Theory of belief functions: application to classification and clustering

Thierry Denœux[1]

[1]Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)

Erasmus University,
Rotterdam, April 2, 2009

Classification and clustering

Classical framework

- We consider a collection $\mathcal{L}$ of n objects.
- Each object is assumed to belong to one of c groups (classes).
- Each object is described by
  - An attribute vector $\mathbf{x} \in \mathbb{R}^p$ (attribute data), or
  - Its similarity to all other objects (proximity data).
- The class membership of objects may be:
  - Completely known, described by class labels (supervised learning);
  - Completely unknown (unsupervised learning);
  - Known for some objects, and unknown for others (semi-supervised learning).

Classification and clustering
Problems

- Classification: predict the class membership of objects drawn from the same population as $\mathcal{L}$.
- Clustering: Determine the class membership of objects in $\mathcal{L}$.

	supervised	unsupervised	semi-supervised
Classification	x		x
Clustering		x	x

Motivations

- In real situations, we may have only partial knowledge of class labels: we have uncertainty in the data $\rightarrow$ partially supervised learning.
- The class membership of objects can usually be predicted with some remaining uncertainty: the outputs from classification and clustering algorithms should reflect this uncertainty.
- The theory of belief functions provides a suitable framework for representing uncertain and imprecise class information as input and as output of classification and clustering algorithms.

Outline

1. Theory of belief functions
   - Representing evidence
   - Combining evidence
   - Making decisions
2. Classification: the evidential k-NN rule
   - Principle
   - Extension to partially supervised data
   - Examples
3. Clustering: learning a credal partition
   - Credal partition
   - EVCLUS
   - Evidential c-means

Theory of belief functions

- Introduced by Dempster (1968) and Shafer (1976), further developed by Smets (Transferable Belief Model) in the 1980's and 1990's. Also known as Dempster-Shafer theory or Evidence theory.
- A formal framework for representing and reasoning from partial (uncertain, imprecise) information.
- Generalizes both Set Theory and Probability Theory:
  - A belief function may be viewed both as a generalized set and as a non additive measure.
  - The theory includes extensions of probabilistic notions (conditioning, marginalization) and set-theoretic notions (intersection, union, inclusion, etc.)

Outline

Mass function

- Let X be a variable taking values in a finite set Ω (frame of discernment).
- Mass function: $m : 2^\Omega \to [0,1]$ such that

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

- Every A of Ω such that $m(A) > 0$ is a focal set of m.
- Interpretation: $m(A)$ represents is the probability of knowing only that $X \in A$, given the available evidence.
- $m(\Omega)$ is the probability of knowing nothing (ignorance).

Example

- A murder has been committed. There are three suspects: $\Omega = \{Peter, John, Mary\}$.
- A witness saw the murderer going away, but he is short-sighted and he only saw that it was a man, with 80 % confidence.
- This piece of evidence can be represented by

$$m(\{Peter, John\}) = 0.8,$$

$$m(\Omega) = 0.2$$

- The mass 0.2 is not committed to $\{Mary\}$, because the testimony does not accuse Mary at all!

Special cases

- m may be seen as:
  - A family of weighted sets $\{(A_i, m(A_i)), i = 1, \ldots, r\}$.
  - A generalized probability distribution (masses are distributed in 2^Ω instead of Ω).
- Special cases:
  - $r = 1$: categorical mass function ($\sim$ set). We denote by m_A the categorical mass function with focal set A.
  - $|A_i| = 1, i = 1, \ldots, r$: Bayesian mass function ($\sim$ probability distribution).

Belief function

- Definition:

$$bel(A) = \sum_{\substack{B \subseteq A \\ B \not\subseteq \overline{A}}} m(B) = \sum_{\emptyset \neq B \subseteq A} m(B), \quad \forall A \subseteq \Omega$$

- Interpretation: degree of belief (support) in hypothesis "$X \in A$".
- *bel* is superadditive. In particular,

$$bel(A \cup B) \geq bel(A) + bel(B) - bel(A \cap B).$$

Plausibility function

- Definition:

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B), \quad \forall A \subseteq \Omega$$

- Interpretation: upper bound on the degree of belief that could be assigned to A after taking into account new information.
- pl is subadditive. In particular,

$$pl(A \cup B) \leq pl(A) + pl(B) - pl(A \cap B).$$

- $bel \leq pl$.
- If m is Bayesian, $bel = pl$ (probability measure).

Example

A	$\emptyset$	$\{P\}$	$\{J\}$	$\{P,J\}$	$\{M\}$	$\{P,M\}$	$\{J,M\}$	Ω
$m(A)$	0	0	0	0.8	0	0	0	0.2
$bel(A)$	0	0	0	0.8	0	0	0	1
$pl(A)$	0	1	1	1	0.2	1	1	1

Relations between *m*, *bel* et *pl*

- Relations:

$$bel(A) = pl(\Omega) - pl(\overline{A}), \quad \forall A \subseteq \Omega$$

$$m(A) = \begin{cases} \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} bel(B), & A \neq \emptyset \\ 1 - bel(\Omega) & A = \emptyset \end{cases}$$

- *m*, *bel* et *pl* are thus three equivalent representations of a same piece of information.

Outline

Conditioning

- Let m represent our state of knowledge about X.
- We learn that $X \in B$ with $B \subset \Omega$.
- Impact on $m \rightarrow$ each mass $m(C)$ is transferred to $C \cap B$:

$$m(A|B) = \sum_{\{C|C \cap B = A\}} m(C).$$

- $m(\cdot|B)$ is a new mass function representing our state of knowledge based on m and the fact that $X \in B$.

Example

- We have $m(\{Peter, John\}) = 0.8$, $m(\Omega) = 0.2$.
- We learn that the murderer is blond. John and Mary are blond. $B = \{John, Mary\}$.
- $m(\{Peter, John\}) \rightarrow \{John\}$, $m(\Omega) \rightarrow \{John, Mary\}$.
- New conditional mass function given B.

$$m(\{John\}|B) = 0.8$$

$$m(\{John, Mary\}|B) = 0.2.$$

Properties

- Generalization of intersection: $m_A(\cdot|B) = m_{A\cap B}$.
- Generalisation of probabilistic conditioning:
  - If $m(\emptyset) > 0$, the normalized mass function m^* is

$$m^*(A) = \frac{m(A)}{1 - m(\emptyset)}.$$

  - Normalized conditioning:

$$pl^*(A|B) = \frac{pl(A \cap B)}{pl(B)}$$

  - If m is Bayesian, $pl = P$: same result as probabilistic conditioning.

Dempster's rule

Definition (Dempster's rule of combination)

Let m_1 and m_2 be mass functions induced by distinct (independent) items of evidence.

$$(m_1 \cap\!\!\!\!\bigcirc m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C), \quad \forall A \subseteq \Omega.$$

- Properties:
  - Generalization of conditioning: $m \cap\!\!\!\!\bigcirc m_B = m(\cdot|B)$.
  - Commutativity, associativity.
  - Neutral element: vacuous m_Ω such that $m_\Omega(\Omega) = 1$ (represents total ignorance).
- $K = (m_1 \cap\!\!\!\!\bigcirc m_2)(\emptyset) \geq 0$: degree of conflict.
- Other rules exist (disjunctive rule, cautious rule, etc...).

Example

- We have $m_1(\{Peter, John\}) = 0.8$, $m_1(\Omega) = 0.2$.
- New piece of evidence: the murderer is blond, confidence=0.6 $\rightarrow m_2(\{John, Mary\}) = 0.6$, $m_2(\Omega) = 0.4$.

	$\{Peter, John\}$ 0.8	Ω 0.2
$\{John, Mary\}$ 0.6	$\{John\}$ 0.48	$\{John, Mary\}$ 0.12
Ω 0.4	$\{Peter, John\}$ 0.32	Ω 0.08

Outline

Pignistic transformation

- Assume that our knowledge about X is represented by a mass function m, and we have to bet on the value of X.
- In order to avoid Dutch books (sequences of bets resulting sure loss), we have to base our decisions on a probability distribution on Ω.
- The pignistic transformation from m to a probability distribution $Betp$ can be justified axiomatically:

$$Betp(\omega) = \sum_{\{A \subseteq \Omega | \omega \in A\}} \frac{m^*(A)}{|A|}.$$

Example

- Let $m(\{John\}) = 0.48$, $m(\{John, Mary\}) = 0.12$, $m(\{Peter, John\}) = 0.32$, $m(\Omega) = 0.08$.
- We have

$$Betp(\{John\}) = 0.48 + \frac{0.12}{2} + \frac{0.32}{2} + \frac{0.08}{3} \approx 0.73,$$

$$Betp(\{Peter\}) = \frac{0.32}{2} + \frac{0.08}{3} \approx 0.19$$

$$Betp(\{Mary\}) = \frac{0.12}{2} + \frac{0.08}{3} \approx 0.09$$

Outline

Voting k-NN rule

- Classical non parametric classification method.
- Let Ω denote the set of classes, et $\mathcal{L}$ the learning set

$$\mathcal{L} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$$

with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \Omega$.
- Let $\mathbf{x} \in \mathbb{R}^p$ be the feature vector for a new object, and $\Phi_k(\mathbf{x})$ the set of the k nearest neighbors of $\mathbf{x}$ in $\mathcal{L}$ (according to some distance measure).
- Decision rule: $\mathbf{x}$ is assigned to the majority class in $\Phi_k(\mathbf{x})$.

Evidential k-NN rule (1/2)

- An alternative to the voting k-NN rule based on the theory of belief functions.
- Each $\mathbf{x}_i \in \Phi_k(\mathbf{x})$ is considered as a piece of evidence regarding the class of $\mathbf{x}$.
- The strength of this evidence decreases with the distance $d(\mathbf{x}, \mathbf{x}_i)$ between $\mathbf{x}$ and $\mathbf{x}_i$.
- It can be represented by a mass function

$$m_i(\{y_i\}) = \alpha \cdot \varphi\left(d(\mathbf{x}, \mathbf{x}_i)\right)$$

$$m_i(\Omega) = 1 - \alpha \cdot \varphi\left(d(\mathbf{x}, \mathbf{x}_i)\right).$$

where $\alpha \in (0, 1)$ is a constant, and φ is a decreasing function from $\mathbb{R}_+$ to $[0, 1]$ such that $\lim_{d \to +\infty} \varphi(d) = 0$.

Evidential k-NN rule (2/2)

- The evidence of the k nearest neighbors of $\mathbf{x}$ is pooled using Dempster's rule of combination:

$$m = \bigcirc\!\!\!\!\cap_{\mathbf{x}_i \in \Phi_k(\mathbf{x})} m_i.$$

- m encodes the evidence of the learning set regarding the class of the new object.
- Practical choice for φ: $\varphi(d) = \exp(-\gamma d^2)$.
- Parameters k, α and γ can be fixed heuristically or determined from the data using cross-validation.
- Decision:

$$\widehat{y} = \arg\max_{\omega \in \Omega} Betp(\omega).$$

Outline

Partially supervised data

- We now consider a learning set of the form

$$\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1, \ldots, n\}$$

  where
  - $\mathbf{x}_i$ is the attribute vector for object o_i, and
  - m_i is a mass function representing expert knowledge about the class y_i of object o_i.
- Special cases:
  - $m_i(\{\omega_k\}) = 1$: precise labeling (supervised learning);
  - $m_i(A) = 1$ for $A \subseteq \Omega$: imprecise (set-valued) labeling;
  - m_i is a Bayesian mass function: probabilistic labeling;

Extension of the evidential k-NN rule

- Each example $(\mathbf{x}_i, m_i)$ in $\mathcal{L}$ is an item of evidence regarding y, whose reliability decreases with the distance $d(\mathbf{x}, \mathbf{x}_i)$ between $\mathbf{x}$ and $\mathbf{x}_i$.
- Each mass function m_i is transformed (discounted) into a "weaker" mass function m_i':

$$m_i'(A) = \alpha \cdot \varphi\left(d(\mathbf{x}, \mathbf{x}_i)\right) m_i(A), \quad \forall A \subset \Omega.$$

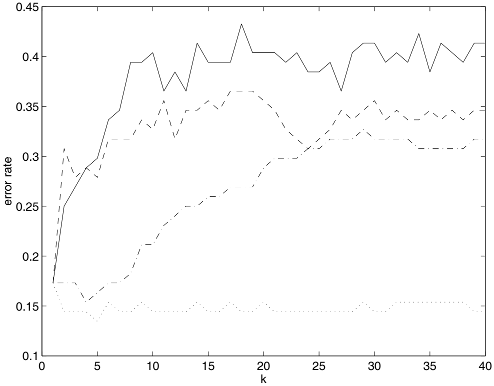
$$m_i'(\Omega) = 1 - \sum_{A \subset \Omega} m_i'(A).$$

- The k mass functions are combined using Dempster's rule:

$$m = \bigcirc\!\!\!\!\cap_{\mathbf{x}_i \in \Phi_k(\mathbf{x})} m_i'.$$

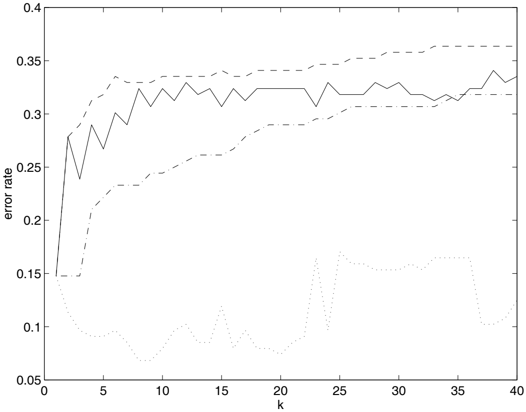
Outline

Example: Sonar data (UCI database)



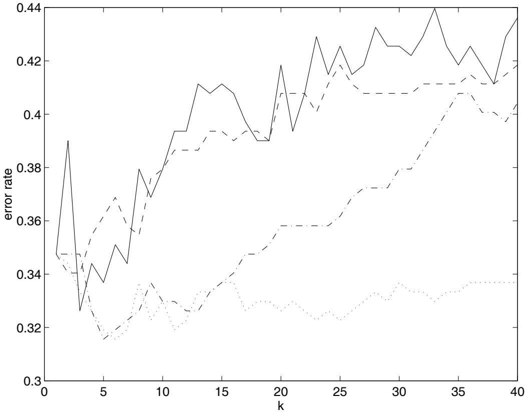
Test error rates as a function of k for the voting (-), evidential (:), fuzzy (–) and distance-weighted (-.) k-NN rules.

Example: Ionosphere data (UCI database)



Test error rates as a function of k for the voting (-), evidential (:), fuzzy (–) and distance-weighted (-.) k-NN rules.

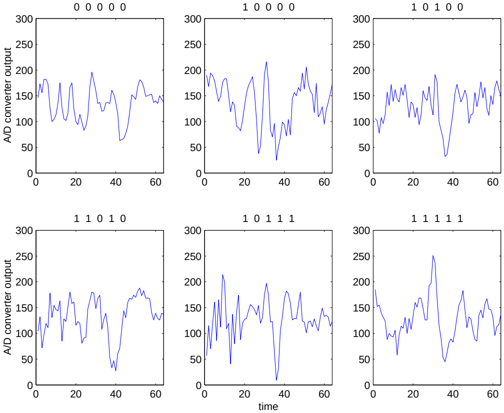
Example: Vehicle data (UCI database)



Test error rates as a function of k for the voting (-), evidential (:), fuzzy (–) and distance-weighted (-.) k-NN rules.

Example: EEG data

500 EEG signals encoded as 64-D patterns, 50 % positive (K-complexes), 50 % negative (delta waves), 5 experts.



Results on EEG data
(Denoeux and Zouhal, 2001)

- $c = 2$ classes, $d = 64$
- data labeled by 5 experts
- Consonant mass functions computed from empirical distribution of expert labels using a probability-possibility transformation.
- $n = 200$ learning patterns, 300 test patterns

k	k-NN	w k-NN	Ev. k-NN (crisp labels)	Ev. k-NN (uncert. labels)
9	0.30	0.30	0.31	0.27
11	0.29	0.30	0.29	0.26
13	0.31	0.30	0.31	0.26

Outline

Credal partition

- n objects described by attribute vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- Assumption: each object belongs to one of c classes in $\Omega = \{\omega_1, ..., \omega_c\}$,
- Goal: express our beliefs regarding the class membership of objects, in the form of mass functions $m_1, \ldots, m_n$ on Ω.
- Resulting structure = credal partition, generalizes hard and fuzzy partitions.

Example

A	$m_1(A)$	$m_2(A)$	$m_3(A)$	$m_4(A)$	$m_5(A)$
$\emptyset$	0	0	0	0	0
$\{\omega_1\}$	0	0	0	0.2	0
$\{\omega_2\}$	0	1	0	0.4	0
$\{\omega_1,\omega_2\}$	0.7	0	0	0	0
$\{\omega_3\}$	0	0	0.2	0.4	0
$\{\omega_1,\omega_3\}$	0	0	0.5	0	0
$\{\omega_2,\omega_3\}$	0	0	0	0	0
Ω	0.3	0	0.3	0	1

Special cases

- Each m_i is a *certain mass function*:

$$m_i(\{\omega_k\}) = 1 \text{ for some } k \in \{1, \ldots, c\}$$

$\rightarrow$ crisp partition of Ω.

- Each m_i is a *Bayesian mass function* (focal sets are singletons) $\rightarrow$ fuzzy partition of Ω

$$u_{ik} = m_i(\{\omega_k\}), \quad \forall i, k$$

$$\sum_{k=1}^{K} u_{ik} = 1.$$

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).

Outline

Proximity Data

Let $\mathcal{P}$ be a collection of n objects $\{o_i\}_{i=1}^n$. The observations consist in pairwise dissimilarities between objects:

$$\begin{array}{c|ccccc} & o_1 & \ldots & o_j & \ldots & o_n \\ \hline o_1 & & & \vdots & & \\ \vdots & & & \vdots & & \\ o_i & & \ldots & d_{ij} & \ldots & \\ \vdots & & & \vdots & & \\ o_n & & & & & \end{array}$$

Learning a Credal Partition from proximity data

- Problem: given th dissimilarity matrix $D = (d_{ij})$, how to build a “reasonable” credal partition ?
- Notion of cluster: objects within a cluster are assumed to be more similar among themselves than with objects from other clusters.
- Compatibility Principle: “The more similar two objects, the more plausible it is that they belong to the same class”.

Formalization

- Let S_{ij} be the event "objects o_i and o_j belong to the same class".
- Let m_i and m_j be mass functions regarding the class membership of objects o_i and o_j.
- It can be shown that

$$pl(S_{ij}) = \sum_{A \cap B \neq \emptyset} m_i(A) m_j(B) = 1 - K_{ij}$$

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

Cost function

- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict K_{ij}, up to an affine transformation (similar to Muldimensional Scaling).
- Example of stress functions:

$$I(M, a, b) = \sum_{i<j} \frac{(aK_{ij} + b - d_{ij})^2}{d_{ij}}$$

- Minimization of I with respect to M and a, b using a gradient-based iterative optimization procedure.

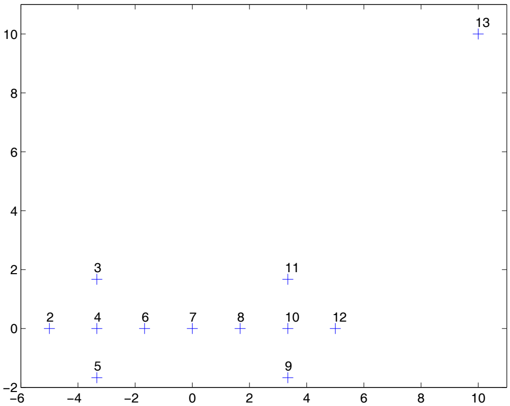
Reducing the complexity

- Learning a credal partition form data may be an ill-posed problem ($O(n2^c)$ parameters, $O(n^2)$ dissimilarities)).
- Solution:
  - Reduce the number of focal elements (e.g. $\{\omega_k\}_{k=1}^c$, $\emptyset$, and Ω)
  - Add constraints to the problem: penalize "uninformative", "complex" credal partitions

$$I' = I + \lambda \sum_{i=1}^{n} H(m_i)$$

where H=generalized entropy function.

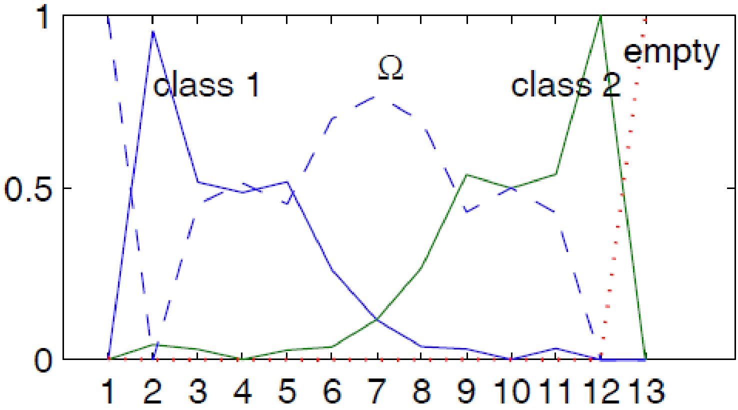
Experiments: Butterfly example
Data



one additional object (#1) similar to all other objects

Experiments: Butterfly example

Results



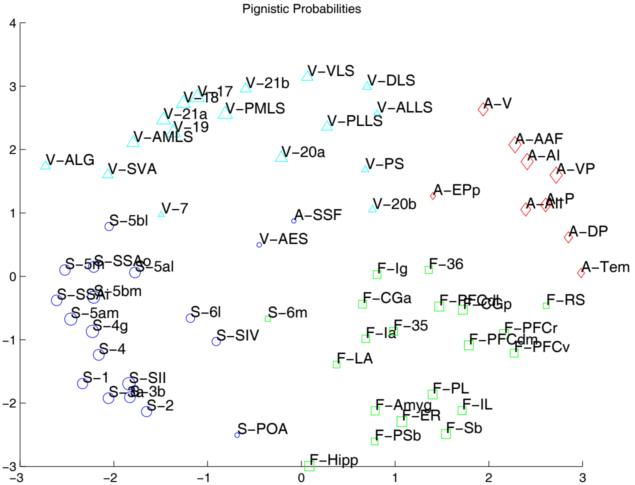
Experiments: Cat cortex dataset

Data

- Objects: 65 cortical areas
- Dissimilarities: connection strength between the cortical areas measured on an ordinal scale (0=self-connection,1=dense connection, 2=intermediate connection, 3=weak connection, 4=absence of connection)
- “True” partition: four functional regions of the cortex (A=auditory, V=visual, S=somatosensory, F=frontolimbic)
- Results:
  - only 3 misclassified regions out 64
  - similar to supervised kernel-based classification algorithms,
  - better than relational fuzzy clustering algorithms).

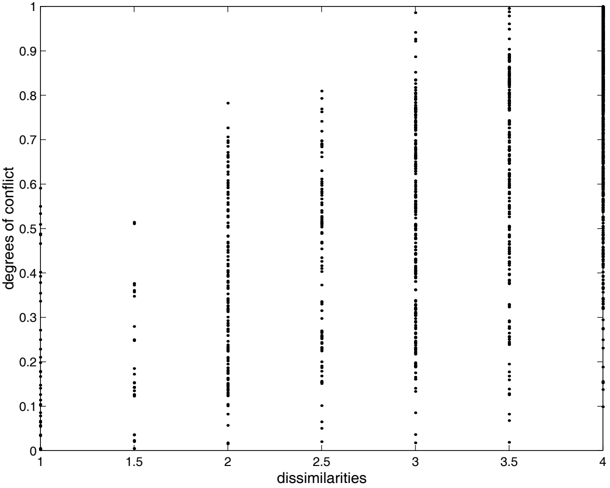
Experiments: Cat cortex dataset

Results



Experiments: Cat cortex dataset

Shepard diagram



Advantages and drawbacks

- Advantages
  - Applicable to proximity data (not necessarily Euclidean).
  - Robust against atypical observations (similar or dissimilar to all other objects).
  - Usually performs better than relational fuzzy clustering procedures.
- Drawback: computational complexity
  - One iteration of a gradient-based optimization procedure: $O(f^3 n^2)$ where $f =$ number of focal sets (usually $c + 2$).
  - Limited to datasets of a few hundred objects and less than 20 classes.
  - Not possible to use the full expressive power of belief functions (only $\{\omega_k\}$, $\emptyset$ and Ω as focal sets).

Outline

Principle

- Problem: generate a credal 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 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 $d_{ik}=||\mathbf{x}_i-\mathbf{v}_k||$ under the constraints $\sum_k u_{ik}=1$, $\forall i$.

- Alternate optimization algorithm:

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

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

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.

Optimization problem

- Minimize

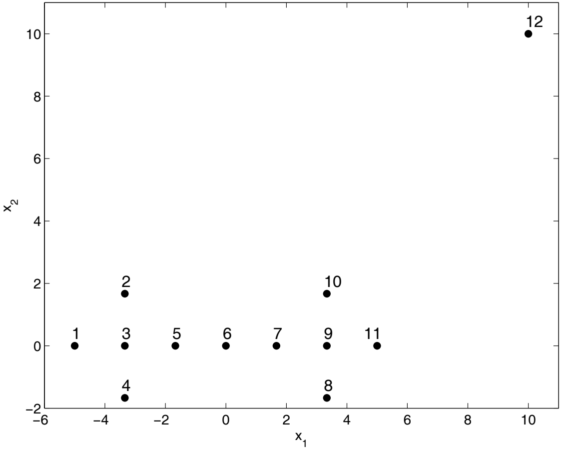
$$J_{\text{ECM}}(M,V) = \sum_{i=1}^{n} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^{2} + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta},$$

subject to

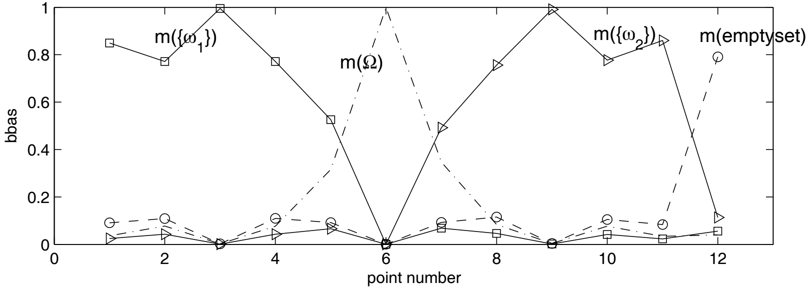
$$\sum_{\{j/A_j \subseteq \Omega, A_j \neq \emptyset\}} m_{ij} + m_{i\emptyset} = 1, \quad \forall i \in \{1,\ldots,n\},$$

- $J_{\text{ECM}}(M,V)$ can be iteratively minimized with respect to M and V using an alternate optimization scheme.

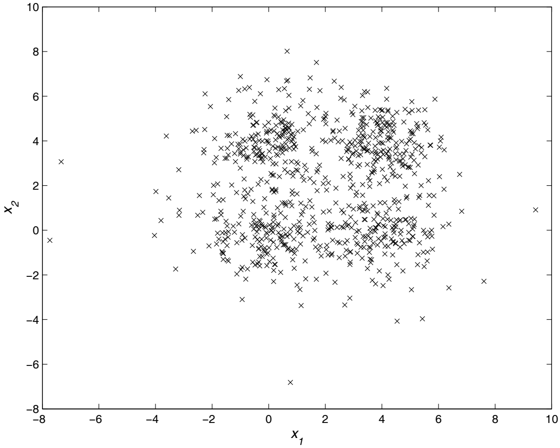
Butterfly dataset



Butterfly dataset
Results

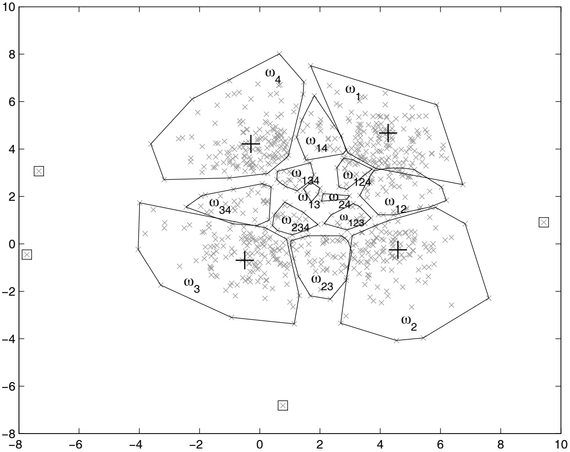


4-class data set



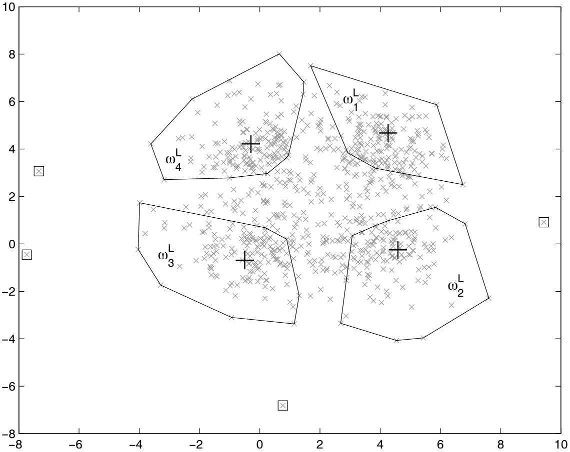
4-class data set

Hard credal partition



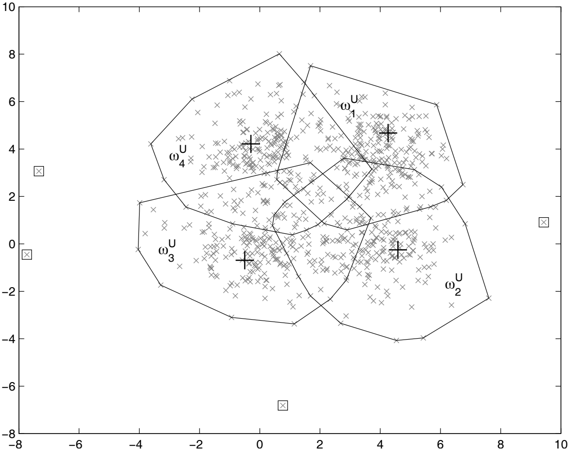
4-class data set

Lower approximation



4-class data set

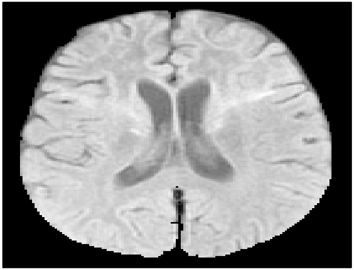
Upper approximation

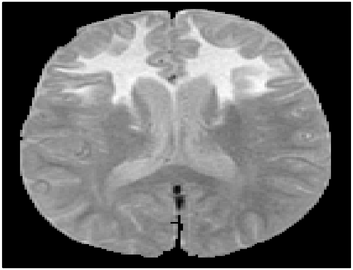


Brain data

Problem

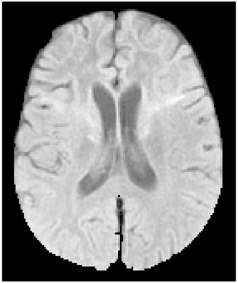
- Magnetic resonance imaging of pathological brain, 2 sets of parameters.
- Three regions: normal tissue (Norm), ventricals + cerebrospinal fluid (CSF/V) and pathology (Path).
- Image 1 highlights CSF/V (dark), image 2 highlights pathology (bright).



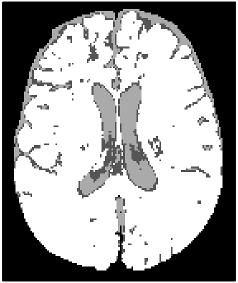


Brain data

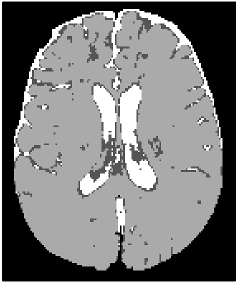
Segmentation of image 1



Initial image



γ_1 = CSF/V



γ_2 = Path ∪ normal

Image 1: 2 classes, coarsening of Ω:

$$\Gamma = \{\gamma_1 = CSF/V, \gamma_2 = \{Path, Normal\}\}$$

Brain data

Segmentation of image 2

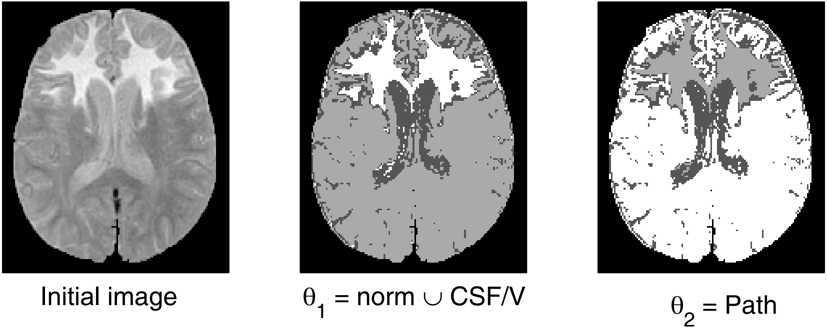


Image 2: 2 classes, coarsening of Ω:

$$\Theta = \{\theta_1 = Path, \theta_2 = \{CSF/V, Normal\}\}$$

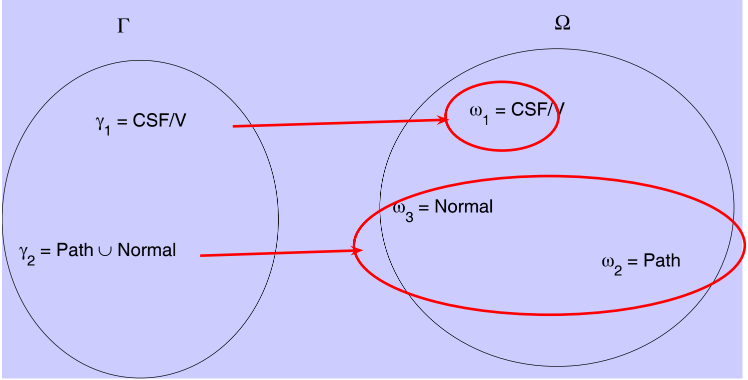
Brain data

Combining the two credal partitions

- Two credal partitions: for each pixel, two mass functions m_1 and m_2 on two different coarsenings of Ω.
- These two mass functions should be combined using Dempster's rule to recover the natural partition in three classes.
- m_1 and m_2 need first to be expressed on a common frame Ω (common refinement of Γ and Θ).

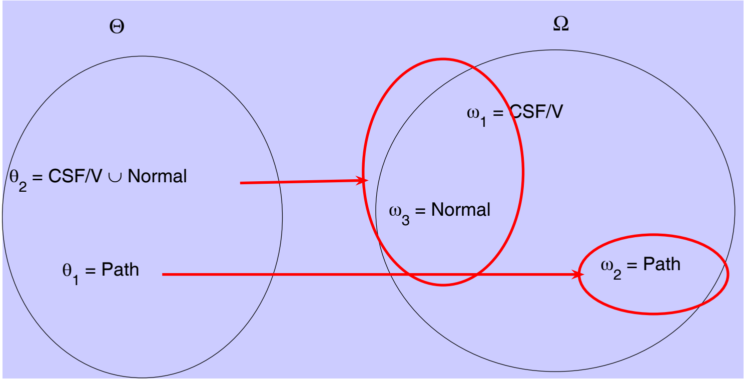
Brain data

Refinement of Γ



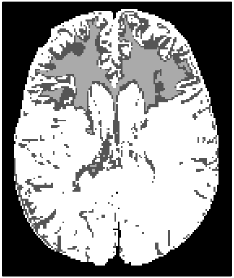
Brain data

Refinement of Θ

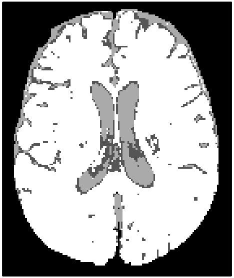


Brain data

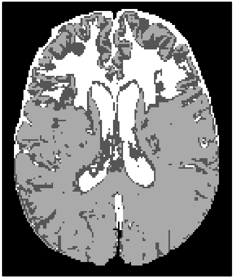
Final result after combination



ω_2 = Path



ω_1 = CSF/V



ω_3 = Normal

Conclusion

- The theory of belief functions extends both set theory and probability theory $\rightarrow$ it allows for the representation of imprecision and uncertainty.
- In classification and clustering, belief functions may be used to represent partial knowledge of class labels.
- Many classification and clustering algorithms can be adapted to
  - handle such class labels (partially supervised learning)
  - generate them from data (credal partition)

References I

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

T. Denœux.

A k-nearest neighbor classification rule based on Dempster-Shafer theory.

IEEE Transactions on Systems, Man and Cybernetics, 25(05):804-813, 1995.

L. M. Zouhal and T. Denoeux.

An evidence-theoretic k-NN rule with parameter optimization.

IEEE Transactions on Systems, Man and Cybernetics C, 28(2):263-271,1998.

References II
cf. http://www.hds.utc.fr/~tdenoeux

T. Denœux.
A neural network classifier based on Dempster-Shafer theory.
IEEE Transactions on Systems, Man and Cybernetics A, 30(2), 131-150, 2000.

T. Denoeux and M. Masson.
EVCLUS: Evidential Clustering of Proximity Data.
IEEE Transactions on Systems, Man and Cybernetics B, (34)1, 95-109, 2004.

References III

cf. `http://www.hds.utc.fr/~tdenoeux`

T. Denœux and P. Smets.
Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches.
IEEE Transactions on Systems, Man and Cybernetics B, 36(6), 1395-1406, 2006.

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

E. Côme, L. Oukhellou, T. Denoeux and P. Aknin.
Learning from partially supervised data using mixture models and belief functions.
Pattern Recognition, 42(3), 334-348, 2009.