Workshop on belief functions

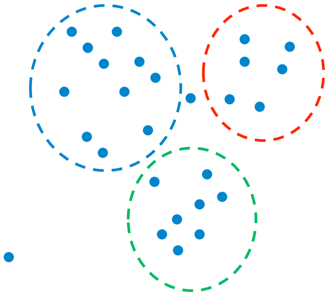
Clustering

Thierry Denœux

Université de Technologie de Compiègne, France
HEUDIASYC (UMR CNRS 7253)
https://www.hds.utc.fr/~tdenoeux

Chiang Mai University
July-August 2017

Clustering



- n objects described by
  - Attribute vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (attribute data) or
  - Dissimilarities (proximity data)
- Goals:
  1. Discover groups in the data
  2. Assess the uncertainty in group membership

Hard and soft clustering concepts

Hard clustering: no representation of uncertainty. Each object is assigned to one and only one group. Group membership is represented by binary variables u_{ik} such that $u_{ik} = 1$ if object i belongs to group k and $u_{ik} = 0$ otherwise.

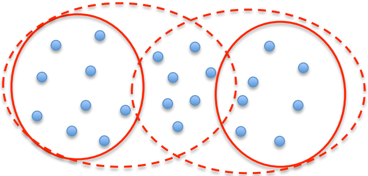
Fuzzy clustering: each object has a degree of membership $u_{ik} \in [0, 1]$ to each group, with $\sum_{k=1}^{c} u_{ik} = 1$. The u_{ik}'s can be interpreted as probabilities.

Fuzzy clustering with noise cluster: the above equality is replaced by $\sum_{k=1}^{c} u_{ik} \leq 1$. The number $1 - \sum_{k=1}^{c} u_{ik}$ is interpreted as a degree of membership (or probability of belonging to) to a noise cluster.

Hard and soft clustering concepts

Possibilistic clustering: the u_{ik} are free to take any value in $[0, 1]^c$. Each number u_{ik} is interpreted as a degree of possibility that object i belongs to group k.

Rough clustering: each cluster ω_k is characterized by a lower approximation $\underline{\omega}_k$ and an upper approximation $\overline{\omega}_k$, with $\underline{\omega}_k \subseteq \overline{\omega}_k$; the membership of object i to cluster k is described by a pair $(\underline{u}_{ik}, \overline{u}_{ik}) \in \{0, 1\}^2$, with $\underline{u}_{ik} \leq \overline{u}_{ik}$, $\sum_{k=1}^{c} \underline{u}_{ik} \leq 1$ and $\sum_{k=1}^{c} \overline{u}_{ik} \geq 1$.



Clustering and belief functions

clustering structure	uncertainty framework
fuzzy partition	probability theory
possibilistic partition	possibility theory
rough partition	(rough) sets
?	belief functions

- As belief functions extend probabilities, possibilities and sets, could the theory of belief functions provide a more general and flexible framework for cluster analysis?
- Objectives:
  - Unify the various approaches to clustering
  - Achieve a richer and more accurate representation of uncertainty
  - New clustering algorithms and new tools to compare and combine clustering results.

Outline

1. Evidential clustering
   - Credal partition
   - Summarization of a credal partition
   - Relational representation of a credal partition
2. Evidential clustering algorithms
   - Evidential c-means
   - EVCLUS
   - Ek-NNclus
3. Comparing and combining the results of soft clustering algorithms
   - The credal Rand index
   - Combining clustering structures

Outline

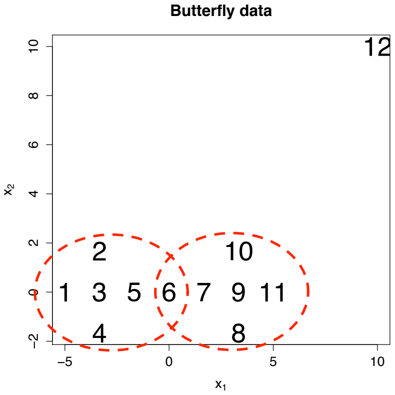
1. Evidential clustering
   - Credal partition
   - Summarization of a credal partition
   - Relational representation of a credal partition
2. Evidential clustering algorithms
   - Evidential c-means
   - EVCLUS
   - Ek-NNclus
3. Comparing and combining the results of soft clustering algorithms
   - The credal Rand index
   - Combining clustering structures

Outline

Evidential clustering

- Let $O = \{o_1, \ldots, o_n\}$ be a set of n objects and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be a set of c groups (clusters).
- Each object o_i belongs to at most one group.
- Evidence about the group membership of object o_i is represented by a mass function m_i on Ω:
  - for any nonempty set of clusters $A \subseteq \Omega$, $m_i(A)$ is the probability of knowing only that o_i belong to one of the clusters in A.
  - $m_i(\emptyset)$ is the probability of knowing that o_i does not belong to any of the c groups.
- The n-tuple $M = (m_1, \ldots, m_n)$ is called a credal partition.

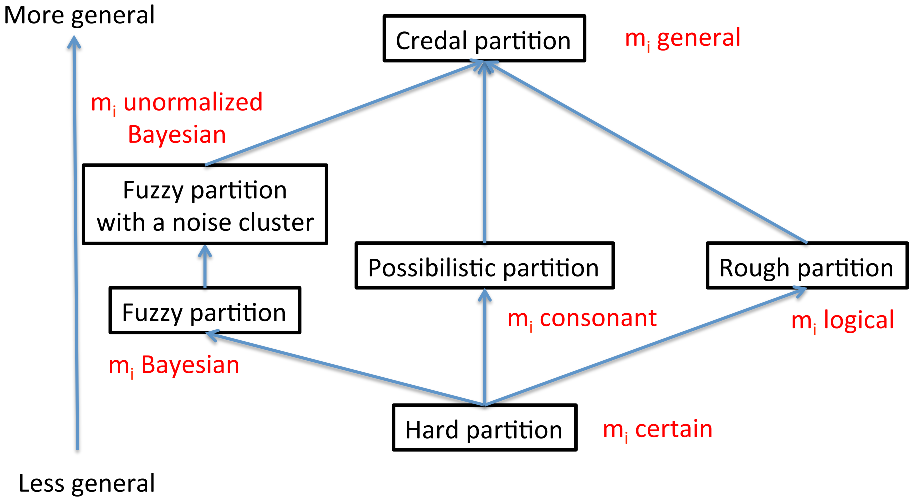
Example



Credal partition

	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_1, \omega_2\}$
m_3	0	1	0	0
m_5	0	0.5	0	0.5
m_6	0	0	0	1
m_{12}	0.9	0	0.1	0

Relationship with other clustering structures

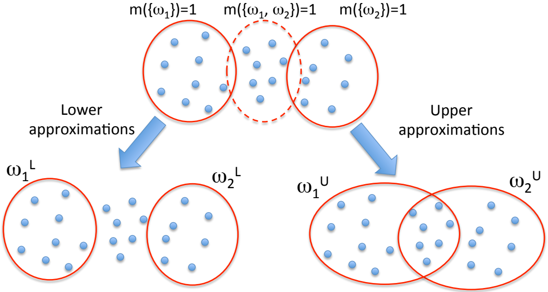


Rough clustering as a special case

- Assume that each m_i is logical, i.e., $m_i(A_i) = 1$ for some $A_i \subseteq \Omega$, $A_i \neq \emptyset$.
- We can then define the lower and upper approximations of cluster ω_k as

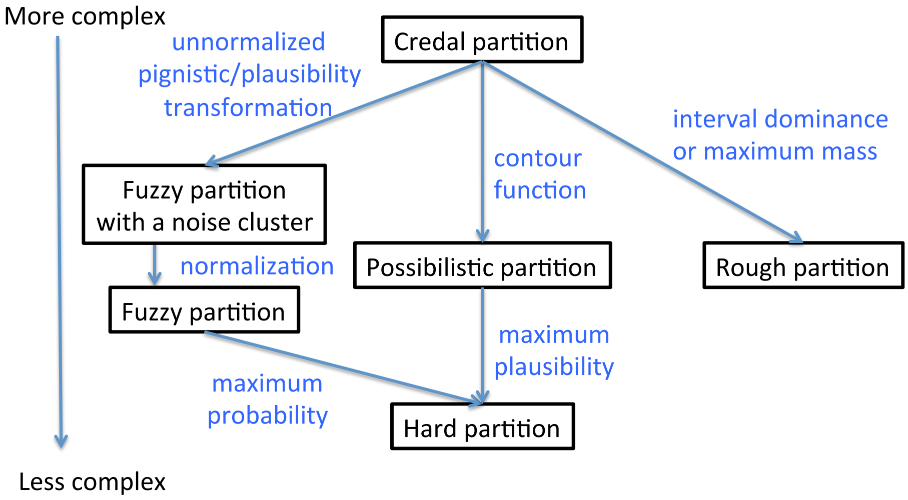
$$\underline{\omega}_k = \{o_i \in O | A_i = \{\omega_k\}\}, \quad \overline{\omega}_k = \{o_i \in O | \omega_k \in A_i\}.$$

- The membership values to the lower and upper approximations of cluster ω_k are $\underline{u}_{ik} = Bel_i(\{\omega_k\})$ and $\overline{u}_{ik} = Pl_i(\{\omega_k\})$.



Outline

Summarization of a credal partition



From evidential to rough clustering

- For each i, let $A_i \subseteq \Omega$ be the set of non dominated clusters

$$A_i = \{\omega \in \Omega | \forall \omega' \in \Omega, Bel_i^*(\{\omega'\}) \leq Pl_i^*(\{\omega\})\},$$

  where Bel_i^* and Pl_i^* are the normalized belief and plausibility functions.
- Lower approximation:

$$\underline{u}_{ik} = \begin{cases} 1 & \text{if } A_i = \{\omega_k\} \\ 0 & \text{otherwise.} \end{cases}$$

- Upper approximation:

$$\overline{u}_{ik} = \begin{cases} 1 & \text{if } \omega_k \in A_i \\ 0 & \text{otherwise.} \end{cases}$$

- The outliers can be identified separately as the objects for which $m_i(\emptyset) \geq m_i(A)$ for all $A \neq \emptyset$.

Outline

Relational representation of a hard partition

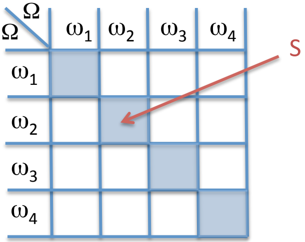
- A hard partition can be represented equivalently by
  - the $n \times c$ membership matrix $U = (u_{ik})$ or
  - an $n \times n$ relation matrix $R = (r_{ij})$ representing the equivalence relation

$$r_{ij} = \begin{cases} 1 & \text{if } o_i \text{ and } o_j \text{ belong to the same group} \\ 0 & \text{otherwise.} \end{cases}$$

- The relational representation R is invariant under renumbering of the clusters, and is thus more suitable to compare or combine several partitions.
- What is the counterpart of matrix R in the case of a credal partition?

Relational representation

- Let $M = (m_1, \ldots, m_n)$ be a credal partition.
- For a pair of objects $\{o_i, o_j\}$, let Q_{ij} be the question "Do o_i and o_j belong to the same group?" defined on the frame $\Theta = \{s, \neg s\}$.
- Θ is a coarsening of Ω^2.



Given m_i and m_j on Ω, a mass function m_{ij} on Θ can be computed as follows:

1. Extend m_i and m_j to Ω^2;
2. Combine the extensions of m_i and m_j by the unnormalized Dempster's rule;
3. Compute the restriction of the combined mass function to Θ.

Pairwise mass function

- Mass function:

$$m_{ij}(\emptyset) = m_i(\emptyset) + m_j(\emptyset) - m_i(\emptyset)m_j(\emptyset)$$
$$m_{ij}(\{s\}) = \sum_{k=1}^{c} m_i(\{\omega_k\})m_j(\{\omega_k\})$$
$$m_{ij}(\{\neg s\}) = \kappa_{ij} - m_{ij}(\emptyset)$$
$$m_{ij}(\Theta) = 1 - \kappa_{ij} - \sum_k m_i(\{\omega_k\})m_j(\{\omega_k\}).$$

where κ_{ij} is the degree of conflict between m_i and m_j.

- In particular,

$$pl_{ij}(s) = 1 - \kappa_{ij}.$$

Return to CECM

Special cases

Hard partition:

$$m_{ij}(\{s\}) = r_{ij}, \quad m_{ij}(\{\neg s\}) = 1 - r_{ij} \quad \text{with } r_{ij} \in \{0, 1\}$$

Fuzzy partition:

$$m_{ij}(\{s\}) = r_{ij}, \quad m_{ij}(\{\neg s\}) = 1 - r_{ij} \quad \text{with } r_{ij} \in [0, 1]$$

Rough partition: Assume $m_i(A_i) = 1$ and $m_j(A_j) = 1$.

$$\begin{aligned} m_{ij}(\{s\}) &= 1 \quad \text{if } A_i = A_j = \{\omega_k\} \\ m_{ij}(\{\neg s\}) &= 1 \quad \text{if } A_i \cap A_j = \emptyset \\ m_{ij}(\Theta) &= 1 \quad \text{otherwise.} \end{aligned}$$

Relational representation of a credal partition

- Let $M = (m_1, \ldots, m_n)$ be a credal partition.
- The tuple $R = (m_{ij})_{1 \leq i < j \leq n}$ is called the relational representation of credal partition M.

$$M = (m_1, m_2, m_3, m_4, m_5) \longrightarrow R = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 \\ 1 & \cdot & m_{12} & m_{13} & m_{14} & m_{15} \\ 2 & \cdot & \cdot & m_{23} & m_{24} & m_{25} \\ 3 & \cdot & \cdot & \cdot & m_{34} & m_{35} \\ 4 & \cdot & \cdot & \cdot & \cdot & m_{45} \\ 5 & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

- Open question: given a relational representation R, can we uniquely recover the credal partition M, up to a permutation of the cluster indices?

Example

- Credal partition:

A	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_1, \omega_2\}$
$m_1(A)$	0.3	0.6	0.1	0.0
$m_2(A)$	0.0	0.7	0.1	0.2
$m_3(A)$	0.0	0.1	0.6	0.3

- Relational representation:

A	$\emptyset$	$\{s\}$	$\{\neg s\}$	$\{s, \neg s\}$
$m_{12}(A)$	0.30	0.43	0.13	0.14
$m_{13}(A)$	0.30	0.12	0.37	0.21
$m_{23}(A)$	0.00	0.13	0.43	0.44

Outline

1. Evidential clustering
   - Credal partition
   - Summarization of a credal partition
   - Relational representation of a credal partition
2. Evidential clustering algorithms
   - Evidential c-means
   - EVCLUS
   - Ek-NNclus
3. Comparing and combining the results of soft clustering algorithms
   - The credal Rand index
   - Combining clustering structures

Main approaches

1. Evidential *c*-means (ECM): (Masson and Denoeux, 2008):
   - Attribute data
   - HCM, FCM family
2. EVCLUS (Denoeux and Masson, 2004; Denoeux et al., 2016):
   - Attribute or proximity (possibly non metric) data
   - Multidimensional scaling approach
3. EK-NNclus (Denoeux et al, 2015)
   - Attribute or proximity data
   - Searches for the most plausible partition of a dataset

Outline

Principle

- Problem: generate a credal partition $M = (m_1, \ldots, m_n)$ from attribute data $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, $\boldsymbol{x}_i \in \mathbb{R}^p$.
- Generalization of hard and fuzzy c-means algorithms:
  - Each cluster is represented by a prototype.
  - Cyclic coordinate descent algorithm: optimization of a cost function alternatively with respect to the prototypes and to the credal partition.

Fuzzy *c*-means (FCM)

- Minimize

$$J_{\text{FCM}}(U,V)=\sum_{i=1}^{n}\sum_{k=1}^{c}u_{ik}^{\beta}d_{ik}^{2}$$

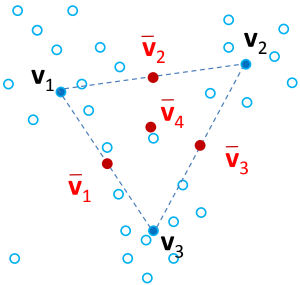
with $d_{ik}=||\boldsymbol{x}_i-\boldsymbol{v}_k||$ subject to the constraints $\sum_k u_{ik}=1$ for all i.
- Alternate optimization algorithm:

$$\boldsymbol{v}_k=\frac{\sum_{i=1}^{n}u_{ik}^{\beta}\boldsymbol{x}_i}{\sum_{i=1}^{n}u_{ik}^{\beta}}$$

$$u_{ik}=\frac{d_{ik}^{-2/(\beta-1)}}{\sum_{\ell=1}^{c}d_{i\ell}^{-2/(\beta-1)}}.$$

ECM algorithm

Principle



- Each cluster ω_k represented by a prototype $\boldsymbol{v}_k$.
- Each nonempty set of clusters A_j represented by a prototype $\bar{\boldsymbol{v}}_j$ defined as the center of mass of the $\boldsymbol{v}_k$ for all $\omega_k \in A_j$.
- Basic ideas:
  - For each nonempty $A_j \subseteq \Omega$, $m_{ij} = m_i(A_j)$ should be high if $\boldsymbol{x}_i$ is close to $\bar{\boldsymbol{v}}_j$.
  - The distance to the empty set is defined as a fixed value δ.

ECM algorithm: objective criterion

- Define the nonempty focal sets $\mathcal{F} = \{A_1, \ldots, A_f\} \subseteq 2^\Omega \setminus \{\emptyset\}$.
- Minimize

$$J_{\text{ECM}}(M, V) = \sum_{i=1}^{n} \sum_{j=1}^{f} |A_j|^\alpha m_{ij}^\beta d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^\beta$$

  subject to the constraints $\sum_{j=1}^{f} m_{ij} + m_{i\emptyset} = 1$ for all i.
- Parameters:
  - α controls the specificity of mass functions (default: 1)
  - β controls the hardness of the credal partition (default: 2)
  - δ controls the proportion of data considered as outliers
- $J_{\text{ECM}}(M, V)$ can be iteratively minimized with respect to M and to V.

ECM algorithm: update equations

Update of M:

$$m_{ij} = \frac{c_j^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{k=1}^{f} c_k^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}},$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, f$, and

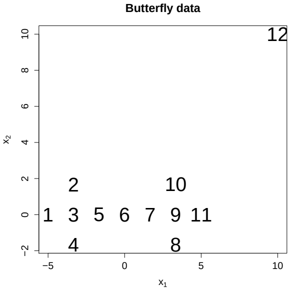
$$m_{i\emptyset} = 1 - \sum_{j=1}^{f} m_{ij}, \quad i = 1, \ldots, n$$

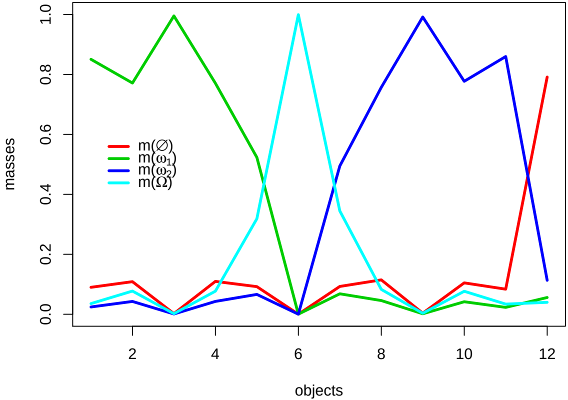
Update of V: solve a linear system of the form

$$HV = B,$$

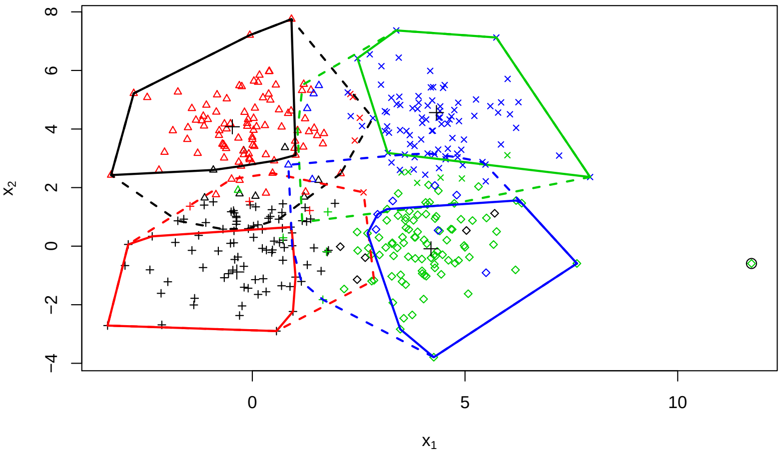
where B is a matrix of size $c \times p$ and H a matrix of size $c \times c$.

Butterfly dataset





4-class data set



Determining the number of groups

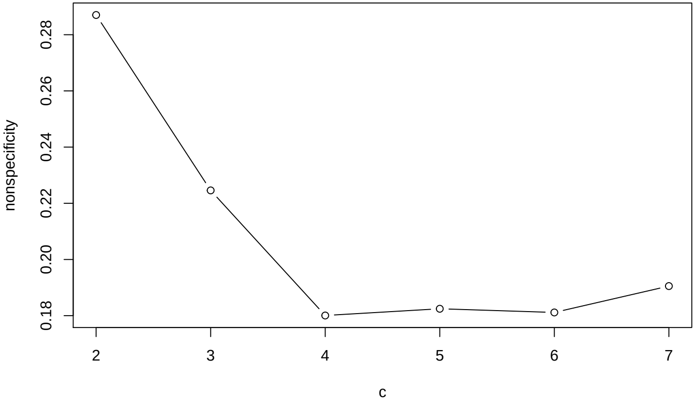
- If a proper number of groups is chosen, the prototypes will cover the clusters and most of the mass will be allocated to singletons of Ω.
- On the contrary, if c is too small or too high, the mass will be distributed to subsets with higher cardinality or to $\emptyset$.
- Nonspecificity of a mass function:

$$N(m) \triangleq \sum_{A \in 2^\Omega \setminus \emptyset} m(A) \log_2 |A| + m(\emptyset) \log_2 |\Omega|$$

- Proposed validity index of a credal partition:

$$N^*(c) \triangleq \frac{1}{n \log_2(c)} \sum_{i=1}^{n} \left[\sum_{A \in 2^\Omega \setminus \emptyset} m_i(A) \log_2 |A| + m_i(\emptyset) \log_2(c) \right]$$

Results for the 4-class dataset



Constrained Evidential *c*-means

- In some cases, we may have some prior knowledge about the group membership of some objects.
- Such knowledge may take the form of instance-level constraints of two kinds:
  1. Must-link (ML) constraints, which specify that two objects certainly belong to the same cluster;
  2. Cannot-link (CL) constraints, which specify that two objects certainly belong to different clusters.
- How to take into account such constraints?

Modified cost-function

- To take into account ML and CL constraints, we can modify the cost function of ECM as

$$J_{\text{CECM}}(M, V) = (1-\xi) J_{\text{ECM}}(M, V) + \xi J_{\text{CONST}}(M)$$

with

$$J_{\text{CONST}}(M) = \frac{1}{|\mathcal{M}| + |\mathcal{C}|} \left[\sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} pl_{ij}(\neg S) + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}} pl_{ij}(S) \right]$$

where
  - $\mathcal{M}$ and $\mathcal{C}$ are, respectively, the sets of ML and CL constraints.
  - $pl_{ij}(S)$ and $pl_{ij}(\neg S)$ are computed from the pairwise mass function m_{ij}

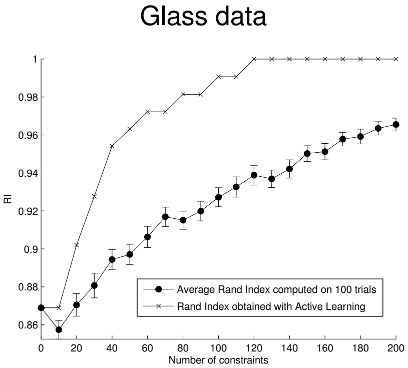
▸ Go back to pairwise mass functions

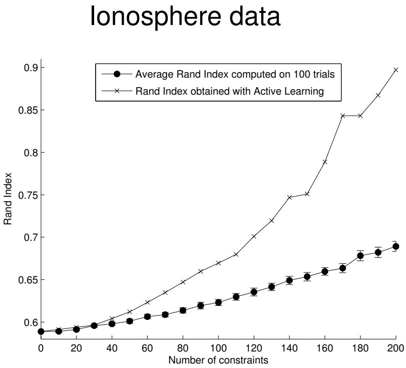
- Minimizing $J_{\text{CECM}}(M, V)$ w.r.t. M is a quadratic programming problem.

Active learning

- ML and CL constraints are sometimes given in advance, but they can sometimes be elicited from the user using an active learning strategy.
- For instance, we may select pairs of object such that
  - The first object is classified with high uncertainty (e.g., an object such that m_i has high nonspecificity);
  - The second object is classified with low uncertainty (e.g., an object that is close to a cluster center).
- The user is then provided with this pair of objects, and enters either a ML or a CL constraint.

Results


Glass data
RI
Average Rand Index computed on 100 trials
Rand Index obtained with Active Learning
Number of constraints


Ionosphere data
Rand Index
Average Rand Index computed on 100 trials
Rand Index obtained with Active Learning
Number of constraints

Other variants of ECM

Relational Evidential *c*-Means (RECM) for (metric) proximity data (Masson and Denœux, 2009).

ECM with adaptive metrics to obtain non-spherical clusters (Antoine et al., 2012). Specially useful with CECM.

Spatial Evidential C-Means (SECM) for image segmentation (Lelandais et al., 2014).

Credal *c*-means (CCM) : different definition of the distance between a vector and a meta-cluster (Liu et al., 2014).

Median evidential *c*-means (MECM) : different cost criterion, extension of the median hard and fuzzy *c*-means (Zhou et al., 2015).

Outline

1. Evidential clustering
   - Credal partition
   - Summarization of a credal partition
   - Relational representation of a credal partition
2. Evidential clustering algorithms
   - Evidential *c*-means
   - EVCLUS
   - E*k*-NNclus
3. Comparing and combining the results of soft clustering algorithms
   - The credal Rand index
   - Combining clustering structures

Learning a Credal Partition from proximity data

- Problem: given the dissimilarity matrix $D = (d_{ij})$, how to build a "reasonable" credal partition ?
- We need a model that relates cluster membership to dissimilarities.
- Basic idea: "The more similar two objects, the more plausible it is that they belong to the same group".
- How to formalize this idea?

Formalization

- Let m_i and m_j be mass functions regarding the group membership of objects o_i and o_j.
- We have seen that the plausibility that objects o_i and o_j belong to the same group is

$$pl_{ij}(S) = \sum_{A \cap B \neq \emptyset} m_i(A) m_j(B) = 1 - \kappa_{ij}$$

  where κ_{ij} = degree of conflict between m_i and m_j.
- Problem: find a credal partition $M = (m_1, \ldots, m_n)$ such that larger degrees of conflict κ_{ij} correspond to larger dissimilarities d_{ij}.

Cost function

- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict κ_{ij}.
- Example of a cost (stress) function:

$$J(M) = \sum_{i<j} (\kappa_{ij} - \varphi(d_{ij}))^2$$

where φ is an increasing function from $[0, +\infty)$ to $[0, 1]$, for instance

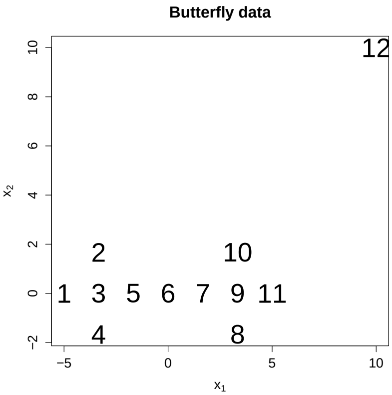
$$\varphi(d) = 1 - \exp(-\gamma d^2).$$

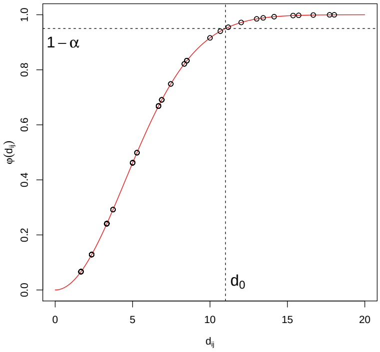
- γ can be determined by fixing $\alpha \in (0, 1)$ and d_0 such that, for any two objects (o_i, o_j) with $d_{ij} \geq d_0$, the plausibility that they belong to the same cluster is at leat $1 - \alpha$.

Butterfly example

Data and dissimilarities

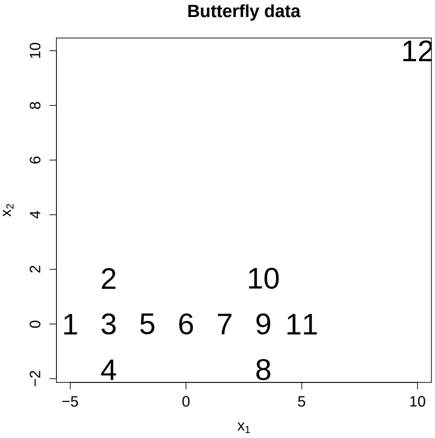
Determination of γ in $\varphi(d) = 1 - \exp(-\gamma d^2)$: fix $\alpha \in (0, 1)$ and d_0 such that, for any two objects (o_i, o_j) with $d_{ij} \geq d_0$, the plausibility that they belong to the same cluster is at least $1 - \alpha$.

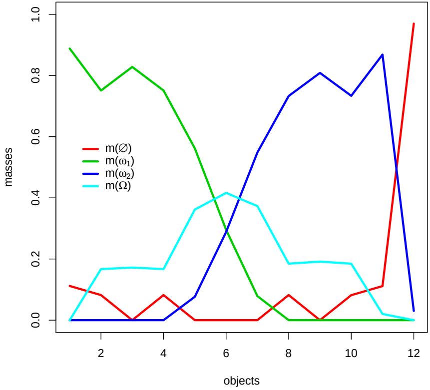




Butterfly example

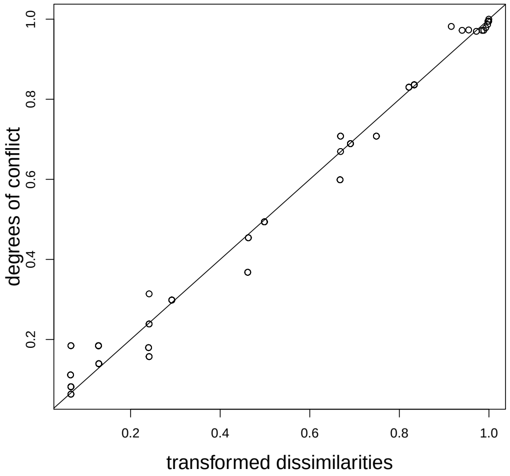
Credal partition



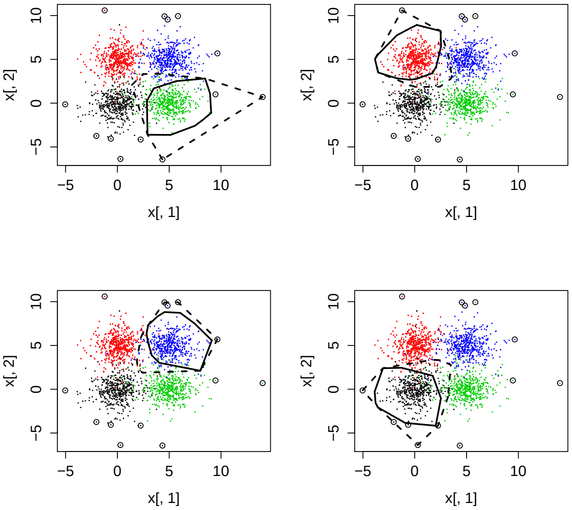


Butterfly example

Shepard diagram



Example with a four-class dataset (2000 objects)



Advantages

- Conceptually simple, clear interpretation.
- EVCLUS can handle non metric dissimilarity data (even expressed on an ordinal scale).
- It was also shown to outperform some of the state-of-the-art relational clustering techniques on a number of datasets (Denoeux and Masson, 2004).

Limitations

- Requires to store the whole dissimilarity matrix; the space complexity is thus $O(n^2)$, where n is the number of objects. Restricts application to datasets with $n \sim 10^2 - 10^3$.
- Each computation of the gradient requires $O(f^3 n^2)$ operations, where f is the number of focal sets of the mass functions. In the worst case, $f = 2^c$.
- To make the method usable even for moderate values of c, we need to restrict the form of the mass functions so that masses are only assigned to focal sets of size 0, 1 or c, which prevents us from fully exploiting the potential generality of the method.

Improvements of EVCLUS

1. Fast optimization algorithm
2. Sample dissimilarities
3. Carefully select the focal sets

Fast optimization

- The optimization algorithm initially used in EVCLUS is a gradient-based procedure.
- Here, we propose to use a cyclic coordinate descent algorithm that minimizes $J(M)$ with respect to each m_i at a time.
- The new method, called Iterative Row-wise Quadratic Programming (IRQP), exploits the particular approach of the problem (a quadratic programming problem is solved at each step), and it is thus much more efficient.

IRQP algorithm

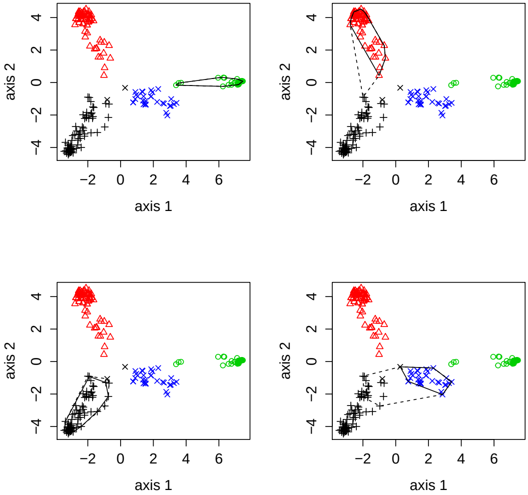
Vector representation of the cost function

- The stress function can be written as

$$J(M) = \sum_{i<j} (\boldsymbol{m}_i^T \boldsymbol{C} \boldsymbol{m}_j - \delta_{ij})^2.$$

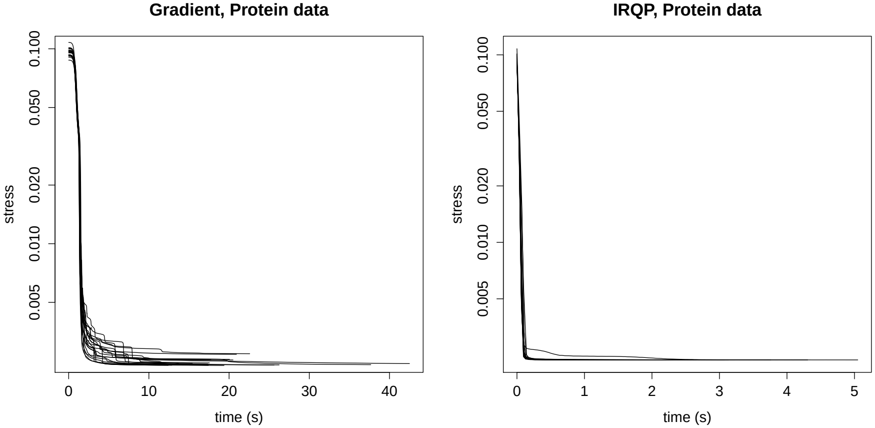
  where
  - $\delta_{ij} = \varphi(d_{ij})$ are the scaled dissimilarities
  - $\boldsymbol{m}_i$ and $\boldsymbol{m}_j$ are vectors encoding mass functions m_i and m_j
  - $\boldsymbol{C}$ is a square matrix, with general term $C_{k\ell} = 1$ if $F_k \cap F_\ell = \emptyset$ and $C_{k\ell} = 0$ otherwise.
- Fixing all mass functions except m_i, the stress function becomes quadratic. Minimizing J w.r.t. $\boldsymbol{m}_i$ is a linearly constrained positive least-squares problem, which can be solved using efficient algorithms.
- By iteratively updating each m_i, the algorithm converges to a local minimum of the cost function.

Experiment 1: Proteins dataset



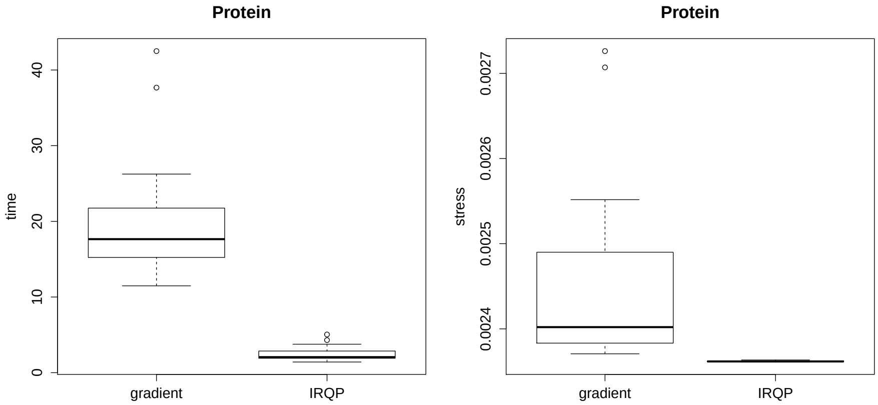
- Nonmetric dissimilarity matrix derived from the structural comparison of 213 protein sequences.
- Ground truth: 4 classes of globins.
- Only 2 errors.

Experiment 1: Proteins dataset



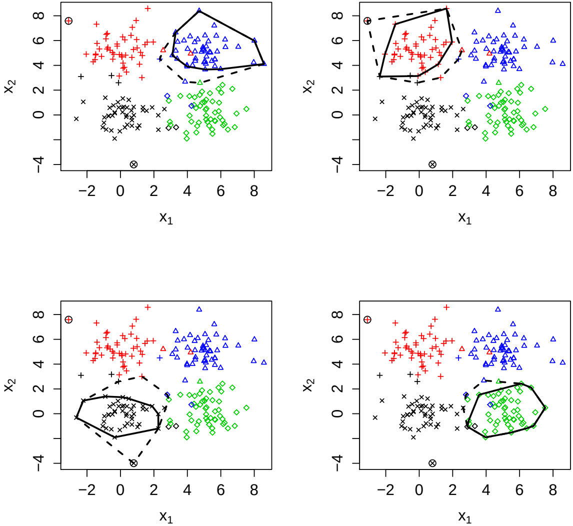
Stress vs. time (in seconds) for 20 runs of the Gradient (left) and IRQP (right) algorithms on the Protein data.

Experiment 1: Proteins dataset

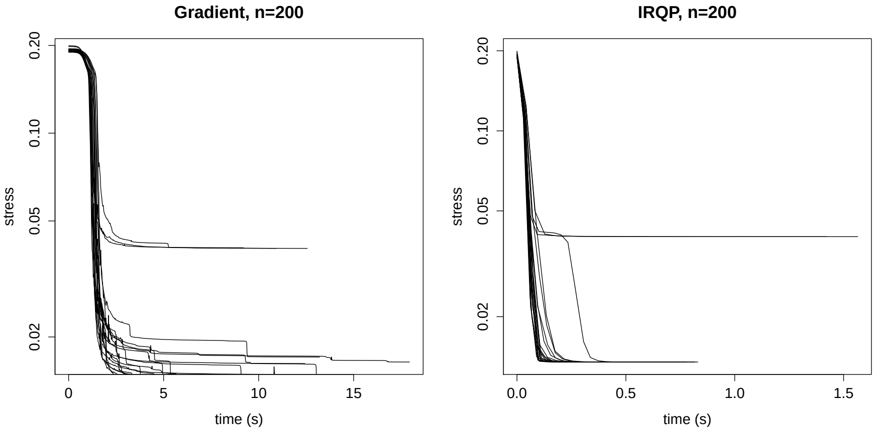


Boxplots of computing time (left) and stress value at convergence (right) for 20 runs of the Gradient and IRQP algorithms on the Protein data.

Experiment 2: simulated data ($n = 200$)

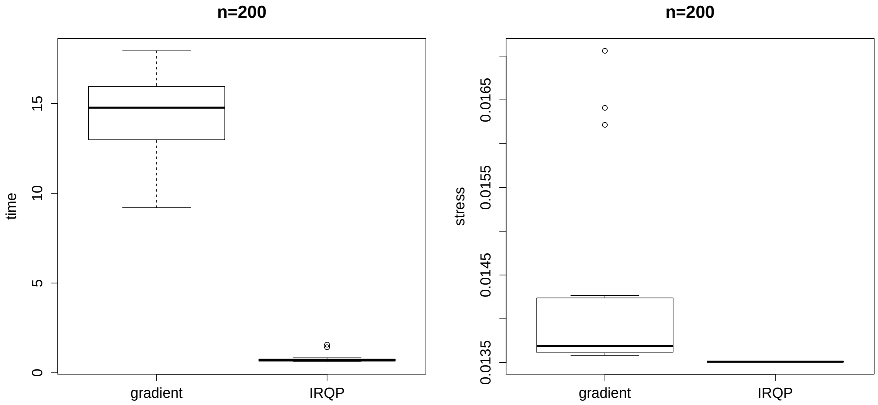


Experiment 2: simulated data ($n = 200$)



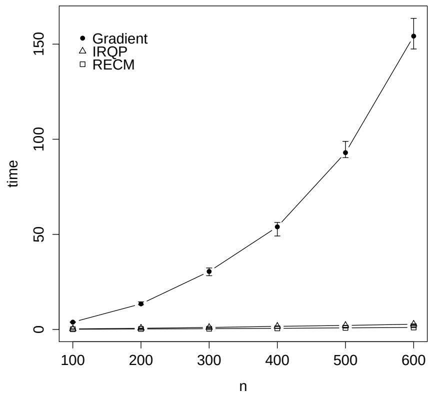
Boxplots of computing time (left) and stress value at convergence (right) for 20 runs of the Gradient and IRQP algorithms on the simulated data.

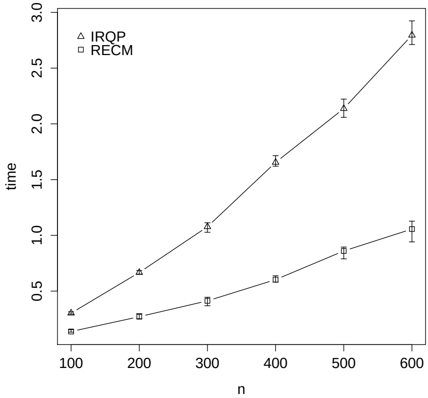
Experiment 2: simulated data ($n = 200$)



Boxplots of computing time (left) and stress value at convergence (right) for 20 runs of the Gradient and IRQP algorithms on the simulated data.

Influence of n





Computing time (in s) as a function of n for EVCLUS with the Gradient and IRQP algorithms and for RECM (left), and zoom on the curves corresponding to IRQP and RECM (right)

Sampling dissimilarities

- EVCLUS requires to store the whole dissimilarity matrix: it is inapplicable to large proximity data.
- However, there is usually some redundancy in a dissimilarity matrix.
- In particular, if two objects o_1 and o_2 are very similar, then any object o_3 that is dissimilar from o_1 is usually also dissimilar from o_2.
- Because of such redundancies, it might be possible to compute the differences between degrees of conflict and dissimilarities, for only a subset of randomly sampled dissimilarities.

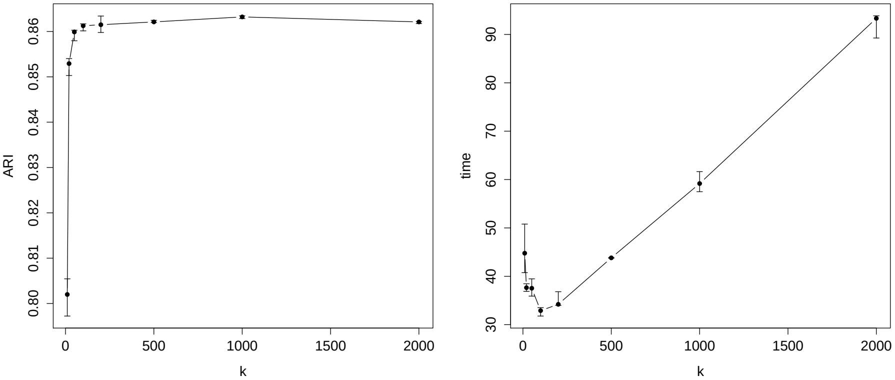
New stress function

- Let $j_1(i), \ldots, j_k(i)$ be k integers sampled at random from the set $\{1, \ldots, i-1, i+1, \ldots, n\}$, for $i = 1, \ldots, n$.
- Let J_k the following stress criterion,

$$J_k(M) = \sum_{i=1}^{n} \sum_{r=1}^{k} (\kappa_{i,j_r(i)} - \delta_{i,j_r(i)})^2.$$

- The calculation of $J_k(M)$ requires only $O(nk)$ operations.
- If k can be kept constant as n increases, then time and space complexities are reduced from quadratic to linear.

Example with simulated data ($n = 10,000$)

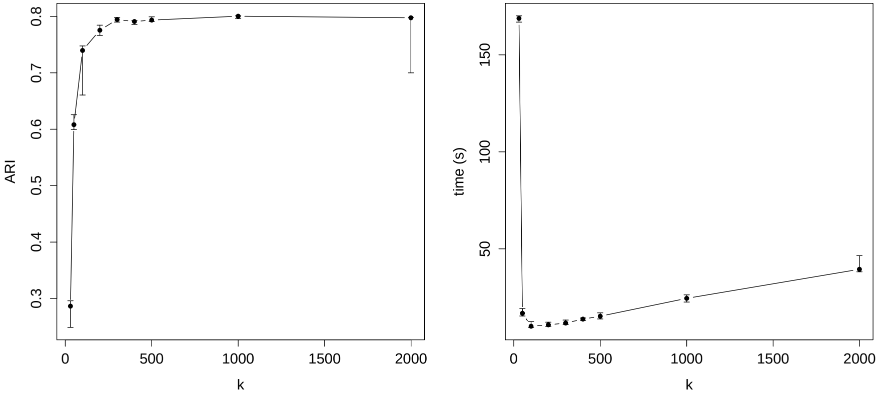


Zongker Digit dissimilarity data

- Similarities between 2000 handwritten digits in 10 classes, based on deformable template matching.
- k-EVCLUS was run with $c = 10$ and differents following values of k.
- Parameter d_0 was fixed to the 0.3-quantile of the dissimilarities.
- k-EVCLUS was run 10 times with random initializations.

Zongker Digit dissimilarity data

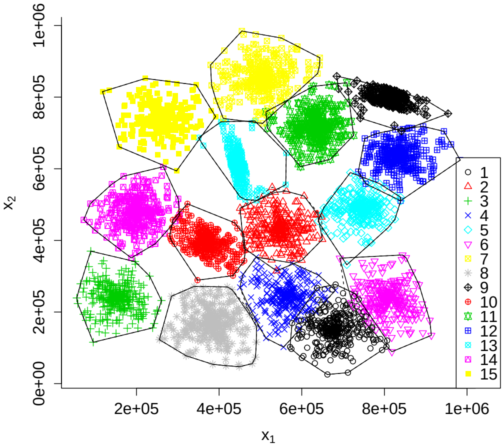
Results



Carefully selecting the focal sets

- If no restriction is imposed on the focal sets, the number of parameters to be estimated in evidential clustering grows exponentially with the number c of clusters, which makes it intractable unless c is small.
- If we allow masses to be assigned to all pairs of clusters, the number of focal sets becomes proportional to c^2, which is manageable for moderate values of c (say, until 10), but still impractical for larger n.
- Idea: assign masses only to pairs of contiguous clusters.
- If each cluster has at most q neighbors, then the number of focal sets is proportional to c.

Example



The S_2 dataset ($n = 5000$) and the 15 clusters found by k-EVCLUS with $k = 100$

Method

Step1: Run a clustering algorithm (e.g., ECM or EVCLUS) with focal sets of cardinalities 0, 1 and (optionally) c. A credal partition M_0 is obtained.

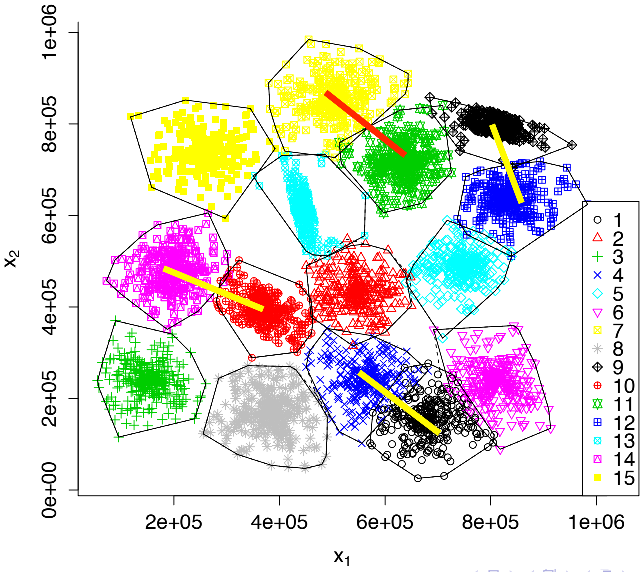
Step 2: Compute the similarity between each pair of clusters (ω_j, ω_ℓ) as

$$S(j, \ell) = \sum_{i=1}^{n} pl_{ij} pl_{i\ell},$$

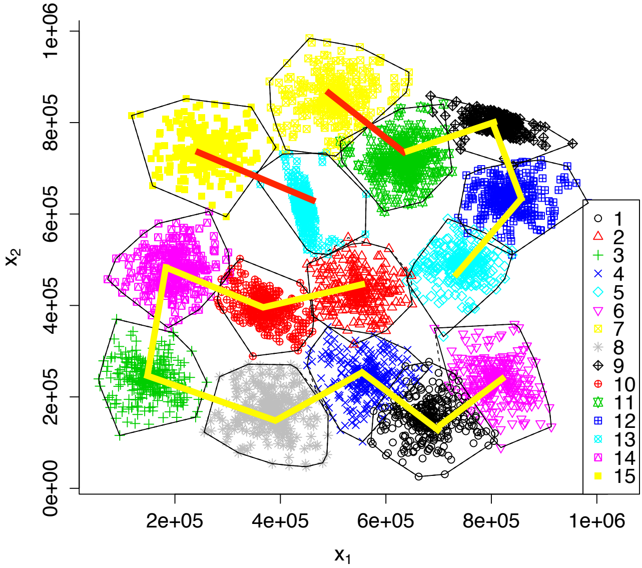
where pl_{ij} and $pl_{i\ell}$ are the normalized plausibilities that object i belongs, respectively, to clusters j and ℓ. Determine the set P_K of pairs $\{\omega_j, \omega_\ell\}$ that are mutual q nearest neighbors.

Step 3: Run the clustering algorithm again, starting from the previous credal partition M_0, and adding as focal sets the pairs in P_K.

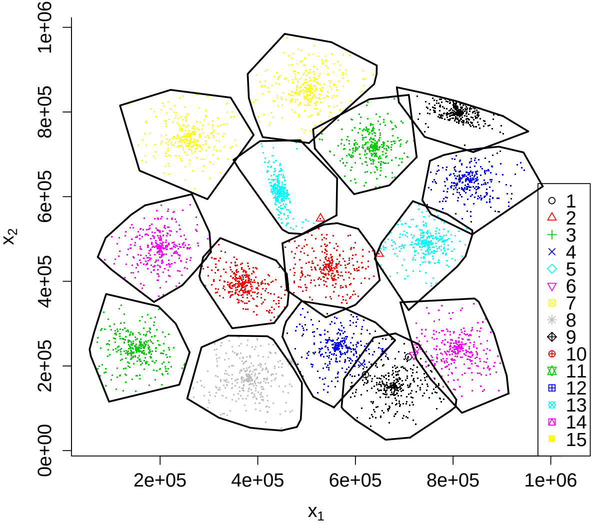
Pairs of mutual neighbors with $q = 1$



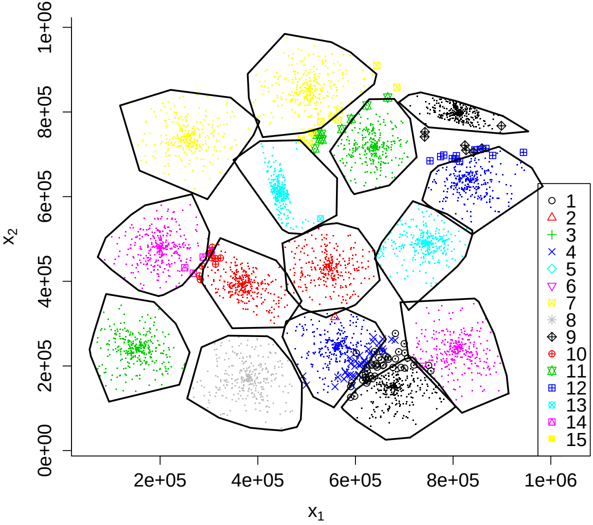
Pairs of mutual neighbors with $q = 2$



Initial credal partition $\mathcal{M}_0$



Final credal partition ($q = 1$)

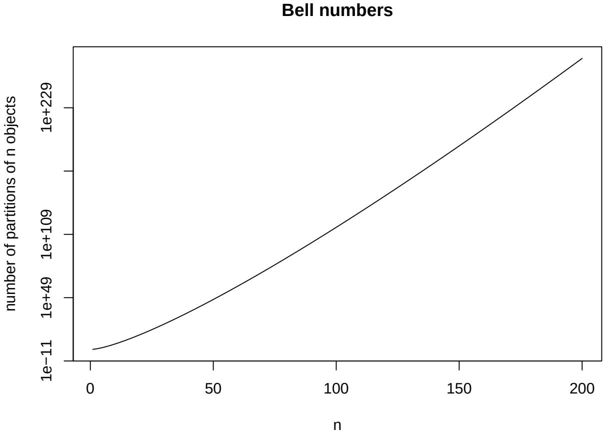


Outline

Reasoning in the space of all partitions

- Assuming there is a true unknown partition, our frame of discernment should be the set $\mathcal{R}$ of all equivalent relations ($\equiv$ partitions) of the set of n objects.
- But this set is huge!

Number of partitions of *n* objects



Can we implement evidential reasoning in such a large space?

Model

- Evidence: $n \times n$ matrix $D = (d_{ij})$ of dissimilarities between the n objects.
- Assumptions
  1. Two objects have all the more chance to belong to the same group, that they are more similar:
  $$m_{ij}(\{S\}) = \varphi(d_{ij}),$$
  $$m_{ij}(\Theta) = 1 - \varphi(d_{ij}),$$
  where φ is a non-increasing mapping from $[0, +\infty)$ to $[0, 1)$.
  2. The mass functions m_{ij} are independent.
- How to combine these $n(n-1)/2$ mass functions to find the most plausible partition of the n objects?

Evidence combination

- Let $\mathcal{R}_{ij}$ denote the set of partitions of the n objects such that objects o_i and o_j are in the same group ($r_{ij} = 1$).
- Each mass function m_{ij} can be vacuously extended to the space $\mathcal{R}$ of equivalence relations:

$$\begin{array}{ccc} m_{ij}(\{S\}) & \longrightarrow & \mathcal{R}_{ij} \\ m_{ij}(\Theta) & \longrightarrow & \mathcal{R} \end{array}$$

- The extended mass functions can then be combined by Dempster's rule.
- Resulting contour function:

$$pl(R) \propto \prod_{i<j} (1 - \varphi(d_{ij}))^{1-r_{ij}}$$

for any $R \in \mathcal{R}$.

Decision

- The logarithm of the contour function can be written as

$$\log pl(R) = -\sum_{i<j} r_{ij} \log(1 - \varphi(d_{ij})) + C$$

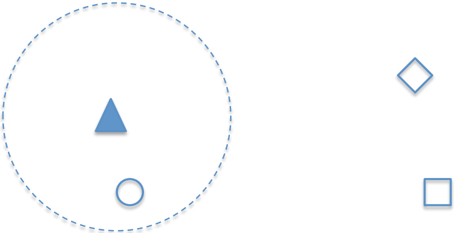
- Finding the most plausible partition is thus a binary linear programming problem. It can be solves exactly only for small n.
- However, the problem can be solved approximately using a heuristic greedy search procedure: the Ek-NNclus algorithm.
- This is a decision-directed clustering procedure, using the evidential k-nearest neighbor (Ek-NN) rule as a base classifier.

Example

Toy dataset

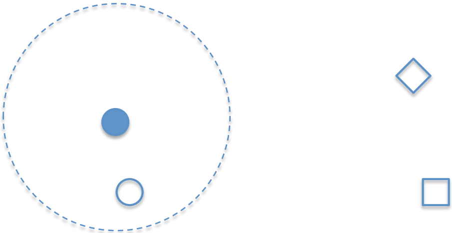
Example

Iteration 1



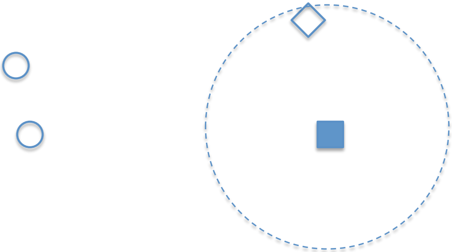
Example

Iteration 1 (continued)



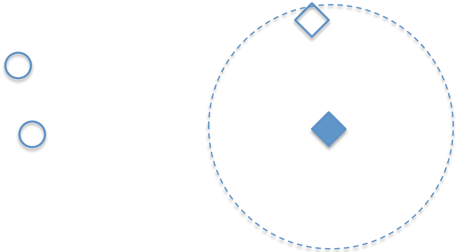
Example

Iteration 2



Example

Iteration 2 (continued)



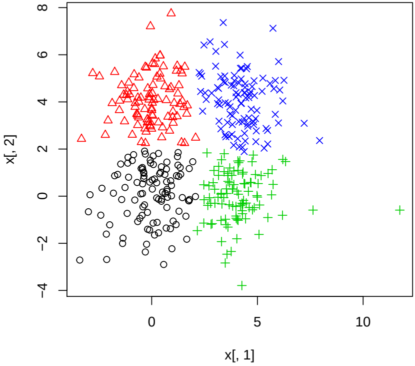
Example

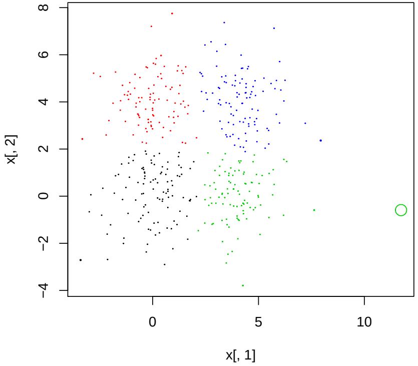
Result

Ek-NNclus

- Starting from a random initial partition, classify each object in turn, using the Ek-NN rule.
- The algorithm converges to a local maximum of the contour function $pl(R)$ if $k = n - 1$.
- With $k < n - 1$, the algorithm converges to a local maximum of an objective function that approximates $pl(R)$.
- Implementation details:
  - Number k of neighbors: two to three times $\sqrt{n}$.
  - $\varphi(d) = 1 - \exp(-\gamma d^2)$, with γ fixed to the inverse of the q-quantile of the distances d_{ij}^2 between an object and its k NN. Typically, $q \geq 0.5$.
  - The number of clusters does not need to be fixed in advance.

Example





Outline

Exploiting the generality of evidential clustering

- We have seen that the concept of credal partition subsumes the main hard and soft clustering structures.
- Consequently, methods designed to evaluate or combine credal partitions can be used to evaluate or combine the results of any hard or soft clustering algorithms.
- Two such methods will be described:
  1. A generalization of the Rand index to compute the distance between two credal partitions;
  2. A method to combine credal partitions.

Outline

Rand index

- The Rand index is a widely used measure of agreement (similarity) tbetween two hard partitions.
- It is defined as

$$RI = \frac{a + b}{n(n-1)/2}$$

  with
  - $a =$ number of pairs of objects that are grouped together in both partitions
  - $b =$ number of pairs of objects that are assigned to different clusters in both partitions.
- How to generalize the Rand Index to credal partitions?

Jousselme's distance

- Let $R = (m_{ij})$ and $R' = (m'_{ij})$ be the relational representations of two credal partitions.
- The assess the distance between R and R', we can average the distances between the m_{ij}'s and m'_{ij}'s.
- A suitable measure is the squared Jousselme's metric, defined as

$$d_{ij} = \left(\frac{1}{2}(\boldsymbol{m}_{ij} - \boldsymbol{m}'_{ij})^T J(\boldsymbol{m}_{ij} - \boldsymbol{m}'_{ij})\right)^{1/2}$$

with $\boldsymbol{m}_{ij} = (m_{ij}(\emptyset), m_{ij}(\{s\}), m_{ij}(\{ns\}), m_{ij}(\Theta))^T$ and

$$J = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1/2 \\ 0 & 0 & 1 & 1/2 \\ 0 & 1/2 & 1/2 & 1 \end{pmatrix}$$

Credal Rand index

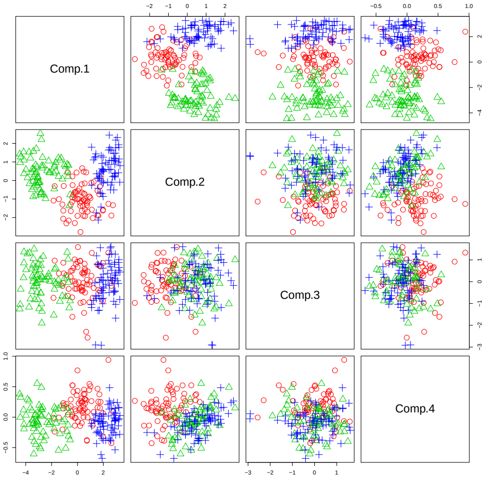
- We define the Credal Rand Index as

$$CRI = 1 - \frac{\sum_{i<j} d_{ij}}{n(n-1)/2}.$$

- Properties:
  - $0 \leq CRI \leq 1$
  - CRI is the Rand index when the two partitions are hard
  - Symmetry: $CRI(R, R') = CRI(R', R)$
  - If $R = R'$, then $CRI(R, R') = 1$
  - 1-CRI is a metric in the space of relational representations of credal partitions (it is reflexive, symmetric, separable and it verifies the triangular inequality).
- The CRI can be used to compare the results of any two hard or soft clustering algorithms.

Example: Seeds data

Seeds from three different varieties of wheat: Kama, Rosa and Canadian, 70 elements each, 7 features. First 4 principal components:

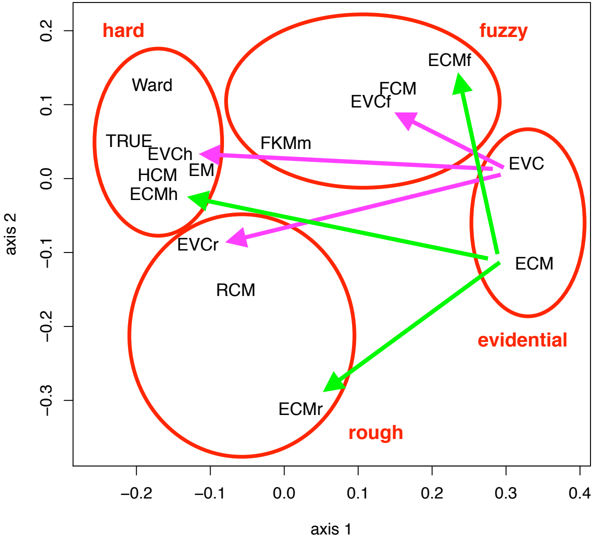


Clustering algorithms

- Evidential clustering (R package `evclust`)
  - ECM, $\mathcal{F} = \{A \subseteq \Omega, |A| \leq 2\}$
  - EVCLUS ($\mathcal{F} = \{A \subseteq \Omega, |A| \leq 1\} \cup \{\Omega\}$; $\mathcal{F} = 2^\Omega$).

  and their derived hard, fuzzy and rough partitions
- Hard clustering: HCM (R package `stats`)
- Fuzzy clustering (R package `fclust`)
  - FCM
  - Fuzzy K medoids
- Rough clustering (R package `SoftClustering`)
  - Peter's rough k-means P-RCM
  - Pi rough k-means π-RCM

Result: MDS configuration

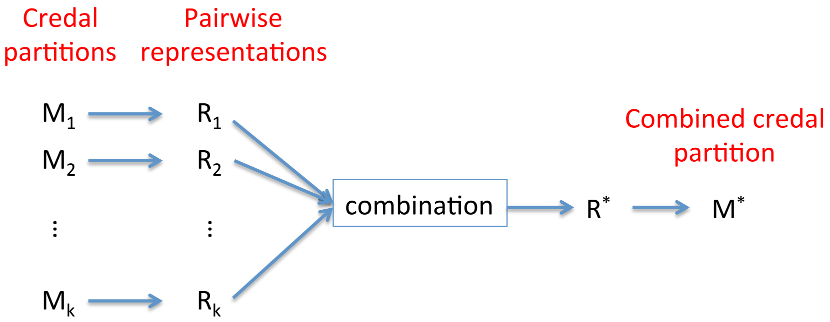


Outline

Motivations for combining clustering structures

- Let $M_1, \ldots, M_N$ be an ensemble of N credal partitions generated by hard or soft (fuzzy, rough, etc.) clustering structures.
- It may be useful to combine these credal partitions:
  - to increase the chance of finding a good approximation to the true partition, or
  - to highlight invariant patterns across the clustering structures.
- Combination is easily carried out using relational representations.

Combination method



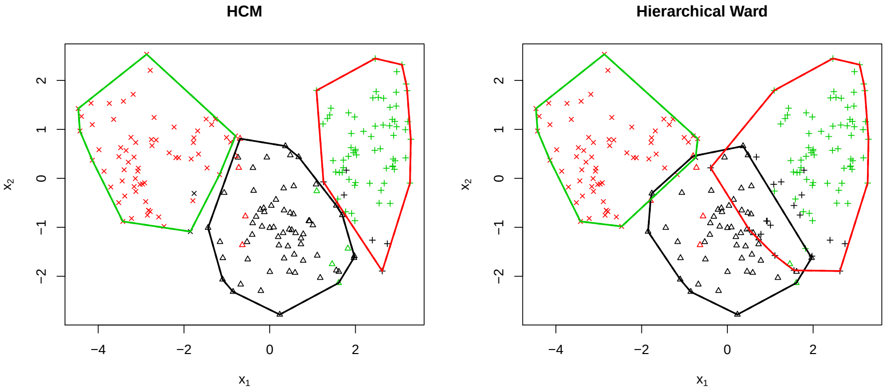
The combined credal partition can be defined as

$$M^* = \arg\max_M CRI(\mathcal{R}(M), R^*),$$

where $\mathcal{R}(M)$ denotes the relational representation of M.

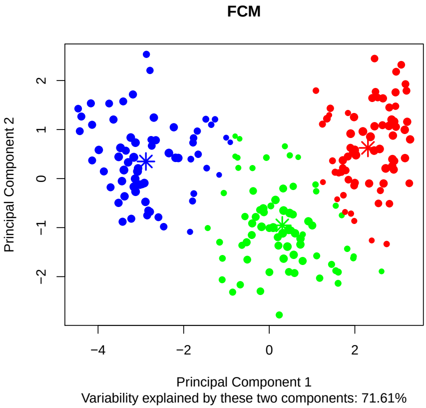
Example: seeds data

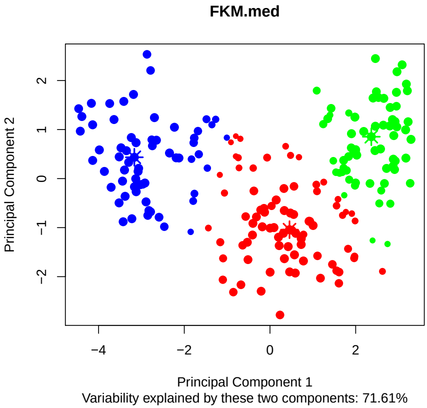
Hard clustering results



Example: seeds data

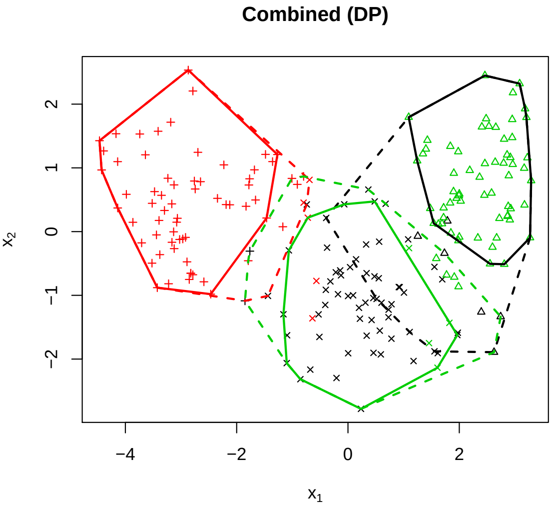
Fuzzy clustering results





Example: seeds data

Combined credal partition (Dubois-Prade rule)



Summary

- The Dempster-Shafer theory of belief functions provides a rich and flexible framework to represent uncertainty in clustering.
- The concept of credal partition encompasses the main existing soft clustering concepts (fuzzy, possibilistic, rough partitions).
- Efficient algorithms exist, allowing one to generate credal partitions from attribute or proximity datasets.
- These algorithms can be applied to large datasets and large numbers of clusters (by carefully selecting the focal sets).
- Concepts from the theory of belief functions make it possible to compare and combine clustering structures generated by various soft clustering algorithms.

Future research directions

- Combining clustering structures in various settings
  - distributed clustering,
  - combination of different attributes, different algorithms,
  - etc.
- Handling huge datasets (several millions of objects)
- Criteria for selecting the number of clusters
- Semi-supervised clustering
- Clustering imprecise or uncertain data
- Applications to image processing, social network analysis, process monitoring, etc.
- Etc...

The evclust package

evclust: Evidential Clustering

Various clustering algorithms that produce a credal partition, i.e., a set of Dempster-Shafer mass functions representing the membership of objects to clusters. The mass functions quantify the cluster-membership uncertainty of the objects. The algorithms are: Evidential c-Means (ECM), Relational Evidential c-Means (RECM), Constrained Evidential c-Means (CECM), EVCLUS and EK-NNclus.

Version:	1.0.3
Depends:	R (≥ 3.1.0)
Imports:	FNN, R.utils, limSolve, Matrix
Suggests:	knitr, rmarkdown
Published:	2016-09-04
Author:	Thierry Denoeux
Maintainer:	Thierry Denoeux <tdenoeux at utc.fr>
License:	GPL-3
NeedsCompilation:	no
In views:	Cluster
CRAN checks:	evclust results

https://cran.r-project.org/web/packages

References on clustering I

cf. https://www.hds.utc.fr/~tdenoeux

M.-H. Masson and T. Denœux.
ECM: An evidential version of the fuzzy c-means algorithm.
Pattern Recognition, 41(4):1384-1397, 2008.

M.-H. Masson and T. Denœux.
RECM: Relational Evidential c-means algorithm.
Pattern Recognition Letters, 30:1015-1026, 2009.

B. Lelandais, S. Ruan, T. Denoeux, P. Vera, I. Gardin.
Fusion of multi-tracer PET images for Dose Painting.
Medical Image Analysis, 18(7):1247-1259, 2014.

V. Antoine, B. Quost, M.-H. Masson and T. Denoeux.
CECM: Constrained Evidential C-Means algorithm.
Computational Statistics and Data Analysis, 56(4):894-914, 2012.

References on clustering II

cf. https://www.hds.utc.fr/~tdenoeux

- T. Denœux and M.-H. Masson.
  EVCLUS: Evidential Clustering of Proximity Data.
  IEEE Transactions on SMC B, 34(1):95-109, 2004.
- T. Denœux, S. Sriboonchitta and O. Kanjanatarakul
  Evidential clustering of large dissimilarity data.
  Knowledge-Based Systems, 106:179–195, 2016.
- T. Denoeux, O. Kanjanatarakul and S. Sriboonchitta.
  EK-NNclus: a clustering procedure based on the evidential K-nearest neighbor rule.
  Knowledge-Based Systems, Vol. 88, pages 57-69, 2015.
- T. Denoeux, S. Li and S. Sriboonchitta.
  Evaluating and Comparing Soft Partitions: an Approach Based on Dempster-Shafer Theory.
  IEEE Transactions on Fuzzy Systems, submitted, 2017.