A Formal Learning Theory for Three-Way Clustering

Andrea Campagner(✉) and Davide Ciucci

Università degli Studi di Milano-Bicocca, Milan, Italy
a.campagner@campus.unimib.it, davide.ciucci@unimib.it

Abstract. In this work, we study the theoretical properties, from the perspective of learning theory, of three-way clustering and related formalisms, such as rough clustering or interval-valued clustering. In particular, we generalize to this setting recent axiomatic characterization results that have been discussed for classical hard clustering. After proposing an axiom system for three-way clustering, which we argue is a compatible weakening of the traditional hard clustering one, we provide a constructive proof of an existence theorem, that is, we show an algorithm which satisfies the proposed axioms. We also propose an axiomatic characterization of the three-way k-means algorithm family and draw comparisons between the two approaches.

Keywords: Three-way clustering · Rough clustering · Interval-set clustering · Learning theory

1 Introduction

Clustering, that is the unsupervised task of grouping objects into groups by account of their similarity [30], is a popular and important task in data analysis and related fields. Several clustering approaches have been proposed, such as hierarchical and partitive [24], density-based [16] or subspace-based [27], and have been successfully applied to different domains [9].

Compared to other areas in Machine Learning, however, the study of the formal properties of clustering, from a learning theory perspective [26], have been lacking, convergence or soundness results for specific algorithms aside [25].

In the recent years, starting from the seminal work of Kleinberg [10], there has been an increasing interest toward the study of the learnability of clustering, focusing on formal characterizations based on an *axiomatic* perspective: that is, studying systems of axioms that clustering methods should satisfy and then prove either impossibility theorems or characterization results.

A major limitation of these works consists in the fact that they apply only to *hard clustering* methods, that is methods in which each object is definitely assigned to one and only cluster. Today, however, many *soft clustering* [23] approaches have been developed and shown to be effective in practical applications: probabilistic clustering methods [17], fuzzy clustering [3], possibilistic

J. Davis and K. Tabia (Eds.): SUM 2020, LNAI 12322, pp. 128–140, 2020.
https://doi.org/10.1007/978-3-030-58449-8_9

clustering [12], credal clustering [8], three-way clustering [32] and related formalisms [13,31]. Contrary to traditional hard clustering approaches, soft clustering methods allow clusters to overlap or, either, the relation of containment of objects into cluster to be only partially or imprecisely defined.

In this article, we start to address this gap by extending the available results to soft clustering, in particular we will study a formal characterization of three-way clustering and related approaches (e.g. rough clustering, interval clustering). Specifically, in Sect. 2, we present the necessary background about three-way clustering and the learning theory of clustering; in Sect. 3, we present the main content of the paper by generalizing the learning theory of hard clustering to three-way clustering; finally, in Sect. 4, we discuss the proposed approach and describe some possible future work.

2 Background

2.1 Formal Theory of Clustering

Let X be a set of objects and $d : X \times X \mapsto \mathbb{R}$ be a *distance* function, i.e. a function s.t.:

$$d(x, y) \geq 0 \wedge d(x, y) = 0 \text{ iff } x = y \tag{1}$$

$$d(x, y) = d(y, x) \tag{2}$$

Remark. *We notice that, formally, a distance d should also satisfy the triangle inequality $\forall x, y, z \in X\ d(x, y) \leq d(x, z) + d(z, y)$. Functions not satisfying the triangle inequality are more usually denoted as* semi-distances. *However, since [10], in the literature on formal clustering theory it is customary to not make such a distinction.*

Let $\mathcal{D}_X$ be the collection of all distance functions over X and $\Pi(X)$ be the collection of partitions over X. A partition π is *trivial* if $\pi = antidiscr(X) = \{X\}$ or $\pi = discr(X) = \{\{x\} : x \in X\}$ and denote as $\hat{\Pi}(X)$ the collection of non-trivial partitions.

Definition 1. A *clustering algorithm* is a computable function $c : \mathcal{D}_X \mapsto \Pi(X)$.

Given d, then $c(d) = \pi = \{\pi_1, ..., \pi_n\}$ where each π_i is a *cluster*. We denote the case in which two objects x, y belong to the same cluster π_i as $x \sim_\pi y$ The formal study of clustering algorithms, after Kleinberg [10], starts from the definition of characterization axioms:

Axiom 1 (Scale Invariance). *A clustering algorithm c is* scale invariant *if, for any $d \in \mathcal{D}_X$ and $\alpha > 0$, $c(d) = c(\alpha \cdot d)$.*

Axiom 2 (Richness). *A clustering algorithm c is* rich *if $Range(c) = \Pi(X)$: that is, for each $\pi \in \Pi(X)$, $\exists d \in \mathcal{D}_X$ s.t. $c(d) = \pi$.*

Axiom 3 (Consistency). *Let $d, d' \in \mathcal{D}_X$ and $\pi \in \Pi(X)$. Then, d' is a π-transformation of d if*

$$\forall x \sim_\pi y . d'(x,y) \leq d(x,y) \tag{3}$$
$$\forall x \not\sim_\pi y . d'(x,y) \geq d(x,y) \tag{4}$$

A clustering algorithm c is consistent *if, given d s.t. $c(d) = \pi$, for any d' $\pi(X)$-transformation of d it holds that $c(d') = c(d)$.*

The following impossibility theorem represents a seminal result in the formal learning theory of clustering:

Theorem 1 ([10]). *If $|X| \geq 2$ then no clustering algorithm satisfies Axioms 1, 2, 3.*

Corollary 1 ([10]). *For each pair of Axioms among 1, 2, 3 there exists a clustering algorithm that satisfies it.*

Remark. *We note that the axioms, and the proofs of Theorems 1 and Corollary 1, allow one to arbitrarily choose the distance function, irrespective of the nature and topological structures of the instances in X. While this assumption may seem overly general, it is to note that in the definitions the instances of X are completely abstract, and the topological space is entirely determined by the function d. In this respect, letting d vary arbitrarily may be seen as requiring that, no matter the nature of the distance function chosen for the given application, a clustering should respect some properties w.r.t. the chosen distance. This is in analogy with the distribution-independence assumption in the definition of PAC learnability for supervised learning theory [26].*

It is to note that these results can be interpreted similarly to the No Free Lunch theorem for supervised learning: that is, there is no clustering algorithm that (under the requirement of allowing to return every possible clustering) satisfies two intuitively appealing criteria. Indeed, Zadeh et al. [35] have shown that the most problematic constraint is due to the Richness axiom and proposed an alternative formalization based on the concept of k-clustering algorithms (i.e. clustering algorithms which require an additional input $k \in \mathbb{N}^+$) and the *k-Richness* axiom:

Axiom. 2' (k-Richness). *A k-clustering algorithm $c_k : \mathcal{D}_X \mapsto \Pi_k(X)$ is* k-rich *if $\forall k, Range(c_k) = \Pi_k(X)$, where $\Pi_k(X)$ is the collection of k-partitions on X.*

The authors also showed that, considering k-Richness in place of Richness, provides a consistent set of axioms:

Theorem 2 ([35]). *There exists a k-clustering algorithm that satisfies Axioms 1, 2', 3.*

Lastly, we recall the work of Ben-David et al. [2] on *clustering quality measures* (CQM), i.e. functions $q : \Pi(X) \times \mathcal{D}_X \mapsto \mathbb{R}$, which showed that the following set of axioms represents a consistent formalization of these measures:

Axiom 1_q (Scale Invariance). *A CQM q is scale invariant if* $\forall \alpha > 0, \pi \in \Pi(X)$, $q(\pi, d) = q(\pi, \alpha \cdot d)$.

Axiom 2_q (Richness). *A CQM q is rich if* $\forall \pi \in \hat{\Pi}(X)$ *exists* $d \in \mathcal{D}_X$ *s.t.* $\pi = argmax_{\pi' \in \hat{\Pi}(X)} \{q(\pi', d)\}$.

Axiom 3_q (Consistency). *A CQM q is consistent if given* $d \in \mathcal{D}_x$, $\pi \in \Pi(x)$, *for any* $\pi(X)$*-transformation* d' *of* d *it holds that* $q(\pi, d') \geq q(\pi, d)$.

Theorem 3 ([2]). *There exists a CQM that satisfies Axioms* 1_q, 2_q, 3_q.

We can note that, even though Axioms 1_q through 3_q are defined for *clustering quality measures*, they also implicitly define a *clustering algorithm* by $c(d; q) = argmax_{\pi' \in \hat{\Pi}(X)} \{q(\pi', d)\}$

2.2 Three-Way Clustering and Related Formalisms

An orthopair on a universe X is defined as $O = \langle P, N \rangle$, where $P \cap N = \emptyset$. From P and N a third set, can be defined as $Bnd = X \setminus (P \cup N)$. In the setting of clustering an orthopair can be understood as an uncertain or imprecisely known cluster: the objects in P are those that surely belong to the cluster (P = Positive), those in N are the ones that surely do not belong to the cluster (N = negative), and the objects in Bnd are those that may possibly belong to the cluster (Bnd = Boundary). In the setting of three-way clustering P is also called the *Core* region of the cluster, and Bnd as the *Fringe* region. In the following, we will denote a cluster as the orthopair $O_i = (Core_i, Fringe_i)$.

Different clustering frameworks have been proposed based on the idea of employing orthopairs as a representation of clusters, namely rough clustering [15], interval-set clustering [31], three-way clustering [32] and shadowed set clustering [18]. In these frameworks, a variety of different clustering algorithms have been proposed: rough k-means [15,19,21] and variations based on evolutionary computing [14] or the principle of indifference [22] for the optimal selection of the thresholds that define the $Core$ and $Fringe$ regions, three-way c-means [28,36], different three-way clustering algorithms that automatically determine the appropriate thresholds or number of clusters such as gravitational search-based three-way clustering [33], variance-based three-way clustering [1], three-way clustering based on mathematical morphology [29] or density-based [34] and hierarchical [5] three-way clustering, and many others.

In the context of this paper we will not consider specific three-way clustering algorithms, as we will be primarily interested in the general formalism behind these clustering frameworks that we now recall. As highlighted previously different frameworks have been proposed, based on similar but different axiom requirements: rough clustering, interval-set clustering, three-way clustering.

A rough clustering is defined as a collection of $\mathcal{O} = \{O_1, ..., O_n\}$ of orthopairs satisfying:

(R1) $\forall i \neq j, Core_i \cap Core_j = Core_i \cap Fringe_j = Core_j \cap Fringe_i = \emptyset$

(R2) $\forall x \in X$, $\nexists i$ s.t. $x \in Core_i \rightarrow \exists i \neq j$ s.t. $x \in Fringe_i, Fringe_j$.

On the other hand, both interval-set clustering and three-way clustering are defined as collections $\mathcal{O} = \{O_1, ..., O_n\}$ of orthopairs s.t.:

(T1) $\forall i, Core_i \neq \emptyset$
(T2) $\bigcup_i (Core_i \cup Fringe_i) = X$
(T3) $\forall i \neq j$, $Core_i \cap Core_j = Core_i \cap Fringe_j = Core_j \cap Fringe_i = \emptyset$

Finally, shadowed set clustering [18] adopts a framework which is instead based on fuzzy clustering, where the degree of membership of an object $x \in X$ to a cluster $C \in \pi$ is given by a membership function $C : X \mapsto [0,1]$. Compared with standard fuzzy clustering, in shadowed set clustering the membership function for each cluster C are then discretized into three regions, which are then equivalent to the three regions in three-way clustering (i.e., $Core_C, Fringe_C$ and $Ext_C = (Core_C \cup Fringe_C)^c$) [20].

Thus, while the different clustering frameworks are based on the same mathematical representation (i.e., orthopairs), there are some differences: rough clustering allows the core regions to be empty (in this case, the object is required to belong to at least two fringe regions); interval–set (and three–way) clustering require the core regions to be non–empty and allows objects to be in only one fringe region.

Recently, the notion of an *orthopartition* [4] has been proposed as a unified representation for clustering based on orthopairs. Formally, an orthopartition is defined as a collection $\mathcal{O}$ of orthopairs s.t.:

(O1) $\forall i \neq j$ $Core_i \cap Core_j = Core_i \cap Fringe_j = Core_j \cap Fringe_i = \emptyset$
(O2) $\bigcup_i (Core_i \cup Fringe_i) = X$
(O3) $\forall x \in U$ $(\exists i$ s.t. $x \in Fringe_i) \rightarrow (\exists j \neq i$ s.t. $x \in Fringe_j)$

It can easily be seen that the axioms for orthopartitions more closely follow the ones for rough clustering (Axiom O3 does not hold for three-way clustering). However, in [4], it has been shown that every three-way clustering can easily be transformed in an orthopartition by isolating in an ad-hoc cluster the elements not satisfying (O3). As all the different representations can be transformed into each other, in the following, we will thus refer generally to three-way clustering as a general term for clustering based on orthopairs.

Since a three-way clustering represents an incomplete or uncertain state of knowledge about a clustering (i.e. about which specific clusters do the objects belong), we can also represent a three-way clustering as a collection of *consistent clusterings*, that is given a three-way clustering $\mathcal{O}$:

$$\Sigma(\mathcal{O}) = \{\pi \in \Pi(X) : \pi \text{ is consistent with } \mathcal{O}\} \tag{5}$$

where π is consistent with $\mathcal{O}$ iff $\forall O_i \in \mathcal{O}, \exists \pi_j \in \pi$ s.t. $\pi_j \subseteq Core_i \cup Fringe_i$.

We notice that, in general, a collection of clusterings $\mathcal{C}$ does not necessarily represents the collection of consistent clusterings for any given three-way clustering $\mathcal{O}$. However, it can also be easily seen that each collection of clusterings

$\mathcal{C}$ can be extended to a collection of consistent clusterings (for a given three-way clustering $\mathcal{O}$). Thus, when we refer to a collection of clusterings, we will implicitly refer to its extension that we denote as $tw(\Sigma)$, where Σ is a collection of clusterings. The vacuous three-way clustering is defined as $\mathcal{O}_v$ s.t. $\Sigma(\mathcal{O}_v) = \Pi(X)$. Let $\mathcal{O}$ be a three-way clustering, we denote by $Core(\mathcal{O}) \subset X$ the collection of objects in the core regions of the clusters of $\mathcal{O}$.

If we denote as $\mathbb{O}(X)$ the set of three-way clusterings over X, then a *three-way clustering algorithm* is a computable function $c_{tw} : \mathcal{D}_X \mapsto \mathbb{O}(X)$.

3 Formal Theory of Three-Way Clustering

Our aim is to study the learnability properties and formal characterization of three-way clustering. One aspect that should be considered, in this respect, is the increased flexibility derived from adopting the three-way formalism, which is due not only to the increased model complexity, but also to the fact that it allows to conceive *weakenings* of the axioms proposed for hard clustering, as long as they retain compatibility with the standard case. In particular, since as previously argued, the Richness axiom represents the most problematic constraint, we will study possible weakenings of it which are meaningful in the three-way clustering setting. This, however, should be done with care, e.g.. the following naive consistent weakening of the Richness axiom:

$$\bigcup_{\mathcal{O} \in Range(c_{tw})} \Sigma(\mathcal{O}) = \Pi(X)$$

would clearly be too permissive, as it would admit always returning the *vacuous* three-way clustering as output. Similarly, requiring that $Range(c_{tw}) = \mathbb{O}(X)$ would be too strong a requirement: as a consequence of Theorem 1 it would result in an unsatisfiable constraint.

The following axiom, which is intermediate in strength between Axiom 2 and Axiom 2_q (as shown previously, any clustering quality measure implicitly defines a clustering algorithm), represents a weakening of the Richness axiom which is coherent with the three-way clustering setting:

Axiom 2_{tw} (Almost Richness). *A three-way clustering algorithm c_{tw} is* almost rich *if*

$$\bigcup_{\mathcal{O} \in Range(c_{tw})} \Sigma(\mathcal{O}) = \Pi(X) \tag{6}$$

and

$$\forall \pi \in \hat{\Pi}(X), \exists d \in \mathcal{D}_X \; s.t. \; c_{tw}(d) = \pi \tag{7}$$

First, we notice that, obviously, when restricted to hard clustering algorithms Almost Richness reduces to Richness.

Proposition 1. *If c is a clustering algorithm, then it satisfies Axiom 2 iff it satisfies Axiom 2_{tw}.*

Proof. For a clustering algorithm c its output $c(d)$ is always a single partition π. Thus, $\Sigma(\{\pi\}) = \pi$ and thus, if c is almost rich it is also rich. Equation (7) becomes redundant in this particular case. The converse (richness implies almost richness) is evident.

Second, we note that in Axiom 2_{tw}, we restrict the range to $\hat{\Pi}(X)$ rather than $\Pi(X)$, in analogy with Axiom 2_q; in this sense, as stated above, the proposed Axiom is intermediate in strength between Axioms 2 and 2_q.

On the other hand, as regards Axioms 1 and 3, we simply require that they hold for each possible three-way clustering $\mathcal{O}$ (thus, we do not weaken these two axioms).

In order to prove that Axiom 2_{tw}, together with Axioms 1 and 3, characterizes three-way clustering, we first introduce the notion of a CQM s.t. the resulting c^q_{tw} is almost rich.

For a pair of clusters π_i, π_j let $s(\pi_i) = \frac{1}{2|\pi_i|}\sum_{x \neq y \in \pi_i} d(x, y)$ be the mean distance of the elements inside cluster π_i and $d(\pi_i, \pi_j) = \frac{1}{|\pi_i||\pi_j|}\sum_{x\in\pi_i}\sum_{y\in\pi_j} d(x, y)$ the mean distance between elements belonging to two different clusters π_i, π_j. Given a partition π and a distance d, we define a CQM $q_{tw} : \Pi(X) \times \mathcal{D}_X \mapsto \mathbb{R}^2$ as

$$q_{tw}(\pi, d) = \langle q_{intra}, q_{inter}\rangle \tag{8}$$

where

$$q_{intra}(\pi, d) = \frac{1}{|\pi|}\sum_{\pi_i\in\pi} s(\pi_i) - min_{x\neq y\in X} d(x, y) \tag{9}$$

$$q_{inter}(\pi, d) = \frac{1}{|\pi|^2}\sum_{\pi_i\neq\pi_j\in\pi} d(\pi_i, \pi_j) - min_{x\neq y\in X} d(x, y) \tag{10}$$

Remark. *We notice that, strictly speaking, the introduced quality measure q_{tw} is not a CQM, as a CQM is defined as a function $q : \Pi(X) \times \mathcal{D}_X \mapsto \mathbb{R}$ while $q_{tw} : \Pi(X) \times \mathcal{D}_X \mapsto \mathbb{R}^2$.*

Definition 2. Given two clustering π^1, π^2 and a distance function, we say that $q_{tw}(\pi^1, d) < q_{tw}(\pi^2, d)$ if both:

$$q_{intra}(\pi^1, d) \geq q_{intra}(\pi^2, d) \tag{11}$$

$$q_{inter}(\pi^1, d) \leq q_{inter}(\pi^2, d) \tag{12}$$

and at least one of the two is strict. Then, we say that $\pi^1 <_q \pi^2$ if $q_{tw}(\pi^1, d) < q_{tw}(\pi^2, d)$.

The idea is that if $\pi^1 <_q \pi^2$ then instances in π_1 have greater intra-cluster distance and smaller inter-cluster distance.

The following result shows that, indeed, the three Axioms provide a characterization for three-way clustering.

Algorithm 1. Three-way Clustering based on q_{tw}

```
Require: d distance function
  Σ = ∅
  for π ∈ Π(X) do
      check := ⊤
      for π' ≠ π ∈ Π(X) do
          if π' >_q π then
              check := ⊥
              break
          end if
      end for
      if check then
          Σ.append(π)
      end if
  end for
  Return tw(Σ)
```

Theorem 4. *There exists a three-way clustering algorithm that satisfies Axioms 1, 2_{tw} and 3.*

Proof. The Theorem can be proven based on the previously defined CQM q_{tw}.

Indeed, from q_{tw}, we can define a three-way clustering algorithm c^q_{tw} as shown in Algorithm 1. It can be verified that, $\forall d \in \mathcal{D}_X$, $c^q_{tw}(d) = \{\pi \in \Pi(X) : \nexists \pi' \in \Pi(X), \pi' >_q \pi\}$. It is easily shown that c^q_{tw} satisfies Scale-Invariance and Consistency (when restricted to pairs of objects x, y in the core regions of $tw(\Sigma)$, where Σ is the result of c^q_{tw}). We thus show only the proof for Almost Richness.

Let π be a given non-trivial clustering and let d be the distance function defined $\forall x, y$ as

$$d(x,y) = \begin{cases} \epsilon & \text{if } x \sim_\pi y \\ \alpha & \text{if } x \not\sim_\pi y \end{cases}$$

with $\alpha \gg \epsilon$. Then, $q_{tw}(\pi, d) = \langle 0, \alpha - \epsilon\rangle$ and evidently, for any other π', $q_{tw}(\pi', d) < q_{tw}(\pi, d)$ (hence, $\pi' <_q \pi$). This satisfies the second condition of the Axiom.

As regards the first condition, let d be s.t. $\forall x, y$, $d(x, y) = \epsilon$. Then, for any $\pi \in \Pi(X)$ $q_{tw}(\pi, d) = \langle 0, 0\rangle$. The condition, and hence the result, follows. $\square$

The proof of Theorem 4 is constructive and directly provides a three-way clustering algorithm satisfying Axioms 1, 2_{tw}, 3. Further, evidently $\hat{\Pi}(X) \subset Range(c^q_{tw}) \subset \mathbb{O}(X)$ but future work should study how to provide a more precise specification of the range of c^q_{tw}.

As a limitation of this result, it is easy to observe that an exact implementation of this algorithm is not practical from a time complexity perspective: indeed, as the algorithm requires to compute the value of q for all possible clusterings its complexity is evidently exponential in $|X|$. A possible solution to this problem would be to define heuristic or randomized algorithms for implementing approximation of the c^q_{tw} three-way clustering algorithm (possibly, with proven quality bounds).

A different approach, instead, consists in studying other, more efficient three-way clustering algorithms and providing their axiomatic characterization, in order to understand their flexibility when compared with c^q_{tw}. We will provide such a characterization for the three-way k-means algorithm [28,36] (and related ones, such as rough k-means [15,21]). In order to prove this result, we will focus on a simplified single-step version of the algorithm, as defined in Algorithm 2.

Algorithm 2. Single-Step Three-way K-Means

Require: d distance function
Require: $x_1, ..., x_k \in X$ cluster centroids
Require: $\delta, \varDelta \in [0, 1]$ parameters
  Let $\hat{d}(x, y) := \frac{d(x,y)}{max_{a,b \in X} d(a,b)}$
  for $x \in X$ **do**
    $I := \{i : \hat{d}(x, x_i) \leq \varDelta\}$
    for $i \in I$ **do**
      if $(|I| = 1) \wedge (\hat{d}(x, x_i) \leq \delta)$ **then**
        $x \in Core_i$
      else
        $x \in Fringe_i$
      end if
    end for
  end for

Evidently, Algorithm 2 is Scale-Invariant (as it only uses the normalized distance) but it can easily be shown that is neither Consistent, nor Almost Rich, but it is k-Rich.

Example 1. For Consistency, consider a distance function d, let $a, b \in X$ s.t. $d(a, b) = max_{x,y \in X} d(x, y)$ and suppose that the result of Algorithm 2, denoted as $\mathcal{O}$, assigns a, b to two different clusters. Further, let d' s.t. $\forall x, y \in X \setminus \{a, b\} d'(x, y) = d(x, y)$, while $d(a, b) \ll d'(a, b)$. Then, evidently, d' is a $\mathcal{O}$-transformation of d, but Consistency is violated. For Almost Richness, it easily follows from the fact the result of Algorithm 2 contains exactly k clusters.

We can characterize this Algorithm (and similar algorithms such as three-way k-means [28] or rough k-means [13]) via the following two Axioms (together with Scale-Invariance):

Axiom 2_{twk} (Three-way k-Richness). *A three-way k-clustering algorithm $c^k_{tw} : \mathcal{D}_X \mapsto \mathbb{O}_k(X)$ is k-rich if $\forall k, Range(c^k_{tw}) = \mathbb{O}_k(X)$ where $\mathbb{O}_k(X) = \{\mathcal{O} \in \mathbb{O}(X) : |\mathcal{O}| = k\}$.*

That is, a three-way clustering algorithm is k-rich if as possible outputs (by changing the distance function) we can obtain all the orthopartions made exactly of k orthopairs. This requirement is a natural generalization of k-Richness to the setting of three-way clustering.

Axiom $\mathit{3}_{twk}$ **((δ, Δ)-Consistency).** *Let $d, d' \in \mathcal{D}_X$ and $\delta < \Delta \in [0,1]$. Then, d' is a (δ, Δ)-transformation of d if:*

$$sign(\frac{d(x,y)}{max_{a,b}d(a,b)} - \delta) = sign(\frac{d'(x,y)}{max_{a,b}d'(a,b)} - \delta) \tag{13}$$

$$sign(\frac{d(x,y)}{max_{a,b}d(a,b)} - \Delta) = sign(\frac{d'(x,y)}{max_{a,b}d'(a,b)} - \Delta) \tag{14}$$

A three-way clustering algorithm c_{tw} is (δ, Δ)-consistent if, given d s.t. $c_{tw}(d) = \mathcal{O}$, for any d' (δ, Δ)-transformation of d it holds that $c_{tw}(d') = \mathcal{O}$.

So, (δ, Δ)-consistency means that small changes in the distance function do not alter the clustering result. The notion of (δ, Δ)-consistency can be seen as a restricted form of Consistency, determined by two thresholds that are used to describe three different regions (a natural requirement in the setting of three-way clustering): the objects whose normalized distance is lower than δ; those for which the normalized distance is between Δ and δ; and those for which the normalized distance is greater than Δ).

Theorem 5. *Algorithm 2 satisfies Axioms 1, 2_{twk}, 3_{twk}.*

Proof. Evidently, Algorithm 2 is Scale-Invariant. Further, by construction, it is also (δ, Δ)-consistent w.r.t. its input parameters δ, Δ. Thus, we only need to show that it is Three-way k-Rich.

Let $\mathcal{O} \in \mathbb{O}_k(X)$ be the target three-way clustering and $\delta < \Delta$ the input parameters. For each cluster $O_i \in \mathcal{O}$ select one element $x_i \in O_i$. Then, for each x, if $x \in Core_i$ set $d(x, x_i) < \delta$ and if $x \in Fringe_i$ set $\delta < d(x, x_i) < \Delta$. For any two x, y s.t. they belong to different clusters set $d(x, y) = 1$. Then the output of Algorithm 2 in this case is exactly $\mathcal{O}$.

We can thus compare the two algorithms, c^q_{tw} and Three-way K-Means, through their characterization. Indeed, we can observe that the two algorithms can be seen as offering a trade-off between representational flexibility and computational efficiency. Indeed, c^q_{tw} is more flexible (as it does not require to set, a-priori, the number of clusters k) and satisfies a stricter notion of consistency. Thus, its result is more well-behaved w.r.t. coherent modifications to the distance function. However, it has exponential complexity in the size of X, as it requires an enumeration of all $\pi \in \Pi(X)$. On the other hand, having fixed both k, the number of clusters, and the cluster centroids (i.e., $x_1, ..., x_k$ in Algorithm 2), the complexity of Three-way K-Means is linear in $|X|$. However, the Three-way K-Means family of algorithms satisfies a weaker form of consistency (i.e., (δ, Δ)-consistency) and, further, requires to set both the number of clusters (which in practice is selected heuristically using criteria such as the Silhouette or cross-validation [11]) and the cluster centroids: this usually involves an iterative approach which, however, only guarantees convergence to a local optimum (as, even for traditional k-Means, the problem of finding the optimal k-clustering is NP-hard [6]).

4 Conclusion

In this article, we set out the foundations for the study of the theoretical properties of three-way clustering and related formalisms, from the perspective of computational learning theory. We provided an axiomatic characterization of three-way clustering and proved that, contrary to the case of traditional clustering, these requirements are consistent, i.e., there exists a three-way clustering algorithm satisfying them, which however has exponential time complexity. We then studied an axiomatic characterization of the popular Three-way k-Means family of clustering algorithm, showing that it provides a trade-off, favoring better time complexity against reduced flexibility. Our results represent a first step towards a formal study of three-way clustering and, as such, we think that the following open problems may be important to understand the formal properties of this increasingly popular clustering framework:

– What is the exact characterization of $Range(c^q_{tw})$? As we previously argued, it can easily be shown that $\hat{\Pi}(X) \subsetneq Range(c^q_{tw}) \subsetneq \mathbb{O}(X)$, but it is not clear which proper three-way clusterings can be represented by c^q_{tw};
– Is there a three-way clustering algorithm satisfying the following generalized Almost Richness axiom (together with Consistency and Scale-Invariance):

$$\forall \mathcal{O} \in \hat{\mathbb{O}}(X) \exists d \in \mathcal{D}_X \text{ s.t } c_{tw}(d) = \mathcal{O} \tag{15}$$

where $\hat{\mathbb{O}}(X)$ is the set of non-trivial three-way clusterings? Otherwise, what is the greatest subset of $\mathbb{O}(X)$ which admits a consistent and scale-invariant three-way clustering algorithm?
– While the time complexity of c^q_{tw} is exponential in $|X|$, can we find an approximation or randomization scheme for c^q_{tw} with provable error bounds?
– What is the learning-theoretic axiomatic characterization of other three-way clustering algorithms, such as three-way density-based clustering [34] or rough-set hierarchical clustering [5]?

More generally, and observing that rough k-means can be seen as a particular case of both evidential clustering [7] and possibilistic clustering [12], we can think to extend the learning-theoretic axiomatic characterization to these other soft clustering approaches.

References

1. Afridi, M.K., Azam, N., Yao, J.: Variance based three-way clustering approaches for handling overlapping clustering. Int. J. Approximate Reasoning **118**, 47–63 (2020)
2. Ben-David, S., Ackerman, M.: Measures of clustering quality: A working set of axioms for clustering. Proc. NIPS **2009**, 121–128 (2009)
3. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2), 191–203 (1984)

4. Campagner, A., Ciucci, D.: Orthopartitions and soft clustering: Soft mutual information measures for clustering validation. Knowl. Based Syst. **180**, 51–61 (2019)
5. Chen, D., Cui, D.W., Wang, C.X., Wang, Z.R.: A rough set-based hierarchical clustering algorithm for categorical data. Int. J. Inf. Technol. **12**(3), 149–159 (2006)
6. Dasgupta, S.: The hardness of k-means clustering. Department of Computer Science and Engineering, University of California, San Diego, Technical report (2008)
7. Denœux, T., Kanjanatarakul, O.: Beyond fuzzy, possibilistic and rough: An investigation of belief functions in clustering. In: Ferraro, M.B., Giordani, P., Vantaggi, B., Gagolewski, M., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Data Science. AISC, vol. 456, pp. 157–164. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-42972-4_20
8. Denœux, T., Masson, M.H.: Evclus: Evidential clustering of proximity data. IEEE Trans. Syst. Man Cybern. **34**(1), 95–109 (2004)
9. Kameshwaran, K., Malarvizhi, K.: Survey on clustering techniques in data mining. IJCSIT **5**(2), 2272–2276 (2014)
10. Kleinberg, J.M.: An impossibility theorem for clustering. Pro. NIPS **2003**, 463–470 (2003)
11. Kodinariya, T.M., Makwana, P.R.: Review on determining number of cluster in k-means clustering. Int. J. Adv. Res. Comput. Sci. Manag. Stud. **1**(6), 90–95 (2013)
12. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Trans. Syst. **1**(2), 98–110 (1993)
13. Lingras, P., Peters, G.: Rough clustering. WIREs Data Min. Knowl. Discov. **1**, 65–72 (2011)
14. Lingras, P.: Evolutionary rough K-means clustering. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS (LNAI), vol. 5589, pp. 68–75. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02962-2_9
15. Lingras, P., West, C.: Interval set clustering of web users with rough k-means. J. Intell. Inf. Syst. **23**(1), 5–16 (2004)
16. Loh, W.K., Park, Y.H.: A survey on density-based clustering algorithms. In: Ubiquitous Information Technologies and Applications, pp. 775–780. Springer, Berlin (2014)
17. MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, New York (2002)
18. Mitra, S., Pedrycz, W., Barman, B.: Shadowed c-means: Integrating fuzzy and rough clustering. Pattern Recogn. **43**(4), 1282–1291 (2010)
19. Murugesan, V.P., Murugesan, P.: A new initialization and performance measure for the rough k-means clustering. Soft Comput. 1–15 (2020)
20. Pedrycz, W.: Interpretation of clusters in the framework of shadowed sets. Pattern Recogn. Lett. **26**(15), 2439–2449 (2005)
21. Peters, G.: Some refinements of rough k-means clustering. Pattern Recogn. **39**(8), 1481–1491 (2006)
22. Peters, G.: Rough clustering utilizing the principle of indifference. Inf. Sci. **277**, 358–374 (2014)
23. Peters, G., Crespo, F., Lingras, P., Weber, R.: Soft clustering-fuzzy and rough approaches and their extensions and derivatives. Int. J. Approximate Reason. **54**(2), 307–322 (2013)
24. Reddy, C.K., Vinzamuri, B.: A survey of partitional and hierarchical clustering algorithms. In: Data Clustering, pp. 87–110. Chapman and Hall/CRC (2018)

25. Selim, S.Z., Ismail, M.A.: K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. IEEE Trans. Pattern Anal. Mach. Intell. **1**, 81–87 (1984)
26. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
27. Vijendra, S.: Efficient clustering for high dimensional data: Subspace based clustering and density based clustering. Inf. Technol. J. **10**(6), 1092–1105 (2011)
28. Wang, P., Shi, H., Yang, X., Mi, J.: Three-way k-means: Integrating k-means and three-way decision. Int. J. Mach. Learn. Cybern. **10**(10), 2767–2777 (2019). https://doi.org/10.1007/s13042-018-0901-y
29. Wang, P., Yao, Y.: Ce3: A three-way clustering method based on mathematical morphology. Knowl. Based Syst. **155**, 54–65 (2018)
30. Xu, R., Wunsch, D.: Clustering, vol. 10. Wiley, Hoboken (2008)
31. Yao, Y., Lingras, P., Wang, R., Miao, D.: Interval set cluster analysis: a re-formulation. In: Sakai, H., Chakraborty, M.K., Hassanien, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS (LNAI), vol. 5908, pp. 398–405. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-10646-0_48
32. Yu, H.: A framework of three-way cluster analysis. In: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Ślęzak, D., Zielosko, B. (eds.) IJCRS 2017. LNCS (LNAI), vol. 10314, pp. 300–312. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60840-2_22
33. Yu, H., Chang, Z., Wang, G., Chen, X.: An efficient three-way clustering algorithm based on gravitational search. Int. J. Mach. Learn. Cybern. **11**(5), 1003–1016 (2019). https://doi.org/10.1007/s13042-019-00988-5
34. Yu, H., Chen, L., Yao, J., Wang, X.: A three-way clustering method based on an improved dbscan algorithm. Physica A Stat. Mech. Appl. **535**, 122289 (2019)
35. Zadeh, R.B., Ben-David, S.: A uniqueness theorem for clustering. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, pp. 639–646 (2009)
36. Zhang, K.: A three-way c-means algorithm. Appl. Soft Comput. **82**, 105536 (2019)