Theory of belief functions

An introduction

Thierry Denœux[1]

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

Workshop on the Theory of Belief Functions
Brest, March 31, 2010

Objectives of this tutorial

1. Provide an introduction to the Theory of Belief Functions
2. Present some recent advances, with emphasis on information modeling in view of practical applications.

Outline

Outline

Uncertain reasoning

- In Information Technology, we often need to process and reason with information coming from various sources (sensors, experts, models, ...)
- Information is almost always tainted with various kinds of imperfection: imprecision, uncertainty, ambiguity,...
- We need a theoretical framework general enough to allow for the representation, propagation and combination of all kinds of imperfect information.
- The theory of belief functions is one such framework.

Imperfections of information

A typology (Dubois and Prade, 1988)

- Let X be a variable taking values in Ω (domain, frame of discernment).
- An item of information about X may be represented as a pair (value, confidence):
  - The "value" component corresponds to a subset of Ω;
  - The "confidence" component is an indication of the reliability of the item of information.
- Imprecision is related to the content of an item of information (the "value" component).
- Uncertainty is related to its conformity to a reality (the "confidence" component).

Imperfections of information

A simple example

- Let X = the temperature of this room
- "It is between 15 and 25 degrees" = $([15, 25]$, certain): certain but imprecise.
- "It is probably 20 degrees" = (20, probable): precise but uncertain.
- "It is probably between 15 and 25 degrees" = $([15, 25]$, probable): both uncertain and imprecise.

Classical frameworks

1. Set-membership approach
   - Interval analysis, bounded error estimation
   - Natural representation of information imprecision
   - Cannot express uncertainty (unreliability)
   - Lacks robustness, too conservative.
2. Probability theory
   - Well-suited for modeling aleatory uncertainty (variability in a population or across repetitions of a random experiment).
   - Does not express any notion of imprecision.

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 the Set-membership approach 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.)

Mass function

Definition

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

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

- Every A of Ω such that $m(A) > 0$ is a focal set of m. Let $A_1, \ldots, A_r$ be focal sets.
- 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 (probability) mass function.

Mass function

Multi-valued mapping interpretation

- A mass function m on Ω may be viewed as arising from
  - A set $\Theta = \{\theta_1, \ldots, \theta_r\}$ of interpretations of the available evidence;
  - A probability measure P on Θ;
  - A multi-valued mapping $\Gamma : \Theta \rightarrow 2^\Omega$.
- Meaning:
  - Under interpretation θ_i, the evidence tells us that $X \in \Gamma(\theta_i)$, and nothing more.
  - The probability $P(\{\theta_i\})$ is transferred to $A_i = \Gamma(\theta_i)$: $m(A_i) = P(\{\theta_i\})$
- In this framework, $m(A)$ may be then viewed as the probability of knowing only that $X \in A$, given the available evidence.
- In particular, $m(\Omega)$ is the probability of knowing nothing.

Example

- A murder has been committed. There are three suspects: $\Omega = \{Peter, John, Mary\}$.
- A witness saw the murderer going away in the dark, and he can only assert that it was man. How, we know that the witness is drunk 20 % of the time.
- 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!

Belief, plausibility

- Belief function:

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

- Plausibility function:

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

- Interpretations:
  - $bel(A) =$ degree to which the evidence supports A.
  - $pl(A) =$ upper bound on the degree of support that could be assigned to A after taking into account new information ($\geq bel(A)$).
- 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 piece of evidence or, equivalently,
    - a state of belief induced by this evidence.

Relationship with Possibility theory

- Assume that the focal sets of m are nested: $A_1 \subset A_2 \subset \ldots \subset A_r \rightarrow m$ is said to be consonant.
- The following relations hold:

$$pl(A \cup B) = \max\left(pl(A), pl(B)\right), \quad \forall A, B \subseteq \Omega.$$

- *pl* is this a possibility measure, and *bel* is the dual necessity measure.
- The possibility distribution is the contour function:

$$\pi(x) = pl(\{x\}), \quad \forall x \in \Omega.$$

- The theory of belief function can thus be considered as more expressive than possibility theory.

Outline

Conditioning

Definitions

- Let m be a mass function on Ω representing some evidence about X.
- Additional evidence tells us that $X \in B$ for sure. How to update m?
- Two basic rules:
  1. Unnormalized conditioning:
  $$m(A|B) = \sum_{\{C|C\cap B=A\}} m(C).$$
  2. Normalized conditioning:
  $$m^*(A|B) = \begin{cases} \frac{m(A|B)}{1-m(\emptyset|B)} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases}$$

Conditioning

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.

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

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

Conditioning

Justification using the multi-valued mapping interpretation

- Assume that m is induced by a probability measure on Θ and a multi-valued mapping $\Gamma : \Theta \rightarrow 2^{\Omega}$.
- After knowing that $X \in B$, each interpretation θ_i that pointed to $A_i = \Gamma(\theta_i)$ now points to $A_i \cap B$.
- New mapping $\Gamma_B : \theta_i \rightarrow A_i \cap B$.
- What to do with θ_is such that $\Gamma_B(\theta_i) = \emptyset$?
  1. Discard them and condition P on the remaining one: normalized rule of conditioning (Dempster's rule of conditioning).
  2. Keep them to keep track of the conflict between pieces of evidence. $\rightarrow$ unnormalized rule of conditioning.

Conditioning

Properties

- Extension of set intersection:

$$m_A(\cdot|B) = m_{A\cap B}.$$

- Extension of Bayesian conditioning:
  - Expression of normalized conditioning in terms of plausibility:

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

  - If m is Bayesian, pl is a probability measure: probabilistic conditioning is recovered.

Plausibility, communality

- Interpretation of $pl(A)$:
  - $pl(A) = bel(A|A) = \max_B bel(A|B)$
  - maximal degree of support that can be assigned to A after conditioning.
- Commonality function: let $q : 2^\Omega \rightarrow [0,1]$ be defined as $q(A) = m(A|A)$:
  - Mass attached to the largest possible subset of Ω (degree of ignorance) after conditioning on A.
  - Other expression:

$$q(A) = \sum_{B \supseteq A} m(B).$$

Conjunctive combination

Definitions

- Let m_1 and m_2 be two mass functions on Ω induced by two distinct items of evidence. How should they be combined?
- Two basic conjunctive operators:
  1. TBM conjunctive rule

$$(m_1 \bigcirc\!\!\!\!\cap\, m_2)(A) = \sum_{B \cap C = A} m_1(B) m_2(C)$$

  2. Dempster's rule

$$(m_1 \oplus m_2)(A) = \begin{cases} \frac{(m_1 \bigcirc\!\!\!\!\cap\, m_2)(A)}{1 - K_{12}} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases}$$

  with $K_{12} = (m_1 \bigcirc\!\!\!\!\cap\, m_2)(\emptyset)$: degree of conflict.

Conjunctive combination

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

Conjunctive rules

Justification using the multi-valued mapping interpretation

- Let $(\Theta_1, P_1, \Gamma_1)$ and $(\Theta_2, P_2, \Gamma_2)$ be the multi-valued mapping frameworks associated to the two pieces of evidence.
- If interpretations $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ both hold, then we can conclude that $X \in \Gamma_1(\theta_1) \cap \Gamma_2(\theta_2)$.
- If the two pieces of evidence are independent, then this happens with probability $P_1(\{\theta_1\})P_2(\{\theta_2\})$.
- Two solutions:
  1. Transfer the mass $P_1(\{\theta_1\})P_2(\{\theta_2\})$ to $\Gamma_1(\theta_1) \cap \Gamma_2(\theta_2)$: TBM conjunctive rule;
  2. First, discard inconsistent interpretations (θ_1, θ_2) such that $\Gamma_1(\theta_1) \cap \Gamma_2(\theta_2) = \emptyset$ and condition the probability on $\Theta_1 \times \Theta_2$ on the remaining ones: Dempster's rule.

Dempster's rules

Properties

- Generalization of conditioning:

$$m(\cdot|B) = m \textcircled{\cap} m_B, \quad m^*(\cdot|B) = m \oplus m_B$$

- Both $\textcircled{\cap}$ and $\oplus$ are commutative and associative
- Neutral element:

$$m \textcircled{\cap} m_\Omega = m \oplus m_\Omega = m.$$

- $(q_1 \textcircled{\cap} q_2) = q_1 \cdot q_2$

TBM disjunctive rule

Definition and justification

- Let $(\Theta_1, P_1, \Gamma_1)$ and $(\Theta_2, P_2, \Gamma_2)$ be the multi-valued mapping frameworks associated to two pieces of evidence.
- If interpretation $\theta_k \in \Theta_k$ holds and piece of evidence k is reliable, then we can conclude that $X \in \Gamma_k(\theta_k)$.
- If interpretation $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ both hold and we assume that at least one of the two pieces of evidence is reliable, then we can conclude that $X \in \Gamma_1(\theta_1) \cup \Gamma_2(\theta_2)$.
- This leads to the TBM disjunctive rule:

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

TBM disjunctive rule

Properties

- commutativity, associativity.
- neutral element: $m_\emptyset$
- Let $b = bel + m(\emptyset)$ (implicability function). We have:

$$(b_1 \textcircled{\cup} b_2) = b_1 \cdot b_2$$

- De Morgan laws for $\textcircled{\cap}$ and $\textcircled{\cup}$:

$$\begin{aligned} \overline{m_1 \textcircled{\cup} m_2} &= \overline{m_1} \textcircled{\cap} \overline{m_2}, \\ \overline{m_1 \textcircled{\cap} m_2} &= \overline{m_1} \textcircled{\cup} \overline{m_2}, \end{aligned}$$

where $\overline{m}$ denotes the complement of m defined by $\overline{m}(A) = m(\overline{A})$ for all $A \subseteq \Omega$.

Selecting a combination rule

- All three rules $\textcircled{\cap}$, $\oplus$ and $\textcircled{\cup}$ assume the pieces of evidence to be independent.
- The conjunctive rules $\textcircled{\cap}$ and $\oplus$ further assume that the pieces of evidence are both reliable;
- The TBM disjunctive rule $\textcircled{\cup}$ only assumes that at least one of the items of evidence combined is reliable (weaker assumption).
- $\textcircled{\cap}$ vs. $\oplus$:
  - $\textcircled{\cap}$ keeps track of the conflict between items of evidence: very useful in some applications.
  - $\textcircled{\cap}$ also makes sense under the open-world assumption.
  - The conflict increases with the number of combined mass functions: normalization is often necessary at some point.
- What to do with dependent items of evidence? $\rightarrow$ Cautious rule

Outline

Operations in product frames

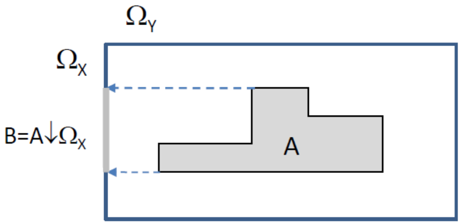
Notations

- In many applications, we need to express uncertain information about several variables taking values in different domains.
- Let X and Y be two variables defined on frames Ω_X and Ω_Y.
- Let $\Omega_{XY} = \Omega_X \times \Omega_Y$ be the product frame.
- A mass function $m^{\Omega_{XY}}$ on Ω_{XY} can be seen as an uncertain relation between variables X and Y.
- Two basic operations on product frames:
    1. Express a joint mass function $m^{\Omega_{XY}}$ in the coarser frame Ω_X or Ω_Y (marginalization);
    2. Express a marginal mass function m^{Ω_X} on Ω_X in the finer frame Ω_{XY} (vacuous extension).

Operations in product frames

Marginalization

- Problem: express $m^{\Omega_{XY}}$ in Ω_X.
- Solution: transfer each mass $m^{\Omega_{XY}}(A)$ to the projection of A on Ω_X:



- Marginal mass function

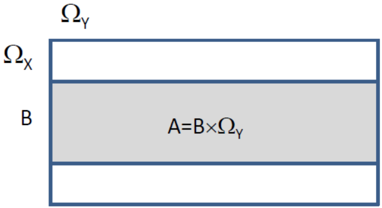
$$m^{\Omega_{XY}\downarrow\Omega_X}(B) = \sum_{\{A\subseteq\Omega_{XY}, A\downarrow\Omega_X = B\}} m^{\Omega_{XY}}(A) \quad \forall B \subseteq \Omega_X.$$

- Generalizes both set projection and probabilistic marginalization.

Operations in product frames

Vacuous extension

- Problem: express m^{Ω_X} in Ω_{XY}.
- Solution: transfer each mass $m^{\Omega_X}(B)$ to the cylindrical extension of B: $B \times \Omega_Y$.



- Vacuous extension:

$$m^{\Omega_X \uparrow \Omega_{XY}}(A) = \begin{cases} m^{\Omega_X}(B) & \text{if } A = B \times \Omega_Y \\ 0 & \text{otherwise.} \end{cases}$$

Operations in product frames

Application to approximate reasoning

- Assume that we have:
  - Partial knowledge of X formalized as a mass function m^{Ω_X};
  - A joint joint mass function $m^{\Omega_{XY}}$ representing an uncertain relation between X and Y.
- What can we say about Y?
- Solution:
  1. Vacuously extend m^{Ω_X} to Ω_{XY};
  2. Combine $m^{\Omega_X \uparrow \Omega_{XY}}$ with $m^{\Omega_{XY}}$;
  3. Marginalize the result on Ω_Y.
- Formally:

$$m^{\Omega_Y} = \left(m^{\Omega_X \uparrow \Omega_{XY}} \bigcirc\!\!\!\!\cap\, m^{\Omega_{XY}} \right)^{\downarrow \Omega_X}.$$

Outline

Decision making

Problem formulation

- A decision problem can be formalized by defining:
  - A set of acts $\mathcal{A} = \{a_1, \ldots, a_s\}$;
  - A set of states of the world Ω;
  - A loss function $L : \mathcal{A} \times \Omega \rightarrow \mathbb{R}$, such that $L(a, \omega)$ is the loss incurred if we select act a and the true state of the world is ω.
- Bayesian framework
  - Uncertainty on Ω is described by a probability measure P;
  - Define the risk of each act a as the expected loss if a is selected: $R(a) = \mathbb{E}_P[L(a, \cdot)]$.
  - Select an act with minimal risk.
- Extension to the belief function framework?

Decision making

TBM solution

- In order to avoid Dutch books (sequences of bets resulting in sure loss), we have to base our decisions on a probability distribution on Ω.
- The TBM postulates that uncertain reasoning and decision making are two fundamentally different operations occurring at two different levels:
  - Uncertain reasoning is performed at the credal level using the formalism of belief functions.
  - Decision making is performed at the pignistic level, after the m on Ω has been transformed into a probability measure.
- The pignistic transformation from m to a probability mass function $Betp$ can be justified axiomatically:

$$Betp(\omega) = \sum_{A \subseteq \Omega} \frac{m(A)}{1 - m(\emptyset)} \frac{1_A(\omega)}{|A|}, \quad \forall \omega \in \Omega.$$

Decision making

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

Informational comparison of belief functions

- Let m_1 et m_2 be two mass functions on Ω.
- In what sense can we say that m_1 is more informative (committed) than m_2?
- Special case:
  - Let m_A and m_B be two categorical mass functions.
  - m_A is more committed than m_B iff $A \subseteq B$.
- Extension to arbitrary mass functions?

Plausibility and commonality orderings

- m_1 is pl-more committed than m_2 (noted $m_1 \sqsubseteq_{pl} m_2$) if

$$pl_1(A) \leq pl_2(A), \quad \forall A \subseteq \Omega.$$

- m_1 is q-more committed than m_2 (noted $m_1 \sqsubseteq_q m_2$) if

$$q_1(A) \leq q_2(A), \quad \forall A \subseteq \Omega.$$

- Properties:
  - Extension of set inclusion:

$$m_A \sqsubseteq_{pl} m_B \Leftrightarrow m_A \sqsubseteq_q m_B \Leftrightarrow A \subseteq B.$$

  - Greatest element: m_Ω t.q. $m_\Omega(\Omega) = 1$ (vacuous mass function).

Strong (specialization) ordering

- m_1 is a specialization of m_2 (noted $m_1 \sqsubseteq_s m_2$) if m_1 can be obtained from m_2 by distributing each mass $m_2(B)$ to subsets of B:

$$m_1(A) = \sum_{B \subseteq \Omega} S(A, B) m_2(B), \quad \forall A \subseteq \Omega,$$

  where $S(A, B) =$ proportion of $m_2(B)$ transferred to $A \subseteq B$.
- S: specialization matrix.
- Properties:
    - Extension of set inclusion;
    - Greatest element: m_Ω;
    - $m_1 \sqsubseteq_s m_2 \Rightarrow m_1 \sqsubseteq_{pl} m_2$ and $m_1 \sqsubseteq_q m_2$.

Least Commitment Principle

Definition

Definition (Least Commitment Principle)

When several belief functions are compatible with a set of constraints, the least informative according to some informational ordering (if it exists) should be selected.

A very powerful method for constructing belief functions!

Outline

Cautious rule

Motivations

- The standard rules $\textcircled{\cap}$, $\oplus$ and $\textcircled{\cup}$ assume the sources of information to be independent, e.g.
  - experts with non overlapping experience/knowledge;
  - non overlapping datasets.
- What to do in case of non independent evidence?
  - Describe the nature of the interaction between sources (difficult, requires a lot of information);
  - Use a combination rule that tolerates redundancy in the combined information.
- Such rules can be derived from the LCP using suitable informational orderings.

Cautious rule

Principle

- Two sources provide mass functions m_1 and m_2, and the sources are both considered to be reliable.
- After receiving these m_1 and m_2, the agent's state of belief should be represented by a mass function m_{12} more committed than m_1, and more committed than m_2.
- Let $\mathcal{S}_x(m)$ be the set of mass functions m' such that $m' \sqsubseteq_x m$, for some $x \in \{pl, q, s, \cdots\}$. We thus impose that $m_{12} \in \mathcal{S}_x(m_1) \cap \mathcal{S}_x(m_2)$.
- According to the LCP, we should select the x-least committed element in $\mathcal{S}_x(m_1) \cap \mathcal{S}_x(m_2)$, if it exists.

Cautious rule

Problem

- The above approach works for special cases.
- Example (Dubois, Prade, Smets 2001): if m_1 and m_2 are consonant, then the q-least committed element in $\mathcal{S}_q(m_1) \cap \mathcal{S}_q(m_2)$ exists and it is unique: it is the consonant mass function with commonality function $q_{12} = q_1 \wedge q_2$.
- In general, neither existence nor uniqueness of a solution can be guaranteed with any of the x-orderings, $x \in \{pl, q, s\}$.
- We need to define a new ordering relation.
- This ordering will be based on the (conjunctive) canonical decomposition of belief functions.

Canonical decomposition

Simple and separable mass functions

- Definition: m is simple mass function if it has the following form

$$\begin{aligned} m(A) &= 1 - w_A \\ m(\Omega) &= w_A, \end{aligned}$$

with $A \subset \Omega$ and $w_A \in [0, 1]$.

- Notation: A^{w_A}.
- Property: $A^{w_1} \bigcirc\!\!\!\!\cap\, A^{w_2} = A^{w_1 w_2}$.
- A mass function is separable if it can be written as the combinaison of simple mass functions:

$$m = \bigcirc\!\!\!\!\cap_{A \subset \Omega} A^{w(A)}$$

with $0 \leq w(A) \leq 1$ for all $A \subset \Omega$.

Canonical decomposition

Subtracting evidence

- Let $m_{12} = m_1 \textcircled{\cap} m_2$. We have $q_{12} = q_1 \cdot q_2$.
- Assume we no longer trust m_2 and we wish to subtract it from m_{12}.
- If m_2 is non dogmatic (i.e. $m_2(\Omega) > 0$ or, equivalently, $q_2(A) > 0, \forall A$), m_1 can be retrieved as

$$q_1 = q_{12}/q_2.$$

- We note $m_1 = m_{12} \textcircled{\not\cap} m_2$.
- Remark: $m_1 \textcircled{\not\cap} m_2$ may not be a valid mass function!

Canonical decomposition

Theorem (Smets, 1995)

Any non dogmatic mass function ($m(\Omega) > 0$) can be canonically decomposed as:

$$m = \left(\textcircled{\cap}_{A \subset \Omega} A^{w_C(A)} \right) \oslash \left(\textcircled{\cap}_{A \subset \Omega} A^{w_D(A)} \right)$$

with $w_C(A) \in (0,1]$, $w_D(A) \in (0,1]$ and $\max(w_C(A), w_D(A)) = 1$ for all $A \subset \Omega$.

- Let $w = w_C / w_D$.
- Function $w : 2^\Omega \setminus \Omega \rightarrow \mathbb{R}_+^*$ is called the (conjunctive) weight function.
- It is a new equivalent representation of a non dogmatic mass function (together with *bel*, *pl*, *q*, *b*).

Properties of w

- Function w is directly available when m is built by accumulating simple mass function (common situation).
- Calculation of w from q:

$$\ln w(A) = -\sum_{B \supseteq A} (-1)^{|B|-|A|} \ln q(B), \quad \forall A \subset \Omega.$$

- Conversely,

$$\ln q(A) = -\sum_{\Omega \supset B \not\supseteq A} \ln w(B), \quad \forall A \subseteq \Omega$$

- TBM conjunctive rule:

$$w_1 \bigcirc\!\!\!\!\cap\, w_2 = w_1 \cdot w_2.$$

w-ordering

- Let m_1 and m_2 be two non dogmatic mass functions. We say that m_1 is w-more committed than m_2 (denoted as $m_1 \sqsubseteq_w m_2$) if $w_1 \leq w_2$.
- Interpretation: $m_1 = m_2 \textcircled{\cap} m$ with m separable.
- Properties:
  - $m_1 \sqsubseteq_w m_2 \Rightarrow m_1 \sqsubseteq_s m_2 \Rightarrow \begin{cases} m_1 \sqsubseteq_{pl} m_2 \\ m_1 \sqsubseteq_q m_2, \end{cases}$
  - m_Ω is the only maximal element of $\sqsubseteq_w$:

$$m_\Omega \sqsubseteq_w m \Rightarrow m = m_\Omega.$$

Cautious rule

Definition

Theorem

Let m_1 and m_2 be two nondogmatic BBAs. The w-least committed element in $\mathcal{S}_w(m_1) \cap \mathcal{S}_w(m_2)$ exists and is unique. It is defined by the following weight function:

$$w_{1 \owedge 2}(A) = w_1(A) \wedge w_2(A), \quad \forall A \subset \Omega.$$

Definition (cautious conjunctive rule)

$$m_1 \owedge m_2 = \bigcirc\!\!\!\!\cap_{A \subset \Omega} A^{w_1(A) \wedge w_2(A)}.$$

Cautious rule

Definition

Theorem

Let m_1 and m_2 be two nondogmatic BBAs. The w-least committed element in $\mathcal{S}_w(m_1) \cap \mathcal{S}_w(m_2)$ exists and is unique. It is defined by the following weight function:

$$w_{1 \textcircled{\wedge} 2}(A) = w_1(A) \wedge w_2(A), \quad \forall A \subset \Omega.$$

Definition (cautious conjunctive rule)

$$m_1 \textcircled{\wedge} m_2 = \textcircled{\cap}_{A \subset \Omega} A^{w_1(A) \wedge w_2(A)}.$$

Cautious rule computation

m-space		w-space
m_1	$\longrightarrow$	w_1
m_2	$\longrightarrow$	w_2
$m_1 \textcircled{\wedge} m_2$	$\longleftarrow$	$w_1 \wedge w_2$

Cautious rule

Properties

- Commutative, associative
- Idempotent : $\forall m$, $m \circledwedge m = m$
- Distributivity of $\circledcap$ with respect to $\circledwedge$:

$$(m_1 \circledcap m_2) \circledwedge (m_1 \circledcap m_3) = m_1 \circledcap (m_2 \circledwedge m_3), \forall m_1, m_2, m_3.$$

The same item of evidence m_1 is not counted twice!

- No neutral element, but $m_\Omega \circledwedge m = m$ iff m is separable.

Related rules

- Normalized cautious rule:

$$(m_1 \owedge^* m_2)(A) = \begin{cases} \frac{(m_1 \owedge m_2)(A)}{1-(m_1 \owedge m_2)(\emptyset)} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset. \end{cases}$$

- Bold disjunctive rule:

$$m_1 \ovee m_2 = \overline{\overline{m_1} \owedge \overline{m_2}}.$$

- Both $\owedge^*$ and $\ovee$ are commutative, associative and idempotent.

Global picture

- Six basic rules:

Sources		independent	dependent
All reliable	open world	$\textcircled{\cap}$	$\textcircled{\wedge}$
	closed world	$\oplus$	$\textcircled{\wedge}^*$
At least one reliable		$\textcircled{\cup}$	$\textcircled{\vee}$

Outline

Belief functions on real numbers

Definitions

- Belief functions can be defined on continuous frames such as $\mathbb{R}$.
- Simplest model: masses are assigned to (closed) intervals (Dempster, 1968).
- Two basic cases:
  - Discrete case: masses are assigned to a finite set of focal intervals;
  - Continuous case: masses are assigned to intervals using a mass density function.

Discrete mass functions

Definitions

- A function m from the set $\mathcal{I}$ of real intervals to $[0,1]$ is a discrete mass function if there exist
  - r intervals $I_1, \ldots, I_r$
  - r positive numbers $m_1, \ldots, m_r$ verifying $\sum_{i=1}^r m_i = 1$

  such that $m(I_i) = m_i$ for all $i \in \{1, \ldots, r\}$ and $m(I) = 0$ for all other $I \in \mathcal{I}$.
- Belief, commonality and plausibility functions:

$$bel(A) = \sum_{\emptyset \neq I_i \subseteq A} m_i, \quad pl(A) = \sum_{I_i \cap A \neq \emptyset} m_i,$$

$$q(A) = \sum_{I_i \supseteq A} m_i,$$

for all $A \subseteq \mathbb{R}$.

Discrete mass functions

Combination and pignistic probability

- Combination using the TBM conjunctive rule:

$$(m \textcircled{\cap} m')(I) = \sum_{\{i,j | I_i \cap I'_j = I\}} m_i \cdot m'_j.$$

- Assuming $0 < |I_i| < +\infty$ for all i, the pignistic probability density associated to m is:

$$Betp(x) = \sum_{i=1}^{r} m_i \frac{1_{I_i}(x)}{|I_i|}, \quad \forall x \in \mathbb{R}.$$

(*Betp* is a finite mixture of continuous uniform distributions.)

Discrete mass functions

Extension of interval arithmetics

- Interval arithmetics is a powerful tool for propagating imprecision in numerical equations.
- If $*$ is a continuous binary operator (e.g., an arithmetic operation) the set

$$[x] * [y] = \{x * y \in \mathbb{R} | x \in [x], y \in [y]\}.$$

  is an interval.
- Arithmetic operations (and other elementary functions) can thus be extended to intervals.
- Examples:

$$\begin{aligned}
[x] + [y] &= [\underline{x} + \underline{y}, \overline{x} + \overline{y}] \\
[x] - [y] &= [\underline{x} - \overline{y}, \overline{x} - \underline{y}] \\
[x] \cdot [y] &= [\min(\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y}), \max(\underline{x}\underline{y}, \underline{x}\overline{y}, \overline{x}\underline{y}, \overline{x}\overline{y})].
\end{aligned}$$

Discrete mass functions

Extension of interval arithmetics

- Let us consider three variables X, Y and Z linked by the relation: $Z = X * Y$.
- Assume that the evidence on X and Y is modeled by discrete mass functions m^X and m^Y with closed focal intervals.
- If the items of evidence regarding X and Y are independent, then uncertainty on X is represented by the following mass function:

$$m^Z([z]) = \sum_{\{i,j | [x_i]*[y_j]=[z]\}} m^X([x_i]) m^Y([y_j]), \quad \forall [z].$$

- Discrete mass functions can be propagated in more complex numerical equations using Interval Analysis techniques.

Extension of interval arithmetics

Example

- Assume that:
  - $m^X([1,2]) = 0.8$, $m^X([0,3]) = 0.2$;
  - $m^Y([4,5]) = 0.6$, $m^Y([0,10]) = 0.4$;
  - $Z = X + Y$.
- A mass function on Z can be computed as:

	$[1,2]$ 0.8	$[0,3]$ 0.2
$[4,5]$ 0.6	$[5,7]$ 0.48	$[4,8]$ 0.12
$[0,10]$ 0.4	$[1,12]$ 0.32	$[0,13]$ 0.08

Belief functions on real numbers

Continuous case

- A (normalized) mass density function m on $\mathbb{R}$ is defined as

$$m([u,v]) = f(u,v), \quad \forall u \leq v,$$

where f is a pdf with support in $\{(u,v) \in \mathbb{R}^2 | u \leq v\}$.

- For all $A \in \mathcal{B}(\mathbb{R})$:

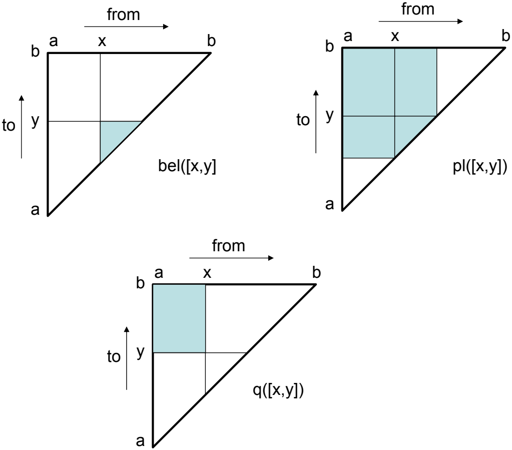
$$bel(A) = \iint_{[u,v]\subseteq A} f(u,v)\, dudv,$$

$$pl(A) = \iint_{[u,v]\cap A \neq \emptyset} f(u,v) dudv,$$

$$q(A) = \iint_{[u,v]\supseteq A} f(u,v) dudv,$$

Belief functions on real numbers

Continuous case



Belief functions on real numbers

Continuous case (continued)

- m can be recovered from bel or q as

$$m([u,v]) = -\frac{\partial^2 bel([u,v])}{\partial u \partial v} = -\frac{\partial^2 q([u,v])}{\partial u \partial v}$$

- TBM conjunctive rule:

$$(q_1 \textcircled{\cap} q_2)([u,v]) = q_1([u,v]) \cdot q_2([u,v]), \quad \forall u \leq v$$

- Pignistic probability density:

$$Betp(x) = \lim_{\epsilon \to 0} \int_{-\infty}^{x} \int_{x+\epsilon}^{+\infty} \frac{f(u,v)}{v-u} dv du.$$

Continuous mass functions on real numbers

Example

- Continuous mass functions naturally arise in statistical inference (Dempster, 1966-1968).
- Let us consider a piece of equipment that fails according to a Bernoulli process with probability p.
- Let X denote the r.v. taking the value 1 if the piece of equipment fails, and 0 otherwise.
- We have made n independent observations $X_1, \ldots, X_n$ of X, in which the piece of equipment has been found to fail r times out of n.
- Opinion about p?

Continuous mass functions on real numbers

Example (continued)

Solution (Dempster, 1966):

- If $0 < r < n$:

$$m([u,v]) = \frac{n!}{(r-1)!(n-r-1)!} u^{r-1}(1-v)^{n-r-1}$$

- If $r = 0$:

$$m([0,v]) = n(1-v)^{n-1}$$

- If $r = n$:

$$m([u,1]) = nu^{n-1}$$

Building belief functions

- The basic theory tells us how to reason and compute with belief functions, but it does not tell us where belief functions come from.
- We need formalized methods for modeling expert opinions and statistical information using belief functions.
- Four general approaches:
  - Least Commitment Principle
  - Using meta-knowledge about information sources (discounting)
  - Predictive belief functions
  - Optimizing a criterion (e.g., clustering)

Outline

Least Commitment Principle

General approach

- General approach:
  1. Express the available information as a set of constraints on an unknown mass function;
  2. Find the least-committed mass function (according to some ordering), compatible with the constraints.
- Three applications:
  - Inverse pignistic transformation
  - Credal ordering constraint
  - Deconditioning

Inverse pignistic transformation

Problem statement

- Assume we want to elicit a mass function m on $\Omega = \{\omega_1, \ldots, \omega_K\}$ from an expert.
- It is easier to elicit the corresponding pignistic probability:
  - For each $\omega_k \in \Omega$ ask for the fair price p_k the expert is willing to pay for a ticket that will allow him to receive 1 euro if $X = \omega_k$, and to receive nothing otherwise.
  - The pignistic probability mass function is $p(\omega_k) = p_k$, $k = 1, \ldots, K$.
- How to compute a mass function m on Ω compatible with p?

Inverse pignistic transformation

Discrete case

- There are infinitely many mass functions m such that $Bet(m) = p$.
- The q-least committed solution is a consonant mass function defined by the following possibility distribution:

$$\pi(\omega_k) = \sum_{\ell=1}^{K} \min(p_k, p_\ell).$$

Inverse pignistic transformation

Example

- Let us consider a frame $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and the pignistic probability mass function

$$p(\omega_1) = 0.7, \quad p(\omega_2) = 0.2, \quad p(\omega_3) = 0.1$$

- We have

$$\begin{aligned}
\pi(\omega_1) &= 0.7 + 0.2 + 0.1 = 1 \\
\pi(\omega_2) &= 0.2 + 0.2 + 0.1 = 0.5 \\
\pi(\omega_3) &= 0.1 + 0.1 + 0.1 = 0.3.
\end{aligned}$$

- The corresponding mass function is

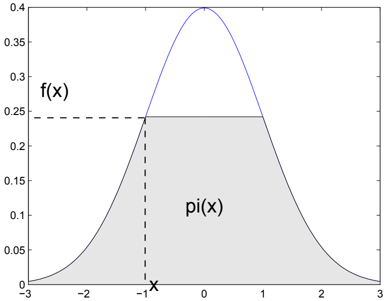
$$m(\{\omega_1\}) = 0.5, \quad m(\{\omega_1, \omega_2\}) = 0.2, \quad m(\Omega) = 0.3.$$

Inverse pignistic transformation

Continuous case

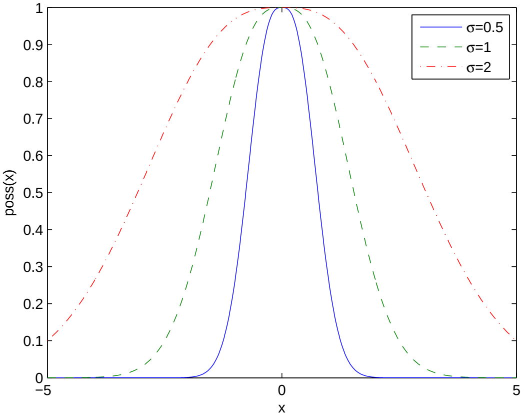
If $\Omega = \mathbb{R}$ and f is a pignistic density, we have

$$\pi(x) = \int_{-\infty}^{+\infty} \min(f(x), f(t))dt.$$



Inverse pignistic transformation

Example: normal distribution



Credal ordering constraint

Problem

- Consider the following problems:
  1. Let X and X' be two variables. Our beliefs on X are represented by m. Additionally, we believe that X' tends to take greater values than X. How to quantify our beliefs on X' using a mass function?
  2. We consider one variable X and two different contexts C and C'. When C holds, our beliefs on X are represented by m. When C' holds, we cannot precisely assess our beliefs on X, but we believe that X tends to take higher values than it does when C holds. How to quantify our beliefs on X in context C'?
- Approach: formalize the notion of “tending to take higher values" as a constraint on a mass function, and find the least-committed solution compatible with that constraint.

Credal ordering constraint

Definitions

- Given two probability distributions P and P' on $\mathbb{R}$, we say that P is stochastically less than or equal to P' if

$$P((x,+\infty)) \leq P'((x,+\infty)), \quad \forall x \in \mathbb{R}$$

- How to extend this notion to compare two mass functions m and m' on $\mathbb{R}$?
- Four definitions (credal orderings):
  1. $m \lesssim m'$ iff $bel((x,+\infty)) \leq pl'((x,+\infty)), \quad \forall x \in \mathbb{R}$;
  2. $m \leqslant m'$ iff $bel((x,+\infty)) \leq bel'((x,+\infty)), \quad \forall x \in \mathbb{R}$;
  3. $m \leqslant m'$ iff $pl((x,+\infty)) \leq pl'((x,+\infty)), \quad \forall x \in \mathbb{R}$;
  4. $m \ll m'$ iff $pl((x,+\infty)) \leq bel'((x,+\infty)), \quad \forall x \in \mathbb{R}$.

Credal ordering constraint

Example of result

Theorem

The pl-least committed element mass function m' such that $m' \geqslant m$ exists and is unique. It is the consonant mass function $m_{\geqslant}$ with possibility distribution $\pi_{\geqslant}$ given by

$$\pi_{\geqslant}(x) = pl((-\infty, x])$$

where pl is the plausibility function associated to m.

Credal ordering constraint

Example

- Assume that m represents the available information regarding the failure probability p of a component in standard operating condition, after observing r failures out of n trials.
- We want to assess our beliefs regarding the failure probability p' of the same component in a more stringent environment, for which we have no data.
- We only know that the failure probability in this new environment tends to be higher than the failure probability in standard operating condition.
- If $r > 0$, we get

$$m_{\geqslant}([u,1]) = \frac{n!}{(r-1)!(n-r)!} u^{r-1}(1-u)^{n-r}, \quad \forall u \in [0,1].$$

Deconditioning

- Let m_0 be a mass function on Ω expressing our beliefs about X in a context where we know that $X \in B$.
- We want to build a mass function m on Ω verifying the constraint

$$m(\cdot|B) = m_0$$

- Any mass function m built from m_0 by transferring each mass $m_0(A)$ to $A \cup C$ for some $C \subseteq \overline{B}$ satisfies the constraint. The largest such set is $A \cup \overline{B}$.
- s-least committed solution: transfer $m_0(A)$ to $A \cup \overline{B}$.

$$m(D) = \begin{cases} m_0(A) & \text{if } D = A \cup \overline{B} \text{ for some } A \subseteq B, \\ 0 & \text{otherwise} \end{cases}$$

Deconditioning

Ballooning extension

- More complex situation: two frames Ω_X and Ω_Y.
- Let $m_0^{\Omega_X}$ be a mass function on Ω_X expressing our beliefs about X in a context where we know that $Y \in B$ for some $B \subseteq \Omega_Y$.
- We want to find $m^{\Omega_{XY}}$ such that

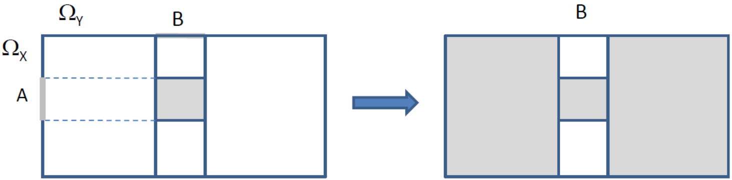
$$\left(m^{\Omega_{XY}} \textcircled{\cap} (m_B^{\Omega_Y})^{\uparrow \Omega_{XY}}\right)^{\downarrow \Omega_X} = m_0^{\Omega_X}$$

Deconditioning

Ballooning extension (continued)

- s-least committed solution:

$$m^{\Omega_{XY}}(D) = \begin{cases} m_0^{\Omega_X}(A) & \text{if } D = (A \times \Omega_Y) \cup (\Omega_X \times \overline{B}) \\ & \text{for some } A \subseteq \Omega_X, \\ 0 & \text{otherwise} \end{cases}$$



- Notation $m^{\Omega_{XY}} = (m_0^{\Omega_X})^{\Uparrow \Omega_{XY}}$ (ballooning extension).

Outline

Discounting

Problem statement

- A source of information provides:
  - a value;
  - a set of values;
  - a probability distribution, etc..
- The information is:
  - not fully reliable or
  - not fully relevant.
- Examples:
  - Possibly faulty sensor;
  - Measurement performed in unfavorable experimental conditions;
  - Information is related to a situation or an object that only has some similarity with the situation or the object considered (case-based reasoning).

Discounting

Formalization

- A source S provides a mass function m_S^Ω.
- S may be reliable or not. Let $\mathcal{R} = \{R, NR\}$.
- Assumptions:
  - If S is reliable, we accept m_S^Ω as a representation of our beliefs:
$$m^\Omega(\cdot|R) = m_S^\Omega$$
  - If S is not reliable, we know nothing:
$$m^\Omega(\cdot|NR) = m_\Omega^\Omega$$
  - The source has a probability $1-\alpha$ of being reliable:
$$m^\mathcal{R}(\{NR\}) = \alpha, \quad m^\mathcal{R}(\{R\}) = 1-\alpha$$

Discounting

Solution

- To exploit this information, we need to
  1. Vacuously extend $m^{\mathcal{R}}$ to $\Omega \times \mathcal{R}$;
  2. Compute the ballooning extension of $m^{\Omega}(\cdot|R)$ and $m^{\Omega}(\cdot|NR)$ in $\Omega \times \mathcal{R}$;
  3. Combine the three mass functions using the TBM conjunctive rule;
  4. Marginalize the combined mass function on Ω.
- Result:

$$^{\alpha}m^{\Omega} = (1-\alpha)m_S^{\Omega} + \alpha m_{\Omega}^{\Omega}$$

- Other expression:

$$^{\alpha}m^{\Omega} = m_S^{\Omega} \textcircled{\cup} m_0^{\Omega}.$$

  with $m_0^{\Omega}(\Omega) = \alpha$ and $m_0^{\Omega}(\emptyset) = 1-\alpha$.
- $^{\alpha}m^{\Omega}$ is a s-less committed than (a generalization of) m_S^{Ω}: $^{\alpha}m^{\Omega} \sqsupseteq_s m_S^{\Omega}$.

Application to classification

Problem statement

- Let Ω be a set of classes, and

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

  a learning set, where $\mathbf{x}_i$ is a feature vector for object o_i, and m_i^Ω a mass function concerning the class of that object.
- Let $\mathbf{x}$ be the feature vector describing a new object o to be classified.
- Problem: Construct a mass function m^Ω relative to the class of o.

Application to classification

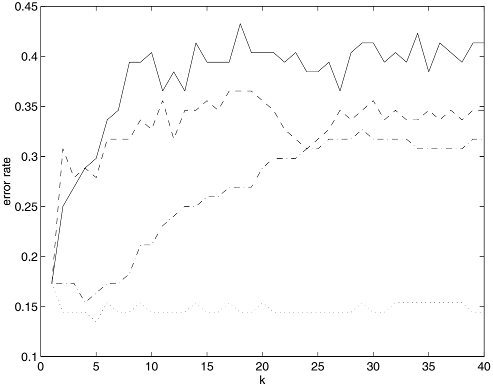
Solution

- Assumption: let α_i be the plausibility that objects o and o_i do not belong to the same class. We assume that $\alpha_i = \phi(d(\mathbf{x}, \mathbf{x}_i))$, where d is a distance, and ϕ is an increasing function from $\mathbb{R}^+$ to $[0, 1]$.
- Each learning instance $(\mathbf{x}_i, m_i^\Omega)$ is a source of information, which must be discounted with discount rate α_i.
- Assuming independence, the n discounted mass functions should be combined using Dempster's rule:

$$m^\Omega = {}^{\alpha_1}m_1^\Omega \oplus \ldots \oplus {}^{\alpha_n}m_n^\Omega$$

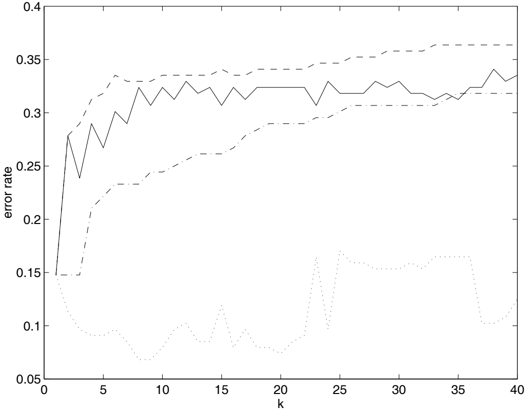
- Alternatively, we may only take into account the k nearest neighbors of $\mathbf{x}$ (evidential k-NN rule).

Sonar data (UCI database)



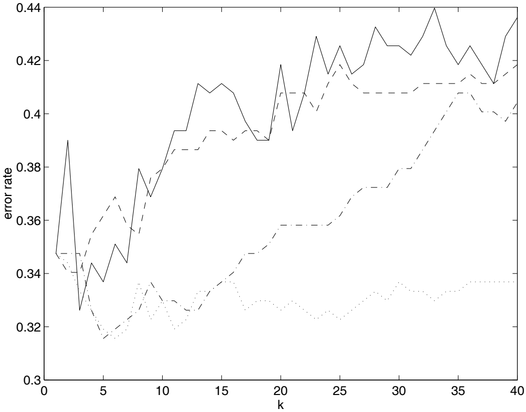
Test error rates as a function of *k* for the voting (-), evidential (:), fuzzy (–) and distance-weighted (-.) *k*-NN rules.

Ionosphere data (UCI database)



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

Vehicle data (UCI database)



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

Generalization: Contextual Discounting

Formalization

- A more general model allowing us to take into account richer meta-information about the source.
- Let $\Theta = \{\theta_1, \ldots, \theta_L\}$ be a partition of Ω, representing different contexts.
- Let $m^{\mathcal{R}}(\cdot|\theta_k)$ denote the mass function on $\mathcal{R}$ quantifying our belief in the reliability of source S, when we know that the actual value of X is in θ_k.
- We assume that:

$$m^{\mathcal{R}}(\{R\}|\theta_k) = 1 - \alpha_k, \quad m^{\mathcal{R}}(\{NR\}|\theta_k) = \alpha_k.$$

  for each $k \in \{1, \ldots, L\}$.
- Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_L)$.

Contextual Discounting

Example

- Let us consider a simplified aerial target recognition problem, in which we have three classes: airplane ($\omega_1 \equiv a$), helicopter ($\omega_2 \equiv h$) and rocket ($\omega_3 \equiv r$).
- Let $\Omega = \{a, h, r\}$.
- The sensor provides the following mass function: $m_S^\Omega(\{a\}) = 0.5$, $m_S^\Omega(\{r\}) = 0.5$.
- We assume that
  - The probability that the source is reliable when the target is an airplane is equal to $1 - \alpha_1 = 0.4$;
  - The probability that the source is reliable when the target is either a helicopter, or a rocket is equal to $1 - \alpha_2 = 0.9$.
- We have $\Theta = \{\theta_1, \theta_2\}$, with $\theta_1 = \{a\}$, $\theta_2 = \{h, r\}$, and $\boldsymbol{\alpha} = (0.6, 0.1)$.

Contextual Discounting

Solution

- To exploit this information, we need to
  1. Compute the ballooning extension of $m^\Omega(\cdot|R)$ and $m^\mathcal{R}(\cdot|\theta_k, k=1,\ldots,L$ in $\Omega\times\mathcal{R}$;
  2. Combine the $L+1$ mass functions conjunctively;
  3. Marginalize the combined mass function on Ω.
- Result:

$$^{\boldsymbol{\alpha}}m^\Omega = m_S^\Omega \cup\!\!\!\!\bigcirc\, m_1^\Omega \cup\!\!\!\!\bigcirc \ldots \cup\!\!\!\!\bigcirc\, m_L^\Omega.$$

  with $m_k^\Omega(\theta_k)=\alpha_k$ and $m_k^\Omega(\emptyset)=1-\alpha_k$.
- Standard discounting is recovered as a special case when $\Theta=\{\Omega\}$.

Contextual Discounting

Example (continued)

- The discounted mass function can be obtained by combining disjunctively 3 mass functions:
  - $m_S^\Omega(\{a\}) = 0.5$, $m_S^\Omega(\{r\}) = 0.5$;
  - $m_1^\Omega(\{a\}) = 0.6$, $m_1^\Omega(\emptyset) = 0.4$;
  - $m_1^\Omega(\{h, r\}) = 0.1$, $m_1^\Omega(\emptyset) = 0.9$.
- Result:

$$^\alpha m^\Omega(\{a\}) = 0.45, \quad ^\alpha m^\Omega(\Omega) = 0.08,$$

$$^\alpha m^\Omega(\{r\}) = 0.18, \quad ^\alpha m^\Omega(\{a, r\}) = 0.27,$$

$$^\alpha m^\Omega(\{h, r\}) = 0.02.$$

Outline

Generalized Bayes Theorem (Smets, 1978)

Problem statement

- Two variables $X \in \Omega_X$ et $\theta \in \Theta = \{\theta_1, \ldots, \theta_K\}$.
- Typically:
  - X is observed (sensor measurement),
  - θ is not observed (class, unknown parameter).
- We know $pl^{\Omega_X}(\{x\}|\theta_k) = pl_k(x)$, $\forall x, k$.
- We have no prior information about θ: $m^{\Theta}(\Theta) = 1$.
- We observe $X = x$. Belief function on Θ?

Generalized Bayes Theorem

Solution and properties

- Solution (derived from the LCP):

$$m^{\Theta}(\cdot|x) = \textcircled{\cap}_{k=1}^{K} \overline{\{\theta_k\}}^{pl_k(x)}.$$

- Property 1: Bayes' theorem is recovered as a special case when $pl_k(x) = P(x|\theta_k)$ (probabilistic information) and $m^{\Theta}(\cdot|x)$ is combined with a prior Bayesian mass function .
- Property 2: If X and Y are cognitively independent conditionally on θ:

$$pl_k(x,y) = pl_k(x) pl_k(y), \quad \forall k$$

then

$$m^{\Theta}(\cdot|x,y) = \textcircled{\cap}_{k=1}^{K} \overline{\{\theta_k\}}^{pl_k(x,y)} = m^{\Theta}(\cdot|x) \textcircled{\cap} m^{\Theta}(\cdot|y).$$

Generalized Bayes Theorem

Application

- Example: Let X_j be a vector of attributes from sensor j, and $f_k(x_j)$ its estimated pdf in class θ_k.
- Definition of $pl_k(x_j)$ (Appriou, 1991):

$$pl_k(x_j) = \alpha_{jk} + (1 - \alpha_{jk})\rho_j f(x_j|\theta_k),$$

where

- ρ_j: normalization coefficient;
- α_{jk}: discount rate expressing our partial knowledge of the distribution of X_j in class θ_k, in a given operational context.

Generalized Bayes Theorem

Sensor fusion

- Independent sensors:

$$m^{\Theta}(\cdot|x_1,\ldots,x_J) = \textcircled{\cap}_{j=1}^{J} m^{\Theta}(\cdot|x_j) = \textcircled{\cap}_{k=1}^{K} \overline{\{\theta_k\}}^{\prod_{j=1}^{J} pl_k(x_j)}.$$

- Dependent sensors:

$$m^{\Theta}(\cdot|x_1,\ldots,x_J) = \textcircled{\wedge}_{j=1}^{J} m^{\Theta}(\cdot|x_j) = \textcircled{\cap}_{k=1}^{K} \overline{\{\theta_k\}}^{\bigwedge_{j=1}^{J} pl_k(x_j)}.$$

Outline

Predictive belief functions

Motivation

- Let X be random variable (defined from a repeatable random experiment), with unknown probability $\mathbb{P}_X$.
- We have observed an independent, identically distributed random sample from X: $\mathbf{X} = (X_1, \ldots, X_n)$.
- Problem: quantify our beliefs regarding a future realization from X using a belief function $bel^{\Omega}(\cdot; \mathbf{X})$: predictive belief function.

Predictive belief functions

Examples

1. Example 1:
   - We have drawn r black balls in n drawings from an urn with replacement:
   - What is our belief that the next ball to be drawn from the urn will be black?
2. Example 2:
   - The lifetimes of 20 bearings have been observed:

     2398, 2812, 3113, 3212, 3523, 5236, 6215,
     6278, 7725, 8604, 9003, 9350, 9460, 11584,
     11825, 12628, 12888, 13431, 14266, 17809.
   - Let X be the lifetime of a bearing taken at random from the same population. Belief function on X?

Predictive belief functions

Requirements

- Requirement 1 (Hacking's frequency principle):
  - If $\mathbb{P}_X$ were known, we would equate our beliefs with probabilities: $bel^{\Omega}(\cdot; \mathbb{P}_X) = \mathbb{P}_X$.
  - Weaker version when $\mathbb{P}_X$ is unknown:

$$\forall A \subset \Omega, \quad bel^{\Omega}(A; \mathbf{X}) \xrightarrow{P} \mathbb{P}_X(A), \text{ as } n \to \infty,$$

- Requirement 2 (LCP):
  - As n is finite, $bel^{\Omega}(\cdot; \mathbf{X})$ should be less committed than $\mathbb{P}$. However, the condition $bel^{\Omega}(\cdot; \mathbf{X}) \leq \mathbb{P}_X$ is too restrictive
  - Weaker requirement:

$$\mathbb{P}\left(bel^{\Omega}(A; \mathbf{X}) \leq \mathbb{P}_X(A), \forall A \subset \Omega\right) \geq 1 - \alpha.$$

Predictive belief functions

Meaning of Requirement 2

$$\mathbf{x} = (x_1, \ldots, x_n) \rightarrow bel^{\Omega}(\cdot, \mathbf{x})$$

$$\mathbf{x}' = (x_1', \ldots, x_n') \rightarrow bel^{\Omega}(\cdot; \mathbf{x}')$$

$$\mathbf{x}'' = (x_1'', \ldots, x_n'') \rightarrow bel^{\Omega}(\cdot; \mathbf{x}'')$$

$$\vdots$$

- As the number of realizations of the random sample tends to ∞, the proportion of belief functions less committed than $\mathbb{P}_X$ should tend to $1 - \alpha$.

Predictive belief functions

Solutions

- If X is discrete, $\Omega = \{\omega_1, \ldots, \omega_K\}$: a solution can be obtained using a confidence region on probabilities $p_k = \mathbb{P}(X = \omega_k)$:

$$\mathbb{P}\left(P_k^- \leq p_k \leq P_k^+, k = 1, \ldots, K\right) = 1 - \alpha$$

(T. Denoeux. International Journal of Approximate Reasoning, 2006).

- If X is absolutely continuous, $\Omega = \mathbb{R}$: a solution can be obtained using a confidence band on the cumulative distribution function F_X of X.
(A. Aregui et T. Denoeux. Proceedings of ISIPTA '07, 2007).

Predictive belief functions

Confidence band

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ be an iid sample from X with cdf F_X.
- A pair of functions $(\underline{F}(\cdot; \mathbf{X}), \overline{F}(\cdot; \mathbf{X}))$ computed from $\mathbf{X}$ and such that $\underline{F}(\cdot; \mathbf{X}) \leq \overline{F}(\cdot; \mathbf{X})$ is a confidence band at level $\alpha \in (0, 1)$ if

$$P\left\{\underline{F}(x; \mathbf{X}) \leq F_X(x) \leq \overline{F}(x; \mathbf{X}), \ \forall x \in \mathbb{R}\right\} = 1 - \alpha,$$

Predictive belief functions

Kolmogorov Confidence band

- A non parametric confidence band can be computed using the Kolmogorov statistic:

$$D_n = \sup_x |S_n(x; \mathbf{X}) - F_X(x)|,$$

where $S_n(\cdot; \mathbf{X})$ is the sample cdf.

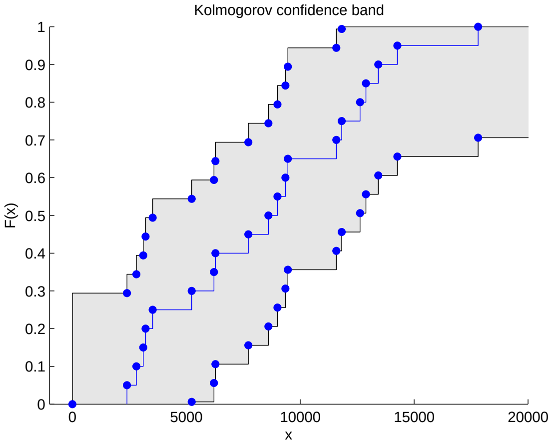
- The probability distribution of D_n can be computed exactly. Let $d_{n,\alpha}$ by the α-critical value of D_n, i.e., $\mathbb{P}(D_n \geq d_{n,\alpha}) = \alpha$.
- The two step functions

$$\begin{aligned} \underline{F}(x; \mathbf{X}) &= \max(0, S_n(x; \mathbf{X}) - d_{n,\alpha}), \\ \overline{F}(x; \mathbf{X}) &= \min(1, S_n(x; \mathbf{X}) + d_{n,\alpha}) \end{aligned}$$

form a confidence band at level $1 - \alpha$.

Kolmogorov Confidence band

Bearing data ($1 - \alpha = 0.95$)



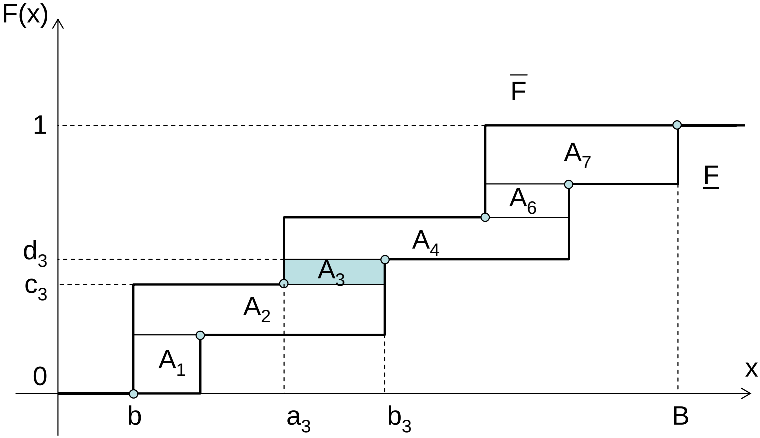
Predictive belief functions

p-boxes and belief functions

- A Kolmogorov confidence band defines a p-box (a set of probability measures with cdf constrained by 2 step functions).
- A p-box can be shown to be equivalent to a discrete mass function.
- The mass function constructed from a Kolmogorov confidence band at level $1-\alpha$ can be shown to be a predictive belief function at level $1-\alpha$.

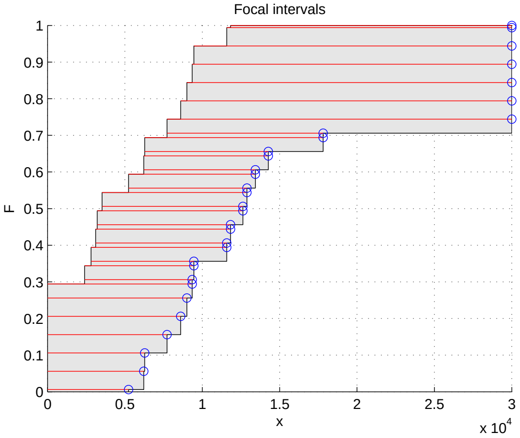
Construction of a mass function from a p-box

Principle



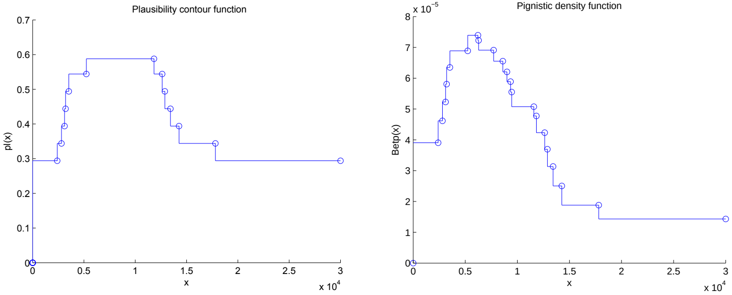
Construction of a mass function from a p-box

Bearing data



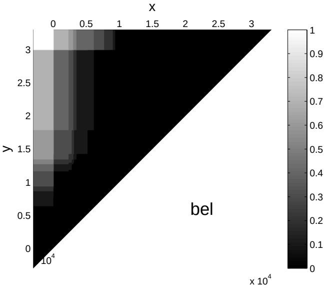
Contour and pignistic density functions

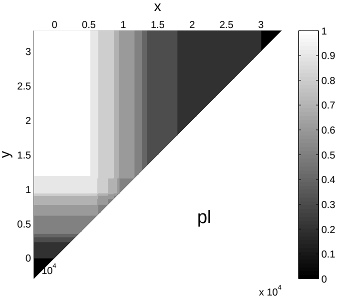
Bearing data



Belief and plausibility functions

Bearing data





Predictive belief functions

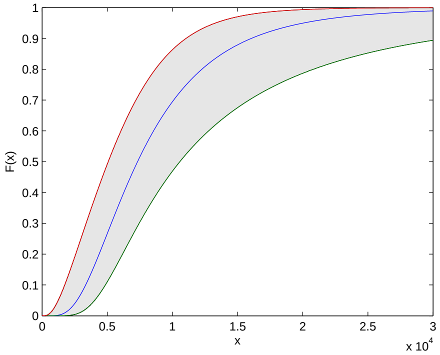
Continuous confidence bands

- Narrower confidence bands can be constructed using parametric methods.
- These methods lead to continuous bounding functions, which can be shown to induce continuous predictive belief functions.

Continuous confidence bands

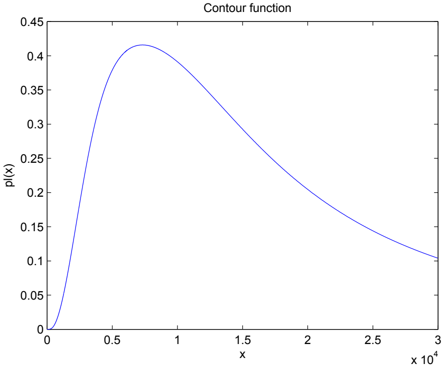
Bearing data

- Parametric confidence band for the Bearing data at level $1 - \alpha = 0.95$, using the Cheng and Yles method, assuming a log-normal distribution:



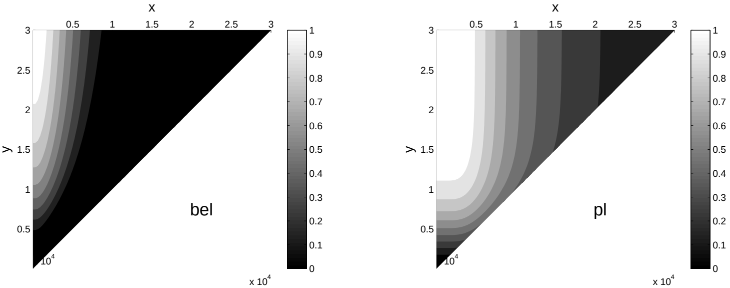
Contour function

Bearing data



Continuous belief and plausibility functions

Bearing data



Outline

Evidential clustering

Problem statement

- A typical application where mass functions can be determined by the solutions of an optimization problem.
- We consider
  - a collection of n objects;
  - a matrix $D = (d_{ij})$ of pairwise dissimilarities between the objects (dissimilarities may or may not correspond to distances in some space of attributes).
- Assumption: each object belongs to one of c classes in $\Omega = \{\omega_1, ..., \omega_c\}$.
- What can we say about the class membership of the objects, knowing only their dissimilarities?

Evidential clustering

Credal partition

- In the belief function framework, uncertain information about the class membership of objects may be represented in the form of mass functions $m_1, \ldots, m_n$ on Ω.
- Resulting structure $M = (m_1, \ldots, m_n)$ is called a credal partition.

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

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

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

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 class membership to dissimilarities.
- Basic idea: "The more similar two objects, the more plausible it is that they belong to the same class".
- How to formalize this idea?

EVCLUS algorithm

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}.

EVCLUS algorithm

Cost function

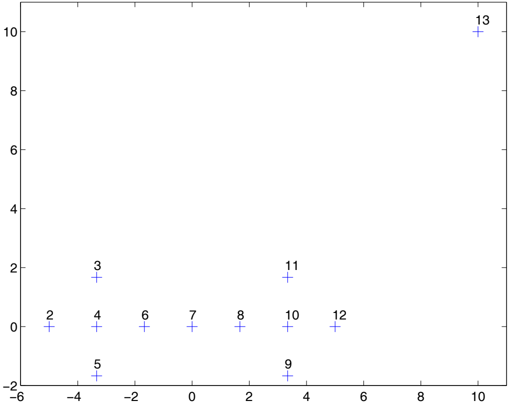
- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict K_{ij}.
- Example of a cost function:

$$J(M) = \sum_{i<j} \left(K_{ij} - d_{ij}\right)^2$$

- M can be determined by minimizing J using a non linear optimization procedure.
- To reduce the complexity, focal sets can be reduced to $\{\omega_k\}_{k=1}^{c}$, $\emptyset$, and Ω

Butterfly example

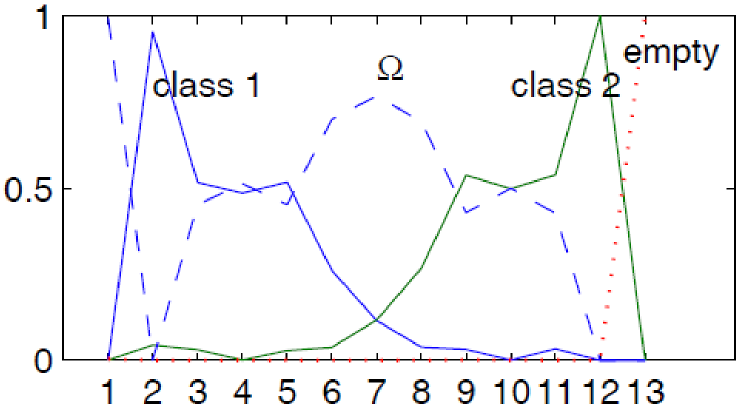
Data



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

Butterfly example

Results



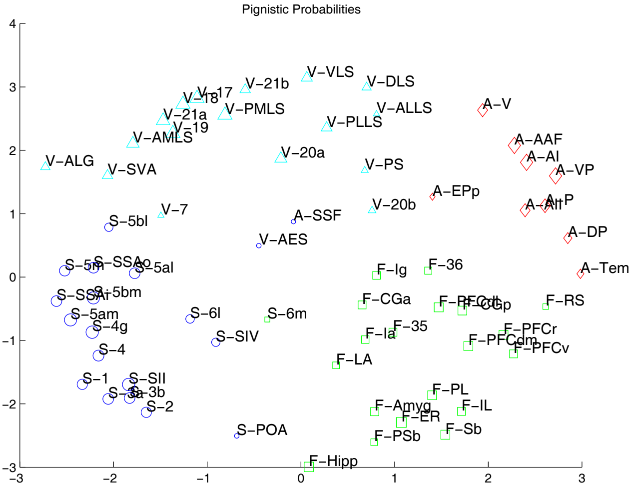
Experiments: Cat cortex dataset

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.

Cat cortex dataset

Results



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.
- More computationally efficient procedures: ECM (Masson and Denoeux, 2008) and RECM (Masson and Denoeux, 2009).

Conclusion

- Belief functions can be seen both as generalized sets and as generalized probability measures:
  - A very general framework for representing imprecision and uncertainty.
  - Reasoning mechanisms extend both set-theoretic operations (intersection, union, cylindrical extension, etc.) and probabilistic operations (conditioning, marginalization, stochastic ordering, etc.).
  - Extension of set-membership approaches (e.g., interval analysis) and probabilistic methods (e.g., classification using the GBT).

Conclusion (continued)

- Developing engineering applications using the belief function framework is still often more art than science BUT ...
- Systematic and principled methods now exist for modeling expert knowledge and statistical information in the belief function framework:
  - Least-commitment principle
  - Discounting
  - GBT
  - Predictive belief functions
  - Optimization of a cost function,
  - etc.
- More research on expert knowledge elicitation and statistical inference is needed.

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.

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

References II

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

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

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.

T. Denœux.
Constructing Belief Functions from Sample Data Using Multinomial Confidence Regions.
International Journal of Approximate Reasoning, Vol. 42, Issue 3, Pages 228-252, 2006.

References III

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

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

T. Denœux.
Conjunctive and Disjunctive Combination of Belief Functions Induced by Non Distinct Bodies of Evidence.
Artificial Intelligence, Vol. 172, pages 234Ű264, 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.

References IV

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

T. Denœux.

Extending stochastic ordering to belief functions on the real line.

Information Sciences, Vol. 179, pages 1362-1376, 2009.

G. Nassreddine, F. Abdallah and T. Denœux.

State estimation using interval analysis and belief function theory: Application to dynamic vehicle localization.

IEEE Transactions on Systems, Man and Cybernetics B, Accepted for publication, 2009.