On Clustering in Qualitative Spatial and Temporal Reasoning

Abderrahmane Boukontar, Jean-François Condotta, Yakoub Salhi
CRIL CNRS UMR 8188 - Université d'Artois
Lens, France

Abstract

Our understanding of the world is intricately linked to both the spatial arrangement of objects and the timing of events. Knowledge-dependent systems employ mechanisms like Qualitative Spatial and Temporal Reasoning (QSTR) to effectively process and interpret this information. This article explores application of QSTR in data clustering, offering several contributions. These include introducing a formal clustering framework for qualitative data, implementing a satisfiability encoding to compute a clustering, introducing two appropriate distance measures for Qualitative Relation Networks, and experimentally validating through adaptations of k-means and Hierarchical Agglomerative Clustering algorithms.

Introduction

Recognizing patterns within complex data is crucial for understanding intricate information. Clustering, a process rooted in discerning inherent structures based on similarities (Kaufman and Rousseeuw 2009), and deals with diverse data forms such as numeric, graphs (Schaeffer 2007), images (Chang et al. 2017), texts (Aggarwal and Zhai 2012), or symbolic data (de Carvalho, Csernel, and Lechevallier 2009; Boudane et al. 2017; Kejžar, Korenjak-Černe, and Batagelj 2021).

Qualitative Spatial and Temporal Reasoning is an AI symbolic framework that is closely aligned with human reasoning, making it easier to interpret extracted knowledge when dealing with complex data. By employing formalisms like Point Algebra (PA) (Vilain, Kautz, and Van Beek 1990), Interval Algebra (IA) (Allen 1983), and Region Connection Calculus (RCC) (Randell, Cui, and Cohn 1992) via qualitative relations, QSTR effectively models and reasons about spatio-temporal knowledge across various domains.

Utilizing qualitative relations as an intermediary representation, QSTR can endow users with an explanatory capability to uncover the rationale behind cluster formation. This facilitates generating insights from data structures, enhancing understanding of discovered patterns. For similar reasons, qualitative approaches have been employed in the field of data mining and knowledge extraction (Wang et al. 2018; Salhi 2019; Homem et al. 2020; Boukontar, Condotta, and Salhi 2022).

Our contributions in this article are manyfold. First, we introduce a formal framework for accomplishing the clustering task in QSTR. Second, we propose a Boolean Satisfiability (SAT) encoding to verify the feasibility of building a clustering comprising exclusively consistent clusters. Third, we introduce distance measures specific to QSTR for their use with conventional clustering algorithms. Finally, we conclude with an experimental study of our framework using an adaptation of the k-means algorithm.

Data Clustering

Let $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ represent a finite non-empty set of items referred to as *data instances*. In traditional data clustering, each data instance x_i comprises a vector of m features $(x_{i,1}, x_{i,2}, \ldots, x_{i,m})$ where $x_{i,j} \in \mathbb{R}$. Clustering data instances aims to maximize similarity within clusters and minimize similarity between them (Aggarwal and Reddy 2013). Similarity, commonly measured using distance metrics like Euclidean and Manhattan distances, evaluates how close or similar data instances are from each other (Sinwar and Kaushik 2014). Similarity can also be expressed as a coefficient (Goyal and Aggarwal 2017) like Simple Matching coefficient, or Hamming coefficient, providing an other perspective by quantifying the number of shared features between two data instances relative to the total number of features. Alternatively, when it is suitable to dismiss the paired absence of features in the data instances, Jaccard's coefficient emerges as an alternative (Ferdous and others 2009; Irani, Pise, and Phatak 2016).

Hierarchical clustering methods like the widely-used *Hierarchical Agglomerative Clustering* (HAC) algorithm (Murtagh and Contreras 2012), progressively forms clusters in ascending order. Starting with each data instance in its own cluster, the algorithm iteratively merges the two closest clusters using a chosen linkage technique until all instances belong to a single cluster. Various linkage techniques, such as the *single linkage* method, which evaluates similarity between two clusters $\mathcal{P}_e, \mathcal{P}_f$ as $dis_{single}(\mathcal{P}_e, \mathcal{P}_f) = \min_{x_i \in \mathcal{P}_e, x_j \in \mathcal{P}_f} d(x_i, x_j)$ (Yim and Ramdeen 2015; Jarman 2020), consider the minimum pairwise distance between data points in the clusters. Other techniques are detailed in (Nielsen 2016).

Partitional clustering methods (Celebi 2014), differ from hierarchical approaches by aiming to create a clustering into a predefined number of clusters k, optimizing a specific objective function to form these k clusters. One well-known partitional clustering method is the k-means algorithm (Krishna and Murty 1999). It begins by initializing k cluster centers, either arbitrarily or using methods studied in the literature (Celebi, Kingravi, and Vela 2013). In each iteration, data instances in $\mathcal{X}$ are assigned to the nearest cluster based on the distance to the center. New cluster centers are then recalculated based on the assigned points. This iterative process continues until convergence, resulting in a clustering into k clusters.

Qualitative Spatial and Temporal Formalisms

Spatial and temporal formalisms are designed to represent and reason about entities like points, lines, intervals, and regions. In this context, let $\mathcal{B}$ be a set of *base relations* describing relationships on a perceived domain of entities $\mathcal{O}$. $\mathcal{B}$ forms a partition of $\mathcal{O} \times \mathcal{O}$, includes the identity relation $\mathsf{id}_{\mathcal{B}} = \{(x, x) \mid x \in \mathcal{O}\}$, and is closed under the converse relation $^{-1}$, where $b^{-1} = \{(y, x) \mid (x, y) \in b\}$ for all $b \in \mathcal{B}$. One base relation from set $\mathcal{B}$ can represent definite knowledge between any two entities, while indefinite knowledge can be described by the union of possible base relations between the entities. This union of base relations is referred to a *qualitative relation* or simply a *relation*. It will be represented by the set of its included base relations. Thus, the set $2^{\mathcal{B}}$ represents the relations in the qualitative formalism based on $\mathcal{B}$. Within the relations of $2^{\mathcal{B}}$, the particular relation $\mathcal{B}$ is known as the *universal relation* representing the absence of any information between two entities. This relation is always satisfied for any pair of entities. On the other hand, the relation $\{\}$ is called the *empty relation* and is never satisfied. The set of relations $2^{\mathcal{B}}$ is equipped with the usual set-theoretic operations (intersection and union). Moreover, the weak composition is defined by: for all $r, r' \in 2^{\mathcal{B}}$, $r \diamond r' = \bigcup_{b \in r, b' \in r'} \{b \diamond b'\}$ (Renz and Ligozat 2005) where $b \diamond b' = \{b'' \in \mathcal{B} \mid \exists x, y, z \in \mathcal{O}$ with $(x, y) \in b, (y, z) \in b'$, and $(x, z) \in b''\}$.

For instance, Interval Algebra (`IA`) (Allen 1983) is a formalism for temporal reasoning using binary relations between time intervals: $\{\mathsf{eq}, \mathsf{p}, \mathsf{pi}, \mathsf{m}, \mathsf{mi}, \mathsf{o}, \mathsf{oi}, \mathsf{s}, \mathsf{si}, \mathsf{d}, \mathsf{di}, \mathsf{f}, \mathsf{fi}\}$ (see Figure 1), while Point Algebra (`PA`) involves binary relations between points on a line $(<, =, >)$.

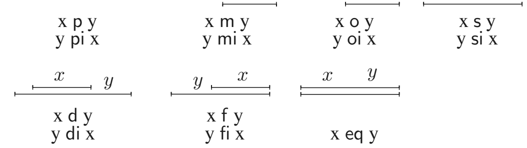

Figure 1: The thirteen base relations of `IA`.

The Region Connection Calculi (`RCC`) (Randell, Cui, and Cohn 1992) are widely used for spatial qualitative reasoning. It is based on binary topological relations between regions. Two well-known formalisms are the `RCC-5` (see Figure 2) and `RCC-8` (Bennett 1994).

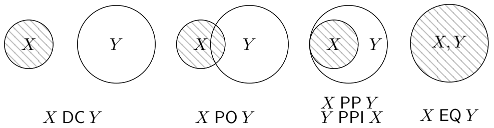

Figure 2: The five base relations of `RCC-5`.

Qualitative information about a set of entities can be modeled as a *Qualitative Relation Network (QRN)*.

A QRN is an ordered pair $\mathcal{N} = (V, R)$ where V is a set of (spatial or temporal) entities and R is a function associating a relation in $2^{\mathcal{B}}$ to each element (i, j) of $V \times V$ s.t. $R(i, i) = \mathsf{id}_{\mathcal{B}}$ and $R(j, i) = (R(i, j))^{-1}$.

A scenario is a QRN where $R(i, j)$ represents exactly one base relation for each (i, j) of $V \times V$.

Given $\mathcal{N}_1 = (V, R_1)$ and $\mathcal{N}_2 = (V, R_2)$, we define $\mathcal{N}_1 \cap \mathcal{N}_2$ as the QRN (V, R) where $R(i, j) = R_1(i, j) \cap R_2(i, j)$ for each (i, j) of $V \times V$.

A QRN $\mathcal{N} = (V, R)$ is *consistent* if there exists a solution s, i.e. a function from V to the domain $\mathcal{O}$ such that, for every $(i, j) \in V \times V$, $(s(i), s(j)) \in b$ for some $b \in R(i, j)$. Hence, the consistency problem is equivalent to the problem of determining whether a QRN has a consistent scenario. Solving the consistency problem is NP-complete in general, whereas is polynomial for the Point Algebra formalism (Van Beek 1992).

A *qualitative database (q-database)* is defined as a finite set of QRNs.

Figure 3 is an illustration of a qualitative database $\mathcal{D}$ of 4 QRNs describing 4 spatial entities using the `RCC-5` formalism. For the sake of simplicity, $id_{\mathcal{B}}$ loops $(R(i, i))$, converse relations and universal relations are omitted, and a node i is omitted iff for every $j \in V \setminus \{i\}$, $R(i, j) = \mathcal{B}$.

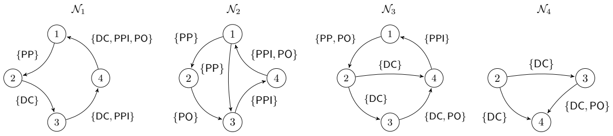

Figure 3: Example of a qualitative database.

QRN Clustering

Within a q-database $\mathcal{D}$, our main goal is to identify clusters based on shared solutions.

Definition 1 (Consistent Partition)**.** *Given a q-database $\mathcal{D}$ and a positive integer k, a consistent k-partition of $\mathcal{D}$ is a partition $\mathcal{P}$ of $\mathcal{D}$ s.t. $|\mathcal{P}| \leq k$ and $\bigcap_{\mathcal{N} \in A} \mathcal{N}$ is consistent for each $A \in \mathcal{P}$.*

It is important to emphasize that the presence of an inconsistent QRN in $\mathcal{D}$ consequently leads to the absence of any consistent partition of $\mathcal{D}$.

Theorem 2. *Let $\mathcal{Q}$ be a QSTR formalism. If the consistency problem in $\mathcal{Q}$ is NP-complete, then the problem of determining whether a q-database admits a consistent k-partition is NP-complete.*

Proof. Clearly, if q-database $\mathcal{D}$ admits a consistent partition with an arbitrary size, then it admits a consistent k-partition with $k \leq |\mathcal{D}|$. A proof that a q-database $\mathcal{D}$ admits a k-partition can be a set of the form $\{(A_1, \sigma_1), \ldots, (A_l, \sigma_l)\}$ where $l \leq min(k, |\mathcal{D}|)$,$\{A_1, \ldots, A_l\}$ is a partition of $\mathcal{D}$, and for every $1 \leq i \leq l$, σ_i is a solution of $\bigcap_{\mathcal{N} \in A_i} \mathcal{N}$, which is verifiable in polynomial time (the consistency problem is in NP). Hence, the positive instances of the consistent partition problem admit proofs that can be verified in polynomial time. Consequently, the problem of determining whether a q-database admits a consistent k-partition is in NP.

To show NP-hardness, we only need to observe that a QRN $\mathcal{N}$ is consistent iff the q-database $\{\mathcal{N}\}$ admits a consistent 1-partition. □

Definition 3 (QRN Clustering)**.** *Given a q-database $\mathcal{D}$ and a positive integer k, a k-clustering of $\mathcal{D}$ is a set $\mathcal{C} \subset 2^{\mathcal{D}}$ where:*

- EXHAUSTIVENESS. $\bigcup_{A \in \mathcal{C}} = \mathcal{D}$
- BOUNDED-SIZE. $|C| \leq k$
- MUTUAL-EXCLUSIVITY. *for every $A, A' \in \mathcal{C}$ if $A \neq A'$ then $A \cap A' = \emptyset$*
- CONSISTENCY. *if $\mathcal{D}$ admits a consistent k-partition, then for every $A \in \mathcal{C}$, $\bigcap_{\mathcal{N} \in A} \mathcal{N}$ is consistent.*

The Exhaustiveness postulate expresses that every QRN within the database $\mathcal{D}$ should manifest within the clustering. The Bounded-Size postulate expresses that the number of clusters must not surpass k. The Mutual-Exclusivity postulate declares the absence of any two clusters sharing the same QRN. Finally, the Consistency postulates affirm that each cluster maintains consistency, i.e. the QRN defined as the intersection of the QRNs of the cluster is consistent.

The subsequent proposition is a direct consequence of the definitions of consistent partition and QRN clustering.

Proposition 4. *If $\mathcal{C}$ is a consistent k-partition of a q-database $\mathcal{D}$, then $\mathcal{C}$ is k-clustering of $\mathcal{D}$.*

Consistent Partition Computation

The Boolean Satisfiability Problem (SAT) consists in determining whether a CNF formula is satisfiable, i.e. there exists an assignment of truth values to the Boolean variables of the formula that satisfies the formula. A propositional formula in Conjunctive Normal Form (CNF) is a conjunction of clauses where a clause is a disjunction of literals, and a literal is either a variable or the negation of a variable.

We present in this section the SAT encoding of the problem seeking to determining whether a q-database admits a consistent k-clustering. We emphasize that the use of the *path-consistency* (PC) method on the QRNs of the q-database, consisting in the calculation of the closure by weak composition can reduce the size of the QRNs by removing basic relations that do not satisfy PC, which can be an improvement when aiming to achieve consistent clusters.

Let $\mathcal{D} = \{\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_n\}$ be a q-database and k a positive integer.

To define our encoding for the instance $(\mathcal{D}, k)$, we associate a propositional variables x_{ij}^b with each base relation $b \in \mathcal{B}$ and each ordered pair of variables (i, j) such that $i < j$. Moreover, for every QRN $\mathcal{N}_\alpha$, we consider k additional variables $y_\alpha^1, \ldots, y_\alpha^k$. Intuitively, if y_α^β is set to true, then $\mathcal{N}_\alpha$ belongs to the β^{th} part. This encoding is an adaptation of the one given in (Pham, Thornton, and Sattar 2008).

For each $\alpha \in \{1, \ldots, n\}$ and each $\beta \in \{1, \ldots, k\}$, we use Φ_α^β to denote the conjunction of the following two formula:

$$\bigwedge_{\substack{i,j \in V \\ i<j}} \bigvee_{b \in R_\alpha(i,j)} x_{ij}^b \tag{1}$$

$$\bigwedge_{\substack{b \in R_\alpha(i,j) \\ b' \in R_\alpha(j,k)}} (x_{ij}^b \wedge x_{jk}^{b'} \rightarrow \bigvee_{b'' \in R_\alpha(i,k) \cap (b \diamond b')} x_{ij}^{b''}) \tag{2}$$

$$\text{for every } i, j, k \in V \text{ with } i < j < k$$

It is wroth noting that Φ_α^β is satisfiable iff $\mathcal{N}_\alpha$ is consistent in the β^{th} part.

The first formula of our encoding states that if y_α^β is true, the QRN $\mathcal{N}_\alpha$ belongs to the β^{th} consistent part:

$$y_\alpha^\beta \rightarrow \Phi_\alpha^\beta \tag{3}$$

$$\text{for every } \alpha \in \{1, \ldots, n\} \text{ and } \beta \in \{1, \ldots, k\}$$

The second formula ensures that every QRN in the q-database belongs to at least one consistent part:

$$\bigwedge_{1 \leq \alpha \leq n} (\bigvee_{1 \leq \beta \leq k} y_\alpha^\beta) \tag{4}$$

We use $\mathsf{ConsPart}(\mathcal{D}, k)$ to denote the encoding that consists of the conjunction of (3) and (4).

Using Formula (3), we have for every $\alpha \in \{1, \ldots, n\}$ and every $\beta \in \{1, \ldots, k\}$, if y_α^β is true then Φ_α^β is also true, and consequently the QRN $\mathcal{N}_\alpha$ is in the β^{th} consistent part. Formula (4) implies that every QRN in $\mathcal{D}$ belongs to at least one consistent part, thus, $\mathcal{D}$ admits a consistent k-partition.

Measuring the distance between two QRNs

In this section, we introduce distance measures designed specifically for QRNs, with the intention of integrating them into established clustering algorithms. Notably, these measures are inspired by the inconsistency measures introduced in (Condotta, Raddaoui, and Salhi 2016).

The subsequent measures depend on solving the Partial MaxSAT problem, a variation of the SAT problem involving two sets of clauses, Σ_1 and Σ_2. The objective is to find the largest subset of *soft clauses* (Σ_1) that can be satisfied along with all the *hard clauses* (Σ_2), where soft clauses are desirable but not mandatory, and hard clauses are clauses that must be satisfied.

Definition 5 (R-Relaxation)**.** *Let $\mathcal{N} = (V, R)$ be a QRN. A R-relaxation of $\mathcal{N}$ is a QRN $\mathcal{N}' = (V', R')$ s.t. $V' = V$ and $\mathcal{N} \subseteq \mathcal{N}'$. A R-relaxation is said to be consistent if it corresponds to a consistent QRN.*

We use $\mathsf{ConsRR}(\mathcal{N})$ to denote the set of consistent R-relaxations of $\mathcal{N}$. Moreover, given a R-relaxation $\mathcal{N}'$, we use $\mathsf{diff}(\mathcal{N},\mathcal{N}')$ to refer to the set of ordered pairs of variables $\{(i,j) \in V \times V : i < j, \mathcal{N}[i,j] \neq \mathcal{N}'[i,j]\}$. It follows from definition 5 the following distance measure:

- $D_{rr}(\mathcal{N},\mathcal{N}') = min\{|\mathsf{diff}(\mathcal{N} \cap \mathcal{N}',\mathcal{N}'')| : \mathcal{N}'' \in \mathsf{ConsRR}(\mathcal{N} \cap \mathcal{N}')\}$

The measure D_{rr} counts the minimum number of constraints that have to be changed to recover consistency.

We now introduce a Partial MaxSAT encoding for computing $D_{cr}(\mathcal{N},\mathcal{N}')$. Let us fix $\mathcal{N} \cap \mathcal{N}' = (V, R)$. Similarly to $\mathsf{ConsPart}(\mathcal{D}, k)$, we associate a propositional variable x^b_{ij} with each base relation $b \in \mathcal{B}$ and each ordered pair of variables $(i,j) \in V \times V$ such that $i < j$. Additionally, we associate a distinct propositional variable r_{ij} with each $(i,j) \in V \times V$ such that $i < j$; these variables are used to measure the distance by capturing the elements in $\mathsf{diff}(\mathcal{N},\mathcal{N}')$.

The hard part of our encoding is defined as the conjunction of the following formulas:

$$\bigwedge_{\substack{i,j \in V\\ i<j}} \bigvee_{b \in \mathcal{B}} x^b_{ij} \tag{5}$$

$$\bigwedge_{b,b' \in \mathcal{B}} (x^b_{ij} \wedge x^{b'}_{jk} \rightarrow \bigvee_{b'' \in b \diamond b'} x^{b''}_{ij}) \tag{6}$$

for every $i,j,k \in V$ with $i < j < k$

$$\bigwedge_{\substack{i,j \in V\\ i<j}} (r_{ij} \rightarrow \bigvee_{b \in R(i,j)} x^b_{ij}) \tag{7}$$

The soft part of our encoding consists simply of the following set of unit clauses $S = \{r_{ij} : i,j \in V \text{ and } i < j\}$. Indeed, by maximizing the number of true element in S, we maximize the number of constraints that share the same base relations with $\mathcal{N} \cap \mathcal{N}'$.

Let us now introduce our second distance measure which is based on the notion of variable relaxation.

Definition 6 (V-Relaxation)**.** *Let $\mathcal{N} = (V, R)$ be a QRN. A V-relaxation of $\mathcal{N}$ is a QRN $\mathcal{N}' = (V', R')$ s.t. $V' \subseteq V$ and $R(i,j) = R'(i,j)$ for every $i,j \in V$. We use $\mathcal{N}'_{\downarrow V'}$ to denote the V-relaxation $\mathcal{N}'$. A V-relaxation is said to be consistent if it is a consistent QRN.*

The set of consistent V-relaxations is denoted by $\mathsf{ConsVR}(\mathcal{N})$.

The distance measure obtained from the notion of V-relaxation is defined as follows:

- $D_{vr}(\mathcal{N},\mathcal{N}') = min\{|V \setminus V'| : \mathcal{N}_{\downarrow V'} \in \mathsf{ConsVR}(\mathcal{N} \cap \mathcal{N}')\}$

This measure counts the minimum number of variables that have to be removed to get consistency.

Let us describe our Partial MaxSAT encoding for computing $D_{vr}(\mathcal{N},\mathcal{N}')$. Let $\mathcal{N} \cap \mathcal{N}' = (V, R)$. To define our encoding, we use variables of the form $v_{i \in V}$ in addition to the variables of the form x^b_{ij}. The variable v_i is used to know whether the variable i is ignored or not.

The hard part is defined as the conjunction of the following formulas:

$$\bigwedge_{\substack{i,j \in V\\ i<j}} \bigvee_{b \in R(i,j)} x^b_{ij} \tag{8}$$

$$\bigwedge_{b,b' \in \mathcal{B}} (x^b_{ij} \wedge x^{b'}_{jk} \wedge v_i \wedge v_j \wedge v_k \rightarrow \bigvee_{b'' \in R(i,j) \cap (b \diamond b')} x^{b''}_{ij}) \tag{9}$$

for every $i,j,k \in V$ with $i < j < k$

The soft part consists of the set of unit clauses $S = \{v_i : i \in V\}$, and by maximizing the true elements of S, we minimise the number of ignored variables.

Adaptation of the k-means Algorithm

We adapt the k-means algorithm by considering the distance measures introduced in the previous section as measures for similarity between QRNs, and by defining the center of a each cluster $\mathcal{P}$ as the QRN $\mathcal{P}^{\odot} = \bigcap_{\mathcal{N} \in \mathcal{P}} \mathcal{N}$. The objective is to minimize the inconsistency of the overall clustering. To achieve this, we define the objective function to be minimized as $\max_{1 \leq l \leq k} inc_{dist}(\mathcal{P}^{\odot}_l)$ where inc_{dist} measures the inconsistency of $\mathcal{P}^{\odot}_l$ using D_{vr} (resp. D_{rr}) as the ratio between the minimum number of edges (resp. edges) that must be removed to make $\mathcal{P}^{\odot}_l$ consistent.

The k-means requires an initialization phase of the k centers. In our context, Algorithm 1 performs this task with the idea of considering the $k-1$ most dispersed QRNs from a QRN $\mathcal{N}_i$ called the *QRN initialization seed*. The QRN initialization seed is generated randomly and considered as the first center $\mathcal{P}^{\odot}_1$, and we proceed iteratively to consider the QRN $\mathcal{N} \in \mathcal{D} \setminus \{\mathcal{P}^{\odot}_1, \mathcal{P}^{\odot}_2, \ldots, \mathcal{P}^{\odot}_{l-1}\}$ as the center of the cluster $\mathcal{P}_l$ if $\mathcal{N}$ is the most distant QRN from $\cap^{l-1}_{h=1} \mathcal{P}^{\odot}_h$ with respect to the chosen distance measure being D_{rr} or D_{vr}.

Algorithm 1: Initialization Algorithm

```
Data: A set $\mathcal{D} = \{\mathcal{N}_1, \ldots, \mathcal{N}_n\}$ of QRNs, number of
      clusters $k$
Result: A set of centers $\{\mathcal{P}^{\odot}_1, \mathcal{P}^{\odot}_2, \ldots, \mathcal{P}^{\odot}_k\}$
$\mathcal{P}^{\odot}_1 \leftarrow randomchoice(\mathcal{D})$
for $l \leftarrow 2$ to $k$ do
    $\mathcal{P}^{\odot}_l \leftarrow arg \max_{\mathcal{N} \in \mathcal{D}} dist(\mathcal{N}, \cap^{l-1}_{h=1} \mathcal{P}^{\odot}_h)$
end
return $\{\mathcal{P}^{\odot}_1, \mathcal{P}^{\odot}_2, \ldots, \mathcal{P}^{\odot}_k\}$
```

Algorithm 2 initiates with k centers representing the initial clusters. The completion phase sequentially assigns each QRN $\mathcal{N}$ not among the initial centers to its nearest cluster, determined by the distance measure (D_{rr} or D_{vr}) between the cluster center and $\mathcal{N}$. After all QRNs in $\mathcal{D}$ are assigned, cluster centers are updated, and the objective function is recalculated. The algorithm then enters a rectification phase, computing distances between each QRN and cluster centers to identify potential moves. Moves are validated with a specific rule: if all QRNs in a cluster are set to change cluster, only the most distant QRN can move, preventing clusters from becoming *black holes* which is arising from QRN reassignments not considering concurrent moves and centers updates occurring after all moves.

The process iterates until the objective function reaches 0 or no further moves can be made, indicating convergence.

```
Algorithm 2: Clustering Algorithm
Data: A set $\mathcal{D} = \{\mathcal{N}_1, \dots, \mathcal{N}_n\}$ of QRNs, number of
      clusters $k$
Result: A set of clusters $\{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_k\}$
$\{\mathcal{P}_1^{\odot}, \mathcal{P}_2^{\odot}, \dots, \mathcal{P}_k^{\odot}\} = Initialization(\mathcal{D}, k)$
for $\mathcal{N} \in \mathcal{D} \setminus \{\mathcal{P}_1^{\odot}, \mathcal{P}_2^{\odot}, \dots, \mathcal{P}_k^{\odot}\}$ do
    $j \leftarrow arg \min_{1 \le l \le k} dist(\mathcal{N}, \mathcal{P}_l^{\odot})$
    $\mathcal{P}_j \leftarrow \mathcal{P}_j \cup \{\mathcal{N}\}$
end
$updateCenters(\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_k)$
while $\max_{1 \le l \le k} inc_{dist}(\mathcal{P}_l^{\odot}) > 0$ do
    $moves = \emptyset$
    for $(\mathcal{N} \in \mathcal{D})$ do
        $j \leftarrow arg \min_{1 \le l \le k} dist(\mathcal{N}, \mathcal{P}_l^{\odot})$
        $moves \cup \{(\mathcal{N}, j)\}$
    end
    if $moves = \emptyset$ then
        break
    end
    for $move = (\mathcal{N}, j) \in areValid(moves)$ do
        Move $\mathcal{N}$ to cluster $\mathcal{P}_j$
    end
    $updateCenters(\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_k)$
end
return $\{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_k\}$
```

Experiments

We conducted experiments with 336 instances, each consisting of 10 QRNs from either RCC−5 or RCC−8. The choice of the two formalisms is just to illustrate our framework. Two sets of instances were considered: one with only consistent QRNs (*COH-instances*) and (*INC-instances*) were inconsistent QRNs are present. These QRNs were generated using the $A(m, d, l)$ model from (Nebel and Renz 2011). The model produces QRNs with m nodes, each with an average degree of d and an average relation size of l on selected edges. We vary m from 7 to 10, while d ranges between $\frac{|m|}{2} - 1$ and $\frac{|m|}{2}$ in increments of 0.1, and l remains fixed at $|\mathcal{B}|/2$ for each formalism. We explore different values of k ranging from 2 to 6. Experiments were performed on a computer with an Intel i9-Core 2.80 GHz processor, 64 GB of RAM, and coded in Python 3.9.16.

To evaluate the results obtained using the HAC and the k-means algorithms, we utilized two metrics to assess the intra-cluster resemblance of the QRNs within the same cluster, and, two metrics to evaluate the inter-cluster dissimilarity. The first metric mi_1, indicates the average number of shared base relations among edges in QRNs within a cluster. A higher mi_1 value implies a greater symbolic similarity among QRNs in the cluster.

```
Algorithm 3: HAC Algorithm
Data: A set $\mathcal{D} = \{\mathcal{N}_1, \dots, \mathcal{N}_n\}$ of $n$ consistent QRNs
Result: A set $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}$ where $\mathcal{C}_k$ is a $k$-clustering
        of $\mathcal{D}$
$\mathcal{C}_n = \{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_n\}$ where $\mathcal{P}_j = \{\mathcal{N}_j\}$
$k \leftarrow n$
while $k > 1$ do
    $\mathcal{P}_i, \mathcal{P}_j = arg \min_{\substack{\mathcal{P}_h, \mathcal{P}_l \in \mathcal{C}_k \\ h \neq l}} dist(\mathcal{P}_h^{\odot}, \mathcal{P}_l^{\odot})$
    $\mathcal{C}_{k-1} \leftarrow (\mathcal{C}_k \setminus \{\mathcal{P}_i, \mathcal{P}_j\}) \cup \{\mathcal{P}_i \cup \mathcal{P}_j\}$
    $k \leftarrow k - 1$
end
return $\{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_n\}$
```

$$mi_1(\mathcal{P}) = \frac{2}{m(m-1)} \sum_{\substack{i,j \in V \\ i<j}} \frac{|R_{\mathcal{P}}(i,j)|}{|\bigcup_{\mathcal{N} \in \mathcal{P}} R(i,j)|}$$

Thus, we evaluate a k-clustering $\mathcal{C}$ through $intra_1(\mathcal{C}) = \min_{\mathcal{P} \in \mathcal{C}} mi_1(\mathcal{P})$. In Figure 4, both the k-means and HAC algorithms, utilizing distances D_{vr} and D_{rr}, yield k-clustering where the QRNs of the least similar cluster exhibit a symbolic similarity of at least 30% on average, as indicated by mean values as black triangles. Notably, the k-means algorithm, especially with the D_{rr} distance, generally outperforms. The second metric, denoted as mi_2, calculates the average number of consistent triplets shared among the triplets of the QRNs of each cluster. We use $T_{\mathcal{N}}(i,j,k)$ to refer to the set of consistent triplets of base relations of $\mathcal{N}$ defined as $\{(b, b', b'') \in R(i,j) \times R(j,k) \times R(i,k) : b'' \in b \diamond b'\}$. Thus,

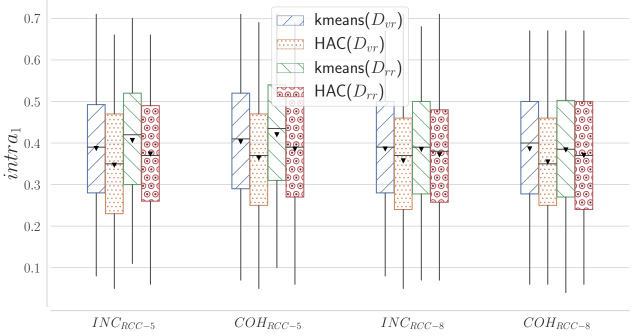

Figure 4: Obtained values of $intra_1$.

$$mi_2(\mathcal{P}) = \frac{1}{\binom{m}{3}} \sum_{\substack{i,j,k \in V \\ i<j<k}} \frac{|T_{\mathcal{P}}(i,j,k)|}{|\bigcup_{\mathcal{N} \in \mathcal{P}} T_{\mathcal{N}}(i,j,k)|}$$

This metric helps gauge the level of consistency in the triplets of the QRNs within each cluster. We evaluate a k-clustering $\mathcal{C}$ through $intra_2(\mathcal{C}) = \min_{\mathcal{P} \in \mathcal{C}} mi_2(\mathcal{P})$.

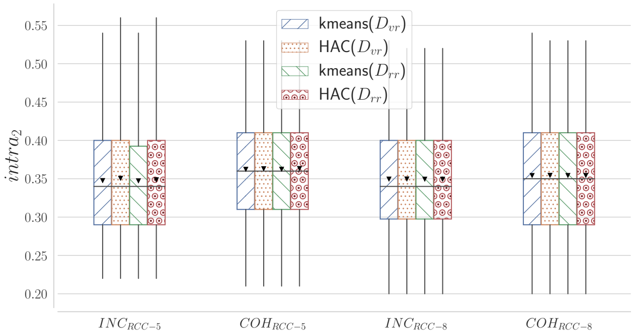

Figure 5: Obtained values of $intra_2$.

In Figure 5, both methods, using distances D_{vr} and D_{rr}, yield similar results. The least similar clusters show a minimum of 30% consistent triplets on average.

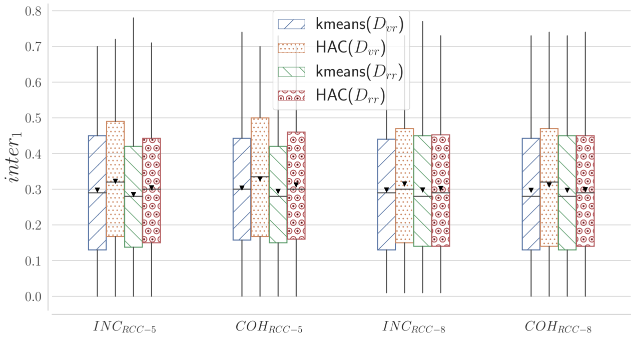

Figure 6: Obtained values of $inter_1$.

The third and fourth metrics, me_1 and me_2, calculate the average number of shared base relations among edges and consistent triplets of base relations among triplets of two cluster centers:

$$me_1(\mathcal{P}_\varphi, \mathcal{P}_\vartheta) = \frac{2}{m(m-1)|\mathcal{B}|} \sum_{\substack{i,j \in V \\ i<j}} R_{\mathcal{P}_\varphi \cap \mathcal{P}_\vartheta}(i,j)$$

$$me_2(\mathcal{P}_\varphi, \mathcal{P}_\vartheta) = \frac{1}{\binom{m}{3}|\mathcal{T}|} \sum_{\substack{i,j \in V \\ i<j}} T_{\mathcal{P}_\varphi \cap \mathcal{P}_\vartheta}(i,j,k)$$

where $\mathcal{T} = \{(b, b', b'') \in \mathcal{B}^3 : b'' \in b \diamond b'\}$. A lower value of me_1 respectively me_3 signifies a greater symbolic dissimilarity between the two clusters respectively a lower number of shared consistent triplets, indicating a reduced likelihood that the two clusters exhibit a form of consistency merged together. These two metrics, me_1 and me_3, offer valuable insights for assessing the divergence of clusters within a k-clustering. Thus, other evaluations of a k-clustering $\mathcal{C}$ are defined as $inter_1(\mathcal{C}) = \min_{\substack{(\mathcal{P}_\varphi, \mathcal{P}_\vartheta) \in \mathcal{C} \times \mathcal{C} \\ \varphi \neq \vartheta}} me_1(\mathcal{P}_\varphi, \mathcal{P}_\vartheta)$ and $inter_2(\mathcal{C}) = \min_{\substack{(\mathcal{P}_\varphi, \mathcal{P}_\vartheta) \in \mathcal{C} \times \mathcal{C} \\ \varphi \neq \vartheta}} me_2(\mathcal{P}_\varphi, \mathcal{P}_\vartheta)$.

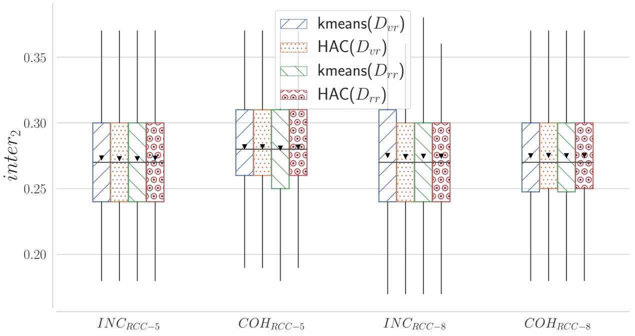

Figure 7: Obtained values of $inter_2$.

Figures 6 and 7 reveal that, the two most similar clusters in each k-clustering exhibit a maximum symbolic similarity and consistent triplets shared copped to 30% on average.

In addition to the mentioned metrics, we consider the minimum of consistency of $\mathcal{C}$, defined as $m_{coh}(\mathcal{C}) = 1 - (\max_{\mathcal{P} \in \mathcal{C}} inc_{D_{vr}}(\mathcal{P}^\odot) \times \max_{\mathcal{P} \in \mathcal{C}} inc_{D_{rr}}(\mathcal{P}^\odot))^{\frac{1}{2}}$. This additional metric provide a more comprehensive evaluation of the consistency of the results.

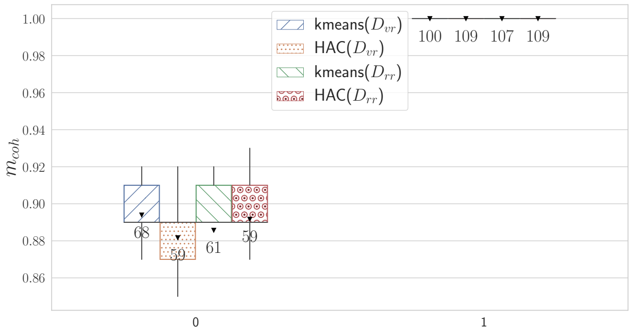

Figure 8: Values of m_{coh} on instances where consistent partitions are feasible.

In Figure 8, all methods identified consistent k-clusterings for at least 100 out of 168 instances with possible consistent partitions. Additionally, when k-clusterings are inconsistent, the average minimum consistency value is $m_{coh} = 0.88$ for the least consistent cluster.

Conclusion

We have presented a formal framework for QSTR clustering, introducing a SAT encoding to verify the feasibility of consistent partitions. We also proposed QRN-specific distance measures to calculate similarity, and conducted experiments using adapted k-means and HAC algorithms using these measures. In future work, we aim to explore other objective functions such as maximizing the number of consistent clusters, develop additional measures to assess QRN similarity and evaluate the quality of a k-clustering, and enhance our k-means adaptation algorithm by leveraging intuitive rules for QRN relocation.

References

Aggarwal, C. C., and Reddy, C. K. 2013. Data clustering: Algorithms and applications, ser. *Data Mining and Knowledge Discovery Series. CRC Press/Taylor & Francis Group*.

Aggarwal, C. C., and Zhai, C. 2012. A survey of text clustering algorithms. *Mining text data* 77–128.

Allen, J. F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11):832–843.

Bennett, B. 1994. Spatial reasoning with propositional logics. In *Principles of Knowledge Representation and Reasoning*, 51–62. Elsevier.

Boudane, A.; Jabbour, S.; Sais, L.; and Salhi, Y. 2017. Clustering complex data represented as propositional formulas. In *Advances in Knowledge Discovery and Data Mining: 21st Pacific-Asia Conference, PAKDD 2017, Jeju, South Korea, May 23-26, 2017, Proceedings, Part II 21*, 441–452. Springer.

Boukontar, A.; Condotta, J.-F.; and Salhi, Y. 2022. Knowledge discovery from qualitative spatial and temporal data. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, 451–458. IEEE.

Celebi, M. E.; Kingravi, H. A.; and Vela, P. A. 2013. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications* 40(1):200–210.

Celebi, M. E. 2014. *Partitional clustering algorithms*. Springer.

Chang, J.; Wang, L.; Meng, G.; Xiang, S.; and Pan, C. 2017. Deep adaptive image clustering. In *Proceedings of the IEEE international conference on computer vision*, 5879–5887.

Condotta, J.-F.; Raddaoui, B.; and Salhi, Y. 2016. Quantifying conflicts for spatial and temporal information. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

de Carvalho, F. d. A.; Csernel, M.; and Lechevallier, Y. 2009. Clustering constrained symbolic data. *Pattern Recognition Letters* 30(11):1037–1045.

Ferdous, R., et al. 2009. An efficient k-means algorithm integrated with jaccard distance measure for document clustering. In *2009 first asian himalayas international conference on internet*, 1–6. IEEE.

Goyal, M., and Aggarwal, S. 2017. A review on k-mode clustering algorithm. *International Journal of Advanced Research in Computer Science* 8(7).

Homem, T. P. D.; Santos, P. E.; Costa, A. H. R.; da Costa Bianchi, R. A.; and de Mantaras, R. L. 2020. Qualitative case-based reasoning and learning. *Artificial Intelligence* 283:103258.

Irani, J.; Pise, N.; and Phatak, M. 2016. Clustering techniques and the similarity measures used in clustering: A survey. *International journal of computer applications* 134(7):9–14.

Jarman, A. M. 2020. Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. *Georgia Southern University* 29.

Kaufman, L., and Rousseeuw, P. J. 2009. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kejžar, N.; Korenjak-Černe, S.; and Batagelj, V. 2021. Clustering of modal-valued symbolic data. *Advances in Data Analysis and Classification* 15(2):513–541.

Krishna, K., and Murty, M. N. 1999. Genetic k-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 29(3):433–439.

Murtagh, F., and Contreras, P. 2012. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(1):86–97.

Nebel, B., and Renz, J. 2011. Efficient methods for qualitative spatial reasoning. *CoRR* abs/1106.0679.

Nielsen, F. 2016. Hierarchical clustering. *Introduction to HPC with MPI for Data Science* 195–211.

Pham, D. N.; Thornton, J.; and Sattar, A. 2008. Modelling and solving temporal reasoning as propositional satisfiability. *Artificial Intelligence* 172(15):1752–1782.

Randell, D. A.; Cui, Z.; and Cohn, A. G. 1992. A spatial logic based on regions and connection. *KR* 92:165–176.

Renz, J., and Ligozat, G. 2005. Weak composition for qualitative spatial and temporal reasoning. In *International Conference on Principles and Practice of Constraint Programming*, 534–548. Springer.

Salhi, Y. 2019. Qualitative reasoning and data mining. In *26th International Symposium on Temporal Representation and Reasoning (TIME 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Schaeffer, S. E. 2007. Graph clustering. *Computer science review* 1(1):27–64.

Sinwar, D., and Kaushik, R. 2014. Study of euclidean and manhattan distance metrics using simple k-means clustering. *Int. J. Res. Appl. Sci. Eng. Technol* 2(5):270–274.

Van Beek, P. 1992. Reasoning about qualitative temporal information. *Artificial intelligence* 58(1-3):297–326.

Vilain, M.; Kautz, H.; and Van Beek, P. 1990. Constraint propagation algorithms for temporal reasoning: A revised report. In *Readings in qualitative reasoning about physical systems*. Elsevier. 373–381.

Wang, Y.; Qiao, M.; Liu, H.; and Ye, X. 2018. Qualitative spatial reasoning on topological relations by combining the semantic web and constraint satisfaction. *Geo-spatial information science* 21(2):80–92.

Yim, O., and Ramdeen, K. T. 2015. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The quantitative methods for psychology* 11(1):8–21.