hathaway, J. W. Davenport, and J. C. Bezdek, "Relational duals of the c-means algorithms," *Pattern Recognition*, vol. 22, pp. 205–212, 1989.

[7] R. J. Hathaway and J. C. Bezdek, "Nerf c-means: Non-euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, p. 429437, 1994.

[8] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," *Pattern Recognition*, vol. 41, no. 1, pp. 176–190, Jan. 2008. [Online]. Available: http://eprints.pascal-network.org/archive/00009117/01/pr08.pdf

[9] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen, "Multiple kernel fuzzy clustering," *IEEE T. Fuzzy Systems*, vol. 20, no. 1, pp. 120–134, 2012.

[10] J. Bezdek and R. Hathaway, "Vat: a tool for visual assessment of (cluster) tendency," *Proc. IEEE Int. Conf. Neural Networks*, vol. 3, pp. 2225–2230, 2002.

[11] D. Titterington, A. Smith, and U. Makov, *Statistical Analysis of Finite Mixture Distributions*. NY: Wiley, 1985.

[12] J. C. Bezdek, J. M. Keller, R. Krishnapuram, and N. R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. NY: Kluwer, 1999.

[13] A. Jain and R. Dubes, *Algorithms for Clustering Data*. NJ: Prentice Hall, 1988.

[14] B. Everitt, *Graphical Techniques for Multivariate Data*. NY: North Holland, 1978.

[15] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*. NY: Academic Press, 2006.

[16] R. Xu and D. C. Wunsch, *Clustering*. NJ: IEEE Press, 2009.

[17] J. V. de Oliveira and W. Pedrycz, *Advances in Fuzzy*

U2	x_1	x_2	x_3	x_4	x_5	x_6
c_1	0.00002	1	1	1	0.00001	0.00001
c_2	0.00001	0.00001	0.00001	1	1	1

U3	x_1	x_2	x_3	x_4	x_5	x_6
c_1	0.00001	1	1	1	0.00001	0.00001
c_2	0.00002	0.00001	0.00001	1	1	1

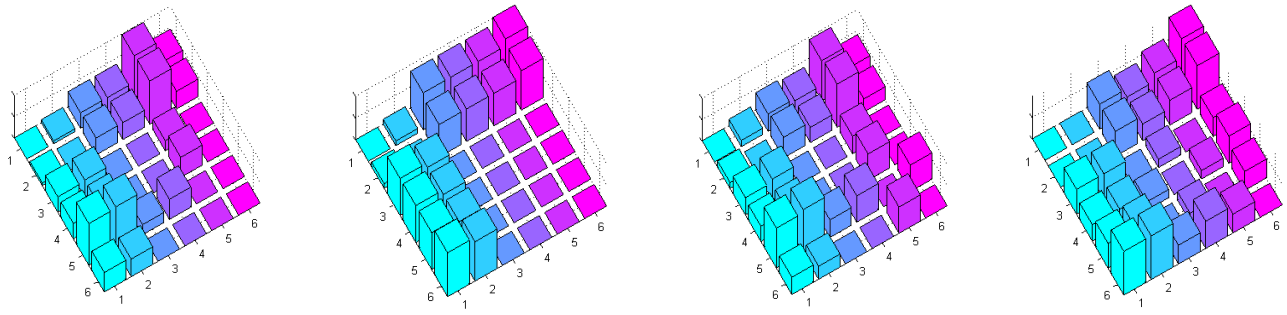
U4	x_1	x_2	x_3	x_4	x_5	x_6
c_1	1	1	1	0.00001	0.00001	0.00001
c_2	0.00001	0.00001	0.00001	1	1	1
c_3	1	1	1	0.00001	0.00001	0.00001

U5	x_1	x_2	x_3	x_4	x_5	x_6
c_1	1	1	1	0.00001	0.00001	0.00001
c_2	0.00001	0.00001	0.00001	1	1	1

U6	x_1	x_2	x_3	x_4	x_5	x_6
c_1	1	1	1	0.00001	0.00001	0.00001
c_2	1	1	1	0.00001	0.00001	0.00001
c_3	1	1	1	0.00001	0.00001	0.00001

U7	x_1	x_2	x_3	x_4	x_5	x_6
c_1	1	1	1	0.00001	0.00001	0.00001
c_2	0.00001	0.00001	0.00001	1	1	1
c_3	1	1	1	0.00001	0.00001	0.00001
c_4	0.00001	0.00001	0.00001	1	1	1
c_5	1	1	1	0.00001	0.00001	0.00001
c_6	0.00001	0.00001	0.00001	1	1	1

Fig. 4. Example $c \times 6$ candidate possibilistic partitions. Note, in $U2$, x_4 belongs to both cluster 1 and cluster 2 and x_1 is *noise*. In $U3$, x_4 belongs to both cluster 1 and cluster 2 and x_1 has a higher membership value in $c1$ than in $U2$. In $U4$, clusters 1 and 3 are co-incident. We consider $U5$ to be the reference partition for the set $\{U4, U5, U6, U7\}$. In $U6$, all clusters are co-incident in a single actual cluster. In $U7$, clusters 1, 3 and 5 are co-incident and clusters 2, 4 and 6 are co-incident.



(a) Anderson et al. with L_P-norm (this paper)

(b) Anderson et al. with soft Jaccard (this paper)

(c) Zhou et al. [37]

(d) Hullermeier et al. fuzzy Rand (reported as $1 - s$) [34]

Fig. 5. Graphical illustration of the dissimilarity matrices for possibilistic partitions. Each matrix has been normalized between its minimum and maximum to assist visual analysis. Index 1 is partition $U2$, 2 is $U3$, etc. Hullermeier's index is converted into a distance using $1 - s$.

Clustering and its Applications. NJ: Wiley, 2007.

[18] P. Arabie, L. J. Hubert, and R. Desoete, *Clustering and Classification*. NJ: World Scientific, 1996.

[19] R. Hoppner, R. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*. UK: Wiley and Sons, 1999.

[20] P. H. A. Sneath and R. R. Sokal, *Numerical Taxonomy - The Principles and Practice of Numerical Classification*. CA: W. H. Freeman, 1973.

[21] J.-P. Mei and L. Chen, "A fuzzy approach for multi-type relational data clustering," *IEEE T. Fuzzy Systems*, vol. 20, no. 2, pp. 358–371, 2012.

[22] G. Beliakov, S. James, and G. Li, "Learning choquet-integral-based metrics for semisupervised clustering," *IEEE T. Fuzzy Systems*, vol. 19, no. 3, pp. 562–574, 2011.

[23] E. Hullermeier, M. Rifqi, S. Henzgen, and R. Senge, "Comparing fuzzy partitions: A generalization of the rand index and related measures," *IEEE Transactions on Fuzzy Systems*, no. 20, pp. 546–556, 2012.

[24] T. Havens, J. Bezdek, and M. Palaniswami, "Cluster validity for kernel fuzzy clustering," *Proc. IEEE Int. Conf. Fuzzy Systems*, 2012.

[25] V. Singh, L. Mukherjee, J. Peng, and J. Xu, "Ensemble clustering using semidefinite programming with applications," *Mach. Learn.*, vol. 79, no. 1-2, pp. 177–200, May 2010. [Online]. Available: http://dx.doi.org/10.1007/s10994-009-5158-y

[26] A. L. N. Fred and A. K. Jain, "Data Clustering Using Evidence Accumulation," *Pattern Recognition, International Conference on*, vol. 4, pp. 40 276+, 2002. [Online]. Available: http://dx.doi.org/10.1109/ICPR.2002.1047450

[27] A. Strehl, J. Ghosh, and C. Cardie, "Cluster ensembles - a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583–617, 2002.

[28] *Adaptive clustering ensembles*, vol. 1, 2004. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1334105

[29] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, 2000.

[30] R. R. Yager, "On ordered weighted averaging aggrega-

tion operators in multicriteria decision making," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 18, pp. 183–190, 1988.

[31] D. T. Anderson, J. C. Bezdek, J. M. Keller, and M. Popescu, "A comparison of five fuzzy rand indices," *Information Processing and Management of Uncertainty*, 2010.

[32] R. J. G. B. Campello, "A fuzzy extension of the rand index and other related indexes for clustering and classification assessment," *Patt. Recog. Letters*, vol. 28, pp. 833–841, 2007.

[33] H. Frigui, C. Hwang, and F. C. H. Rhee, "Clustering and aggregation of relational data with applications to image database categorization," *Pattern Recognition*, vol. 40, p. 30533068, 2007.

[34] E. Hullermeier and M. Rifqi, "A fuzzy variant of the rand index for comparing clustering structures," *Proc. IFSA*, pp. 1–6, 2009.

[35] R. K. Brower, "Extending the rand, adjusted rand, and jaccard indices to fuzzy partitions," *J. Intell. Inf. Systems*, vol. 32, pp. 213–235, 2009.

[36] M. H. Coen, M. H. Ansari, and N. Fillmore, "Comparing clusterings in space," in *ICML*, 2010, pp. 231–238.

[37] D. Zhou, J. Li, and H. Zha, "A new mallows distance based metric for comparing clusterings," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 1028–1035. [Online]. Available: http://doi.acm.org/10.1145/1102351.1102481

[38] E. Levina and P. Bickel, "The earth mover's distance is the mallows distance: Some insights from statistics," *International Conference on Computer Vision*, pp. 251–256, 2001.

[39] S. Shirdhonkar and D. Jacobs, "Approximate earth mover's distance in linear time," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[40] H. Timm, C. Borgelt, C. Doring, and R. Kruse, "An extension to possibilistic fuzzy cluster analysis," *Fuzzy Sets and Systems*, vol. 147, no. 1, pp. 3–16, Oct. 2004. [Online]. Available: http://dx.doi.org/10.1016/j.fss.2003.11.009

[41] M. de Cáceres, F. Oliva, and X. Font, "On relational possibilistic clustering." *Pattern Recognition*, vol. 39, no. 11, pp. 2010–2024, 2006.

[42] M.-S. Yang and C.-Y. Lai, "A robust automatic merging possibilistic clustering method," *Trans. Fuz Sys.*, vol. 19, no. 1, pp. 26–41, Feb. 2011. [Online]. Available: http://dx.doi.org/10.1109/TFUZZ.2010.2077640