Dempster-Shafer theory

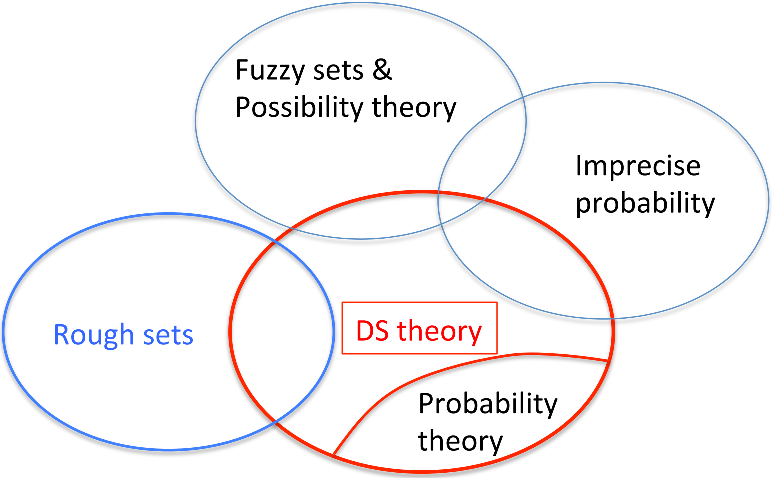
Introduction, connections with rough sets and application to clustering

Thierry Denœux[1]

[1] Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)
http://www.hds.utc.fr/~tdenoeux

RSKT 2014
Shanghai, China
October 25, 2014

Theories of uncertainty



Focus of this talk

- Dempster-Shafer (DS) theory (evidence theory, theory of belief functions):
  - A formal framework for reasoning with partial (uncertain, imprecise) information.
  - Has been applied to statistical inference, expert systems, information fusion, classification, clustering, etc.
- Purpose of these talk:
  - Brief introduction or reminder on DS theory, emphasizing some connections with rough sets;
  - Review the application of belief functions to clustering, showing some connections with fuzzy and rough approaches.

Outline

1. Dempster-Shafer theory
   - Mass function
   - Belief and plausibility functions
   - Connection with rough sets

2. Application to clustering
   - Evidential partition
   - Evidential *c*-means

Outline

Mass function

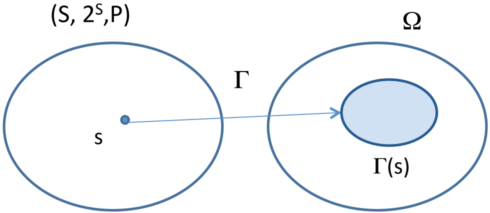
- Let Ω be a finite set called a frame of discernment.
- A mass function is a function $m : 2^\Omega \rightarrow [0, 1]$ such that

$$\sum_{A \subseteq \Omega} m(A) = 1.$$

- The subsets A of Ω such that $m(A) \neq 0$ are called the focal sets of Ω.
- If $m(\emptyset) = 0$, m is said to be normalized (usually assumed).

Source

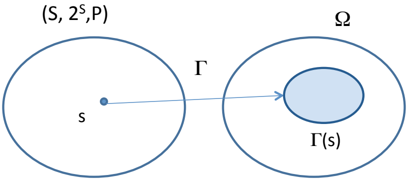
- A mass function is usually induced by a source, defined a 4-tuple $(S, 2^S, P, \Gamma)$, where
  - S is a finite set;
  - P is a probability measure on $(S, 2^S)$;
  - Γ is a multi-valued-mapping from S to 2^Ω.



- Γ carries P from S to 2^Ω: for all $A \subseteq \Omega$,

$$m(A) = P(\{s \in S | \Gamma(s) = A\}).$$

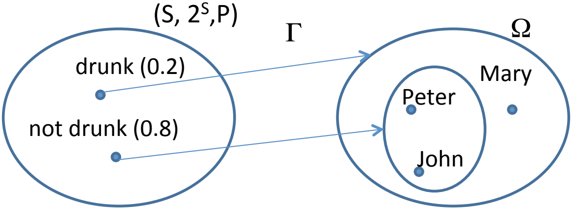
Interpretation



- Ω is a set of possible states of the world, about which we collect some evidence. Let ω be the true state.
- S is a set of interpretations of the evidence.
- If $s \in S$ holds, we know that ω belongs to the subset $\Gamma(s)$ of Ω, and nothing more.
- $m(A)$ is then the probability of knowing only that $\omega \in A$.
- In particular, $m(\Omega)$ is the probability of knowing nothing.

Example

- A murder has been committed. There are three suspects: $\Omega = \{\text{Peter}, \text{John}, \text{Mary}\}$.
- A witness saw the murderer going away, but he is short-sighted and he only saw that it was a man. We know that the witness is drunk 20 % of the time.



- We have $\Gamma(\neg\text{drunk}) = \{\text{Peter}, \text{John}\}$ and $\Gamma(\text{drunk}) = \Omega$, hence

$$m(\{\text{Peter}, \text{John}\}) = 0.8, \quad m(\Omega) = 0.2$$

Special cases

- A mass function m is said to be:
  - logical if it has only one focal set; it is then equivalent to a set.
  - Bayesian if all focal sets are singletons; it is equivalent to a probability distribution.
- A mass function can thus be seen as
  - a generalized set, or as
  - a generalized probability distribution.

Outline

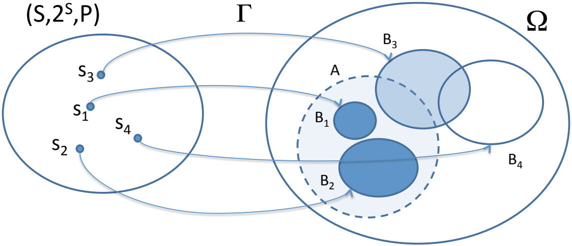
1. Dempster-Shafer theory
   - Mass function
   - Belief and plausibility functions
   - Connection with rough sets

2. Application to clustering
   - Evidential partition
   - Evidential c-means

Belief function

Degrees of support and consistency

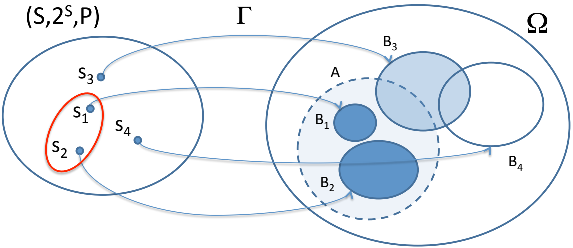
- Let m be a normalized mass function on Ω induced by a source $(S, 2^S, P, \Gamma)$.
- Let A be a subset of Ω.
- One may ask:
  1. To what extent does the evidence support the proposition $\omega \in A$?
  2. To what extent is the evidence consistent with this proposition?



Belief function
Definition and interpretation

- For any $A \subseteq \Omega$, the probability that the evidence implies (supports) the proposition $\omega \in A$ is

$$Bel(A) = P(\{s \in S | \Gamma(s) \subseteq A\}) = \sum_{B \subseteq A} m(B).$$



- The function $Bel : A \to Bel(A)$ is called a belief function.

Belief function
Characterization

- Function $Bel : 2^\Omega \rightarrow [0,1]$ is a completely monotone capacity: it verifies $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ and

$$Bel\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right).$$

for any $k \geq 2$ and for any family $A_1, \ldots, A_k$ in 2^Ω.

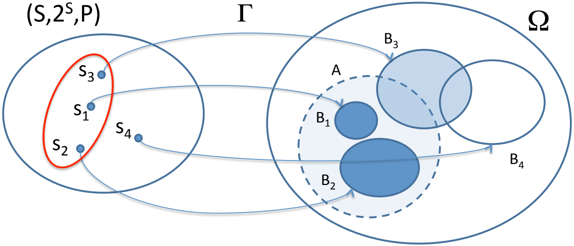
- Conversely, to any completely monotone capacity *Bel* corresponds a unique mass function m such that:

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B), \quad \forall A \subseteq \Omega.$$

Plausibility function

- The probability that the evidence is consistent with (does not contradict) the proposition $\omega \in A$

$$Pl(A) = P(\{s \in S | \Gamma(s) \cap A \neq \emptyset\}) = 1 - Bel(\overline{A})$$



- The function $Pl : A \rightarrow Pl(A)$ is called a plausibility function.

Special cases

- If m is Bayesian, then $Bel = Pl$ and it is a probability measure.
- If the focal sets of m are nested ($A_1 \subset A_2 \subset \ldots \subset A_n$), m is said to be consonant. Pl is then a possibility measure:

$$Pl(A \cup B) = \max\left(Pl(A), Pl(B)\right)$$

  for all $A, B \subseteq \Omega$ and Bel is the dual necessity measure.
- DS theory thus subsumes both probability theory and possibility theory.

Summary

- A probability measure is precise, in so far as it represents the uncertainty of the proposition $\omega \in A$ by a single number $P(A)$.
- In contrast, a mass function is imprecise (it assigns probabilities to subsets).
- As a result, in DS theory, the uncertainty about a subset A is represented by two numbers $(Bel(A), Pl(A))$, with $Bel(A) \leq Pl(A)$.
- This model is thus reminiscent of rough set theory, in which a set is approximated by lower and upper approximations, due to coarseness of a knowledge base.

Outline

Interval rough sets

Belief and plausibility functions induced by an interval relation

- Let S and Ω be two finite sets and $R \subseteq S \times \Omega$. R is called an interval relation (Yao and Lingras, 1998) if

$$\Gamma_R(s) = \{\omega \in \Omega | (s, \omega) \in R\} \neq \emptyset,$$

for all $s \in S$.

- Any $A \subseteq \Omega$ may be approximated in S by an interval rough set defined by:

$$\underline{R}(A) = \{s \in S | \Gamma_R(s) \subseteq A\}$$
$$\overline{R}(A) = \{s \in S | \Gamma_R(s) \cap A \neq \emptyset\}$$

- Let P be a probability measure on $(S, 2^S)$. Then, functions Bel and Pl defined, for all $A \subseteq \Omega$, by

$$Bel(A) = P(\underline{R}(A)), \quad Pl(A) = P(\overline{R}(A))$$

are belief and plausibility functions.

Interval rough sets

Equivalence with belief functions

- Conversely, let m be a normalized mass function on a finite set Ω, induced by a source $(S, 2^S, P, \Gamma)$. The relation

$$R = \{(s, \omega) \in S \times \Omega | \omega \in \Gamma(s)\}$$

is an interval relation, and

$$Bel(A) = P(\underline{R}(A)), \quad Pl(A) = P(\overline{R}(A)), \quad \forall A \subseteq \Omega.$$

Equivalence result

Belief function on Ω = interval relation between S and Ω
+ probability measure on $(S, 2^S)$

Rough mass functions

- Let Ω be the frame of discernment and let R be an equivalence relation on Ω defining a partition of Ω.
- Any $A \subseteq \Omega$ may be approximated by a (Pawlak) rough set defined by:

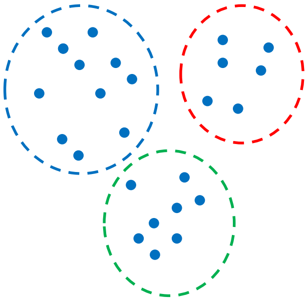
$$\underline{R}(A) = \{\omega \in \Omega | [\omega]_R \subseteq A\}$$

$$\overline{R}(A) = \{\omega \in \Omega | [\omega]_R \cap A \neq \emptyset\}$$

- Given a mass function m with focal sets $A_1, \ldots, A_n$, we can define:
  - Its lower approximation $\underline{m}$ with focal sets $\underline{R}(A_1), \ldots, \underline{R}(A_n)$;
  - Its upper approximation $\overline{m}$ with focal sets $\overline{R}(A_1), \ldots, \overline{R}(A_n)$.
- The pair $(\underline{m}, \overline{m})$ may be called a rough mass function. This notion extends that of rough set.
- Remark: these notions was introduced by Shafer (1976) with a different terminology, before the introduction of rough sets!

Outline

Clustering



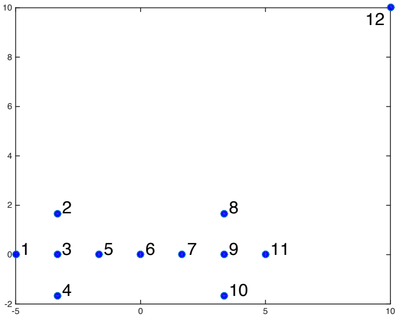
- n objects described by
  - Attribute vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (attribute data) or
  - Dissimilarities (proximity data).
- Goal: find a meaningful structure in the data set, usually a partition into c crisp or fuzzy subsets.
- Belief functions may allow us to express richer information about the data structure.

Different clustering concepts

- Hard clustering: each object belongs to one and only one group. Group membership is expressed 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 $\sum_{k=1}^{c} u_{ik} = 1$.
- Possibilistic clustering: the condition $\sum_{k=1}^{c} u_{ik} = 1$ is relaxed. Each number u_{ik} can be interpreted as a degree of possibility that object i belongs to cluster k.
- Rough clustering: the membership of object i to cluster k is described by a pair $(\underline{u}_{ik}, \overline{u}_{ik}) \in \{0, 1\}^2$ indicating its membership to the lower and upper approximations of cluster k.

Evidential clustering

- In Evidential clustering, the group membership of each object is described by a (not necessarily normalized) mass function m_i over Ω.
- Example:



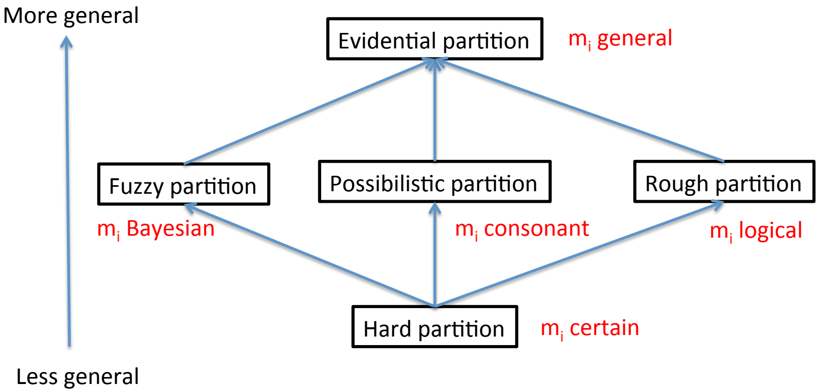
Evidential partition

	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_1, \omega_2\}$
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

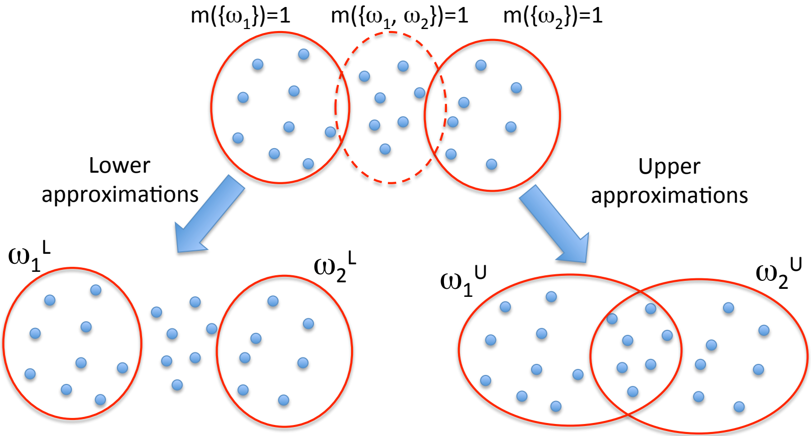
Université de Technologie Compiègne

heudiasyc

Relationship with other clustering structures



Rough clustering as a special case



From evidential to hard/fuzzy/possibilistic clustering

- Let $(m_1, \ldots, m_n)$ be an evidential partition.
- Induced hard partition:

$$u_{ik} = \begin{cases} 1 & \text{if } Pl_i(\{\omega_k\}) = \max_\ell Pl_i(\{\omega_\ell\}) \\ 0 & \text{otherwise.} \end{cases}$$

- Induced fuzzy partition:

$$u_{ik} = \frac{Pl_i(\{\omega_k\})}{\sum_\ell Pl_i(\{\omega_\ell\})}$$

- Induced possibilistic partition:

$$u_{ik} = Pl_i(\{\omega_k\})$$

From evidential to rough clustering

- Let $(m_1, \ldots, m_n)$ be an evidential partition.
- For each i, let $A_i \subseteq \Omega$ such that

$$m_i(A_i) = \max_{A \subseteq \Omega} m_i(A).$$

- Lower approximations:

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

- Upper approximations:

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

Algorithms

- EVCLUS (Denoeux and Masson, 2004):
  - Proximity (possibly non metric) data,
  - Multidimensional scaling approach.
- Evidential *c*-means (ECM): (Masson and Denoeux, 2008):
  - Attribute data,
  - HCM, FCM family (alternate optimization of a cost function).
- Relational Evidential *c*-means (RECM): (Masson and Denoeux, 2009): ECM for proximity data.
- Constrained Evidential *c*-means (CECM) (Antoine et al., 2011): ECM with pairwise constraints.
- Constrained EVCLUS (CEVCLUS) (Antoine et al., 2014): EVCLUS with pairwise constraints.

Outline

Principle

- Problem: generate an evidential partition $M = (m_1, \ldots, m_n)$ from attribute data $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$, $\mathbf{x}_i \in \mathbb{R}^p$.
- Generalization of hard and fuzzy c-means algorithms:
  - Each class represented by a prototype;
  - Alternate optimization of a cost function with respect to the prototypes and to the evidential 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 $d_{ik} = ||\mathbf{x}_i - \mathbf{v}_k||$ under the constraints $\sum_k u_{ik} = 1, \forall i$.

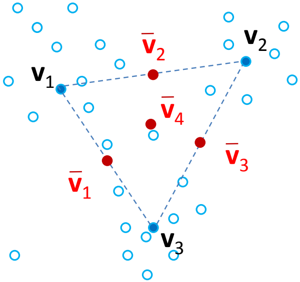
- Alternate optimization algorithm:

$$\mathbf{v}_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta}\mathbf{x}_i}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad \forall k = 1,\ldots,c,$$

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

ECM algorithm

Principle



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

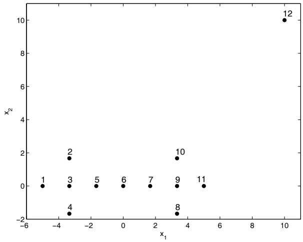
ECM algorithm
Objective criterion

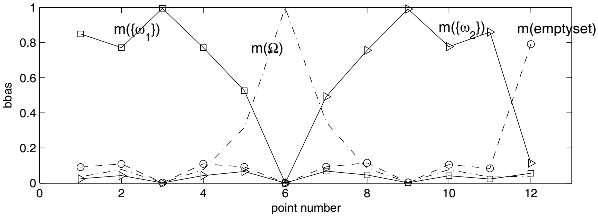
- Criterion to be minimized:

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

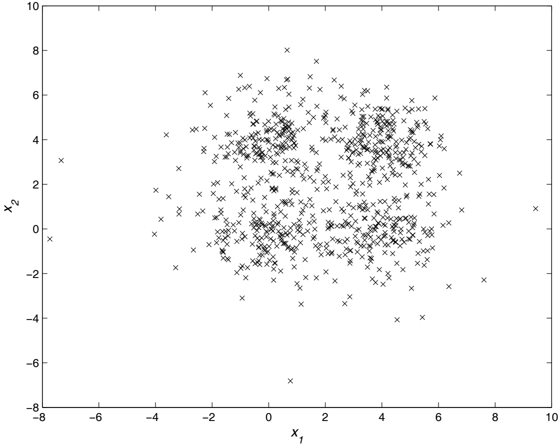
- Parameters:
  - α controls the specificity of mass functions;
  - β controls the hardness of the evidential partition;
  - δ controls the amount of data considered as outliers.
- $J_{\mathrm{ECM}}(M, V)$ can be iteratively minimized with respect to M and V using an alternate optimization scheme.

Butterfly dataset



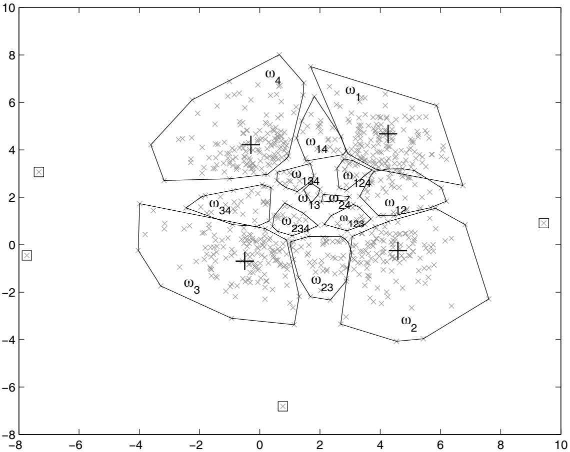


4-class data set



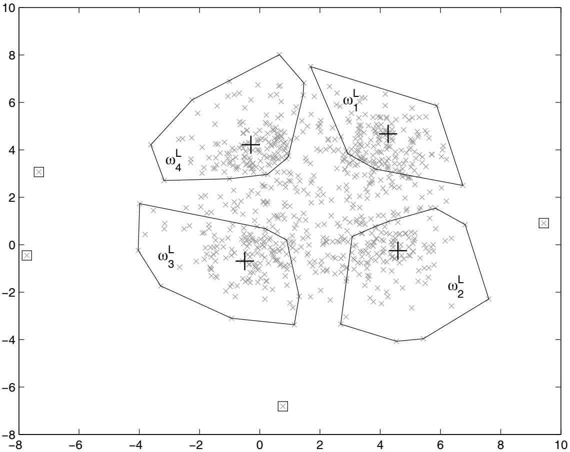
4-class data set

Hard evidential partition

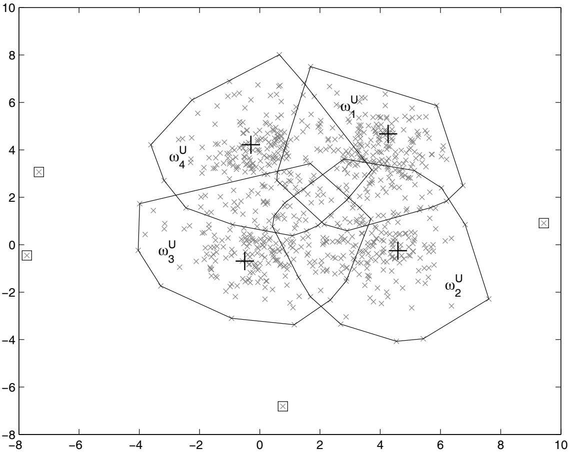


4-class data set

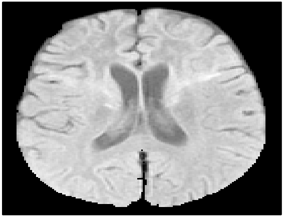
Lower approximations



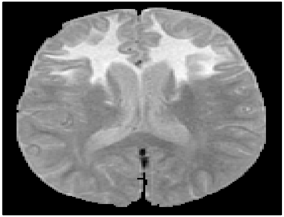
4-class data set
Upper approximations



Brain data
Problem



(a)

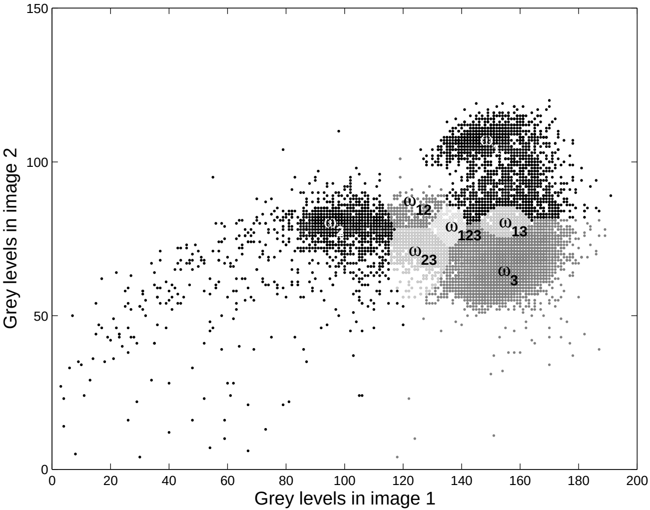


(b)

- Magnetic resonance imaging of pathological brain, 2 sets of parameters.
- Three regions: normal tissue (Norm), ventricles + cerebrospinal fluid (CSF/V) and pathology (Path).
- Image 1 highlights CSF/V (dark), image 2 highlights pathology (bright).

Brain data

Results in grey level space



Brain data
Image segmentation

Pathology (left); CSF and ventricles (center); normal brain tissues (right). The lower approximations of the clusters are represented by light grey areas, the upper approximations by the union of light and dark grey areas.

Determining the number of groups

- If a proper number of classes 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:

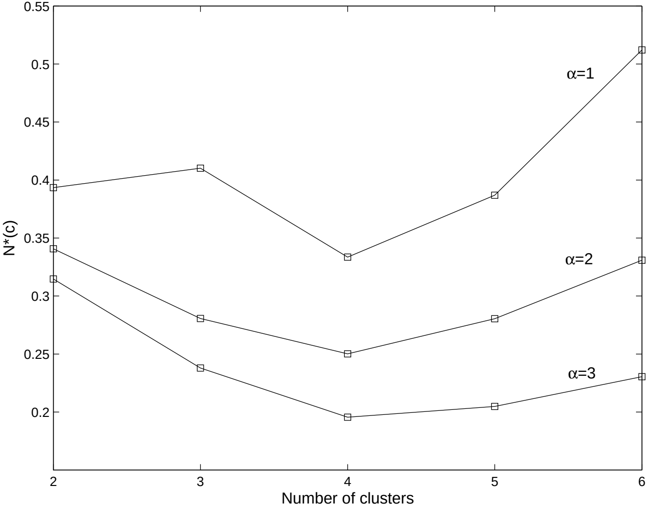
$$N(m) \triangleq \sum_{A \in 2^\Omega \setminus \emptyset} m(A) \log_2 |A| + m(\emptyset) \log_2 |\Omega|,$$

- Proposed validity index of an evidential partition:

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

Determining the number of groups

Result with the 4-class dataset



Conclusion
DS theory vs. Rough set theory

- Dempster-Shafer theory and Rough set theory have different agendas:
  - DS theory formalizes reasoning with uncertainty;
  - Rough set theory is a tool for knowledge extraction from databases.
- However, they are both concerned with coarseness of representation, and they have strong connections from a formal point of view:
  - A belief function Ω can be seen as being generated from a probability measure on some underlying space S and an interval relation between S and Ω.
  - The notions of lower and upper approximations of a set induced by an equivalence relation can be extended to mass functions.

utc Université de Technologie Compiègne

heudiasyc

Conclusion
Evidential vs. rough clustering

- When applied to clustering, DS theory leads to the notion of evidential partition, which generalizes most previous clustering structures, including rough clustering.
- Several algorithms have been proposed to generate an evidential partition from proximity or attribute data:
  - EVCLUS;
  - Evidential c-means and its variants (proximity data, optimized distance measure, etc.)
- These algorithms may also be used to generate a rough clustering structure.
- A detailed comparison with, e.g., the rough c-means algorithm (Lingras and West, 2004) remains to be done (see a first approach in Joshi and Lingras, 2012).

References

cf. `http://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.

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

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.

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.