Theory of belief functions

Introduction, connections with rough sets and some recent advances

Thierry Denœux[1]

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

RST Workshop,
Milano, September 14-16

What is the Theory of belief functions?

- A formal framework for representing and reasoning from partial (uncertain, imprecise) information. Also known as Dempster-Shafer theory or Evidence theory.
- Introduced by Dempster (1968) and Shafer (1976), further developed by Smets (Transferable Belief Model) and others.
- The theory of belief functions extends both the set-membership and probabilistic approaches to uncertain reasoning:
  - A belief function may be viewed both as a generalized set and as a non additive measure;
  - Extension of probabilistic notions (conditioning, marginalization) and set-theoretic notions (intersection, union, inclusion, etc.).

Links with other theories of uncertainty

- The theory of belief functions also has links with other contemporary theories of uncertainty, including:
  - Random sets;
  - Imprecise probabilities;
  - Possibility theory;
  - Rough sets.
- Purpose of these talk:
  - Brief introduction or reminder on belief functions emphasizing some of the known relationships with other theories;
  - Presentation of some new results concerning the manipulation of belief functions in very large universes.

Outline

Outline

Mass function

- Let X be a variable taking values in a finite domain Ω, called the frame of discernment.
- We collect a piece of evidence (information) about X.
- This piece of evidence has different interpretations $\theta_1, \ldots, \theta_r$ with corresponding subjective probabilities $p_1, \ldots, p_r$.
- If interpretation θ_i holds, we only know that $X \in A_i$ for some $A_i \subseteq \Omega$, and nothing more. Let $A_i = \Gamma(\theta_i)$.
- The probability that the evidence means exactly that $X \in A$ is $m(A) = \sum_{\{i|A_i=A\}} p_i$.
- Function $m : 2^\Omega \to [0, 1]$ is called a mass function with focal sets $A_1, \ldots, A_r$.

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. We know that the witness is drunk 20 % of the time.
- Two interpretations:
  1. θ_1= the witness was not drunk, $p_1 = 0.8$;
  2. θ_2= the witness was drunk, $p_2 = 0.2$.
- We have $\Gamma(\theta_1) = \{Peter, John\}$ and $\Gamma(\theta_2) = \Omega$, hence

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

Mass functions
Special cases

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

Mass function

Comparison with the random set framework

- Each mass function m on Ω can thus be seen associated a triple (Θ, P, Γ), where Γ is a multi-valued mapping from Θ to $2^\Omega \setminus \{\emptyset\}$.
- This formally defines a random set: mass functions are thus exactly equivalent to random sets from a mathematical point of view.
- However, they have different interpretations:
  - Random set view: a random mechanism generates each set A with chance $m(A)$. Example: taking a handful of balls from an urn.
  - Belief function view: a given piece of evidence supports different hypotheses with different subjective probabilities. Example: taking a single ball from an urn and partially observing the result.

Outline

Belief function

Definition and interpretation

- The belief function induced by m is defined as

$$bel(A) = \sum_{B \subseteq A} m(B), \quad \forall A \subseteq \Omega.$$

- $bel(A)$ can be seen as the probability that the evidence can be interpreted as implying that $X \in A$:

$$bel(A) = P(\{\theta \in \Theta | \Gamma(\theta) \subseteq A\}.$$

- It can thus be interpreted as:
  - a total degree of support in A provided by the item of evidence;
  - a measure of our total belief committed to A after receiving that item of evidence.

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 plausibility function is defined by

$$pl(A) = 1 - bel(\overline{A}) = \sum_{B \cap A \neq \emptyset} m(B)$$

- Interpretation:
  - degree to which the evidence is not contradictory with A:
  - probability that A cannot be refuted by the available evidence.
- m, bel et pl are thus three equivalent representations of
  - a piece of evidence or, equivalently,
  - a state of belief induced by this evidence.
- If m is Bayesian, then $bel = pl$ is a probability measure.

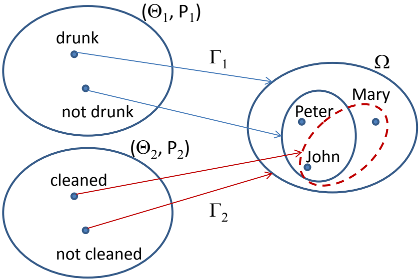
Outline

Dempster's rule

Murder example continued

- The first item of evidence gave us: $m_1(\{Peter, John\}) = 0.8$, $m_1(\Omega) = 0.2$.
- New piece of evidence: a blond hair has been found.
- There is a probability 0.6 that the room has been cleaned before the crime: $m_2(\{John, Mary\}) = 0.6$, $m_2(\Omega) = 0.4$.
- How to combine these two pieces of evidence?

Dempster's rule
Justification



- If $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ both hold, then $X \in \Gamma_1(\theta_1) \cap \Gamma_2(\theta_2)$.
- If the two pieces of evidence are independent, then this happens with probability $P_1(\{\theta_1\})P_2(\{\theta_2\})$.
- If $\Gamma_1(\theta_1) \cap \Gamma_2(\theta_2) = \emptyset$, we know that the pair of interpretations (θ_1, θ_2) is impossible.
- The joint probability distribution on $\Theta_1 \times \Theta_2$ must be conditioned, eliminating such pairs.

Dempster's rule
Expression and example

$$(m_1 \oplus m_2)(A) = \frac{\sum_{B\cap C=A} m_1(B)m_2(C)}{\sum_{B\cap C\neq\emptyset} m_1(B)m_2(C)}, \quad \forall A \neq \emptyset$$

	$\{Peter, John\}$ 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

Dempster's rule

Properties

- Commutativity, associativity. Neutral element: m_Ω.
- Generalization of intersection: if m_A and m_B are categorical mass functions and $A \cap B \neq \emptyset$, then

$$m_A \oplus m_B = m_{A \cap B}$$

- Generalization of probabilistic conditioning: if m is a Bayesian mass function and m_A is a categorical mass function, then $m \oplus m_A$ is a Bayesian mass function that corresponding to the conditioning of m by A.

Dempster's rule

Incompatibility with the imprecise probability interpretation

- To each mass function m on Ω can be associated a set $\mathcal{P}(m)$ of compatible probability measures such that $P(A) \geq bel(A)$ for all $A \subseteq \Omega$. We then have:

$$bel(A) = \inf_{P \in \mathcal{P}} P(A) \text{ and } pl(A) = \sup_{P \in \mathcal{P}} P(A).$$

- However, $m \oplus m_A$ does not correspond to $\{P(\cdot|A), P \in \mathcal{P}(m)\}$.
- Consequently, the imprecise probability interpretation of belief functions is not compatible with Dempster's rule and the DS model is not an imprecise probability model.

Outline

Possibility theory

- When 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.
- Conversely, to any possibility distribution π corresponds a consonant mass function whose focal sets are the α-cuts of π.
- The theory of belief function is thus, in a sense, more general than possibility theory.
- However, consonance is not preserved by Dempster's rule, and the minimum rule of possibility theory has no obvious interpretation from the point of view of belief functions.

Fuzzification of belief functions
Fuzzy mass functions

- Continuing the murder example, assume that we receive a third item of evidence that tells us that “the murderer is tall”.
- Such “fuzzy” evidence may be represented by a probability space $(\Theta, 2^\Theta)$ and a mapping Γ from Θ to the set $[0,1]^\Omega$ of normal fuzzy subsets of Ω.
- $\widetilde{F}_i = \Gamma(\theta_i)$ defines a possibility distribution that constraints the value of X if interpretation θ_i holds.
- This framework induces a mass function with fuzzy focal sets $\Gamma(\Theta) = \{\widetilde{F}_1, \ldots, \widetilde{F}_n\}$, such that $m(\widetilde{F}_i) = P\left(\Gamma^{-1}(\widetilde{F}_i)\right)$.

Fuzzification of belief functions

Fuzzy belief and plausibility functions

- Functions *bel* and *pl* may be defined as

$$pl(\widetilde{A}) = \sum_{i=1}^{n} \Pi(\widetilde{A}|\widetilde{F}_i)m(\widetilde{F}_i) \quad \forall \widetilde{A} \in [0,1]^{\Omega},$$

$$bel(\widetilde{A}) = \sum_{i=1}^{n} N(\widetilde{A}|\widetilde{F}_i)m(\widetilde{F}_i), \quad \forall \widetilde{A} \in [0,1]^{\Omega},$$

  where $\Pi(\widetilde{A}|\widetilde{F}_i) = \max_{\omega\in\Omega} \min(\widetilde{A}(\omega), \widetilde{F}_i(\omega))$ is the possibility of $\widetilde{A}$ given $\widetilde{F}_i$ and $N(\widetilde{A}|\widetilde{F}_i) = 1 - \Pi(\overline{\widetilde{A}}|\widetilde{F}_i)$ is the necessity of $\widetilde{A}$ given $\widetilde{F}_i$.
- The above expressions reduce to the standard definitions when $\widetilde{A}$ and $\widetilde{F}_i$ are crisp.

Comparison between the two approaches

- A possibility distribution π with corresponding fuzzy set $\widetilde{F}$ may then be viewed as
  - a crisp consonant mass function with focal sets ${}^{\alpha}\widetilde{F}$, or as
  - a fuzzy categorical mass function such that $m(\widetilde{F}) = 1$.
- The corresponding plausibility functions coincide on 2^{Ω}, since

$$pl(A) = \max_{\omega \in \Omega} \min(A(\omega), F_i(\omega)) = \max_{\omega \in A} \min F(\omega)$$

  for all $A \subseteq \Omega$.
- Under the latter view, the belief function and possibility frameworks are special cases of a more general model.

Outline

Some references

The relationship between rough sets and belief functions have been studied by several authors, e.g.:

D. Dubois and H. Prade

Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems*, 17, 191-208, 1990

Y. Y. Yao and P. J. Lingras

Interpretations of belief functions in the theory of rough sets. *Information sciences*, 104, 81-106, 1998

W.-Z. Wu, Y. Leung and W.-X. Zhang

Connections between rough set theory and Dempster-Shafer theory of evidence. *International Journal of General Systems*, Vol. 31 (4), pp. 405-430, 2002.

Pawlak rough sets

- Let R be an equivalence relation on Ω. Any $A \subseteq \Omega$ may be approximated by a (Pawlak) rough set defined by:

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

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

- Let P be a probability measure on $(\Omega/R, 2^{\Omega/R})$. The corresponding inner and outer measures are defined by:

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

- $\underline{P}$ is a belief function and $\overline{P}$ is the dual plausibility function. The focal sets are the equivalence classes of R, and $m(F) = P(F)$ for all $F \in \Omega/R$.

Interval rough sets
Definition

- The classical Pawlak rough set model thus corresponds to a special class of belief functions, whose focal sets form a partition of Ω.
- To establish a connection between rough sets and general belief functions, we need a more general notion: interval rough set.
- Let Θ and Ω be two finite sets and $R \in \Theta \times \Omega$. R is called an interval relation if, for all $\theta \in \Theta$, $\Gamma_R(\theta) = \{\omega \in \Omega | (\theta, \omega) \in R\} \neq \emptyset$.
- Any $A \subseteq \Omega$ may be approximated in Θ by an interval rough set defined by:

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

Interval rough sets

Correspondence with belief functions

- Let P be a probability measure on $(\Theta, 2^\Theta)$. Then, the multi-valued mapping Γ_R defines belief and plausibility functions defined as:

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

- Conversely, any belief function on Ω can be seen as being induced by:
  - an interval relation R between a set Θ and Ω (qualitative component);
  - a probability measure P on Θ (quantitative component).

Rough belief functions

Definitions

- Let m be a mass function on Ω and R an equivalence relation on Ω. The quotient space Ω/R is called a coarsening of Ω.
- The inner and outer approximations of m can be defined as:

$$\underline{m}(A) = \sum_{\{B \subseteq \Omega | \underline{R}(B) = A\}} m(B), \quad \overline{m}(A) = \sum_{\{B \subseteq \Omega | \overline{R}(B) = A\}} m(B).$$

- $\underline{m}$ is a specialization of m: $\underline{m} \subseteq m$ and $\overline{m}$ is a generalization of m: $m \subseteq \overline{m}$.

Rough belief functions

Application

- $\underline{m}$ and $\overline{m}$ my be expressed without loss of information in the coarsening Ω/R, making it possible to perform approximate computations with reduced complexity.
- In particular:

$$\underline{m}_1 \cap\!\!\!\!\bigcirc\, \underline{m}_2 \subseteq m_1 \cap\!\!\!\!\bigcirc\, m_2 \subseteq \overline{m}_1 \cap\!\!\!\!\bigcirc\, \overline{m}_2,$$

where $\cap\!\!\!\!\bigcirc$ denotes Dempster's rule without normalization and

$$\underline{pl}(A) \leq pl(A) \leq \overline{pl}(A), \quad A \subseteq \Omega.$$

Rough belief functions
Case of consonant belief functions

- Let m be a consonant mass function. It is equivalent to a possibility distribution π, itself equivalent to a fuzzy subset of Ω.
- It can be shown that the lower and upper approximations of m induced by an equivalence relation R are consonant and correspond to the rough fuzzy set $(\underline{\pi}, \overline{\pi})$ defined by:

$$\underline{\pi}(\omega) = \min_{\omega' \in [\omega]_R} \pi(\omega'), \quad \overline{\pi}(\omega) = \max_{\omega' \in [\omega]_R} \pi(\omega')$$

- A rough consonant mass function is thus equivalent to a rough fuzzy set.

Rough fuzzy belief functions

- It is also possible to mix up the three frameworks of belief functions, fuzzy sets and rough sets.
- Let m be a fuzzy mass function with fuzzy focal sets $\{\widetilde{F}_1, \ldots, \widetilde{F}_n\}$ and R an equivalence relation on Ω.
- A rough approximation of m can be defined as the pair of fuzzy mass functions $(\underline{m}, \overline{m})$ defined by

$$\underline{m}(\widetilde{A}) = \sum_{\{i | \underline{R}(\widetilde{F}_i) = \widetilde{A}\}} m(\widetilde{F}_i), \quad \overline{m}(\widetilde{A}) = \sum_{\{i | \overline{R}(\widetilde{F}_i) = \widetilde{A}\}} m(\widetilde{F}_i),$$

where $(\underline{R}(\widetilde{F}_i), \overline{R}(\widetilde{F}_i))$ is the rough fuzzy set approximating $\widetilde{F}_i$.

Conclusion on the links with related theories

- The belief function, fuzzy and rough frameworks are not competing but complementary theories that model different aspects of imperfect information:
  - Belief functions adequately model uncertainty induced by partial evidence;
  - Fuzzy sets represent vagueness of concepts as typically expressed by natural language;
  - Rough sets model indiscernibility due to coarseness of representation.
- The three formalisms can be mixed up to build more general models of imperfect information.
- Are the most complex models needed in real applications? This remains to be demonstrated (to my knowledge).

Outline

Complexity of evidential reasoning

- In the worst case, representing beliefs on a finite frame of discernment of size K requires the storage of $2^K - 1$ numbers, and operations on belief functions have exponential complexity.
- In most applications of DS theory, the frame of discernment is usually of moderate size (less than 100). Can we address more complex problems, e.g., in machine learning, involving considerably larger frames of discernment?
- Examples of such problems:
  - Multi-label classification (Denœux, *Art. Intell.*, 2010);
  - Ensemble clustering (Masson and Denœux, *IJAR*, 2011).

Belief functions on very large frames

General Approach

- Outline of the approach:
  1. Consider a partial ordering $\leq$ of the frame Ω such that $(\Omega, \leq)$ is a lattice.
  2. Define the set of propositions as the set $\mathcal{I} \subset 2^{\Omega}$ of intervals of that lattice.
  3. Define *m*, *bel* and *pl* as functions from $\mathcal{I}$ to $[0, 1]$ (this is possible because $(\mathcal{I}, \subseteq)$ has a lattice structure).
- As the cardinality of $\mathcal{I}$ is at most proportional to $|\Omega|^2$, all the operations of Dempster-Shafer theory can be performed in polynomial time (instead of exponential when working in $(2^{\Omega}, \subseteq)$).

Outline

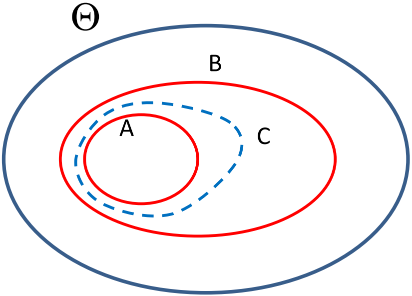
Multi-label classification

- In some problems, learning instances may belong to several classes at the same time.
- For instance, in image retrieval, an image may belong to several semantic classes such as "beach", "urban", "mountain", etc.
- If $\Theta = \{\theta_1, \ldots, \theta_c\}$ denotes the set of classes, the class label of an instance may be represented by a variable y taking values in $\Omega = 2^\Theta$.
- Expressing partial knowledge of y in the Dempster-Shafer framework may imply storing 2^{2^c} numbers.

c	2	3	4	5	6	7	8
2^{2^c}	16	256	65536	4.3e9	1.8e19	3.4e38	1.2e77

Multi-label classification

- The frame of discernment is $\Omega = 2^\Theta$, where Θ is the set of classes.
- The natural ordering in 2^Θ is $\subseteq$, and $(2^\Theta, \subseteq)$ is a (Boolean) lattice.



The intervals of $(2^\Theta, \subseteq)$ are sets of subsets of Θ of the form:

$$[A, B] = \{C \subseteq \Theta | A \subseteq C \subseteq B\}$$

for $A \subseteq B \subseteq \Theta$.

Example (diagnosis)

- Let $\Theta = \{a, b, c, d\}$ be a set of faults.
- Item of evidence 1 $\rightarrow$ a is surely present and $\{b, c\}$ may also be present, with confidence 0.7:

$$m_1([\{a\}, \{a, b, c\}]) = 0.7, \quad m_1([\emptyset_\Theta, \Theta]) = 0.3$$

- Item of evidence 2 $\rightarrow$ c is surely present and either faults $\{a, b\}$ (with confidence 0.8) or faults $\{a, d\}$ (with confidence 0.2) may also be present:

$$m_2([\{c\}, \{a, b, c\}]) = 0.8, \quad m_2([\{c\}, \{a, c, d\}]) = 0.2$$

Example

Combination by Dempster's rule

	$[\{a\},\{a,b,c\}]$ 0.7	$[\emptyset_\Theta,\Theta]$ 0.3
$[\{c\},\{a,b,c\}]$ 0.8	$[\{a,c\},\{a,b,c\}]$ 0.56	$[\{c\},\{a,b,c\}]$ 0.24
$[\{c\},\{a,c,d\}]$ 0.2	$[\{a,c\},\{a,c\}]$ 0.14	$[\{c\},\{a,c,d\}]$ 0.06

Based on this evidence, what is our belief that

- Fault a is present: $bel([\{a\},\Theta]) = 0.56 + 0.14 = 0.70$;
- Fault d is not present: $bel([\emptyset_\Theta,\overline{\{d\}}]) = bel([\emptyset_\Theta,\{a,b,c\}]) = 0.56 + 0.14 + 0.24 = 0.94$.

Multi-label classification

Imprecise labels

- Let us consider a learning set of the form:

$$\mathcal{L} = \{(\mathbf{x}_1, [A_1, B_1]), \ldots, (\mathbf{x}_n, [A_n, B_n])\}$$

  where
  - $\mathbf{x}_i \in \mathbb{R}^p$ is a feature vector for instance i
  - A_i is the set of classes that certainly apply to instance i;
  - B_i is the set of classes that possibly apply to that instance.
- In a multi-expert context, A_i may be the set of classes assigned to instance i by all experts, and B_i the set of classes assigned by some experts.

Multi-label evidential k-NN rule

Construction of mass functions

- Let $\mathcal{N}_k(\mathbf{x})$ be the set of k nearest neighbors of a new instance $\mathbf{x}$, according to some distance measure d.
- Let $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ with label $[A_i, B_i]$. This item of evidence can be described by the following mass function in $(\mathcal{I}, \subseteq)$:

$$
\begin{aligned}
m_i([A_i, B_i]) &= \varphi\left(d_i\right), \\
m_i([\emptyset_\Theta, \Theta]) &= 1 - \varphi\left(d_i\right),
\end{aligned}
$$

  where φ is a decreasing function from $[0, +\infty)$ to $[0, 1]$ such that $\lim_{d \to +\infty} \varphi(d) = 0$.
- The k mass functions are combined using Dempster's rule:

$$
m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i
$$

Multi-label evidential k-NN rule

Decision

- Let $\widehat{Y}$ be the predicted label set for instance **x**.
- To decide whether to include in $\widehat{Y}$ each class $\theta \in \Theta$ or not, we compute
  - the degree of belief $bel([\{\theta\}, \Theta])$ that the true label set Y contains θ, and
  - the degree of belief $bel([\emptyset, \overline{\{\theta\}}])$ that it does not contain θ.
- We then define $\widehat{Y}$ as

$$\widehat{Y} = \{\theta \in \Theta \mid bel([\{\theta\}, \Theta]) \geq bel([\emptyset, \overline{\{\theta\}}])\}.$$

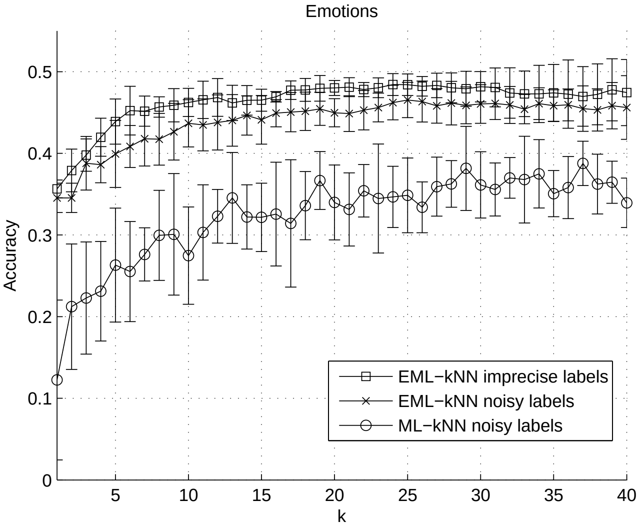
- Other method: find the set of labels $\widehat{Y}$ with the largest plausibility (linear programming problem).

Example: emotions data (Trohidis et al. 2008)

- Problem: Predict the emotions generated by a song.
- 593 songs were annotated by experts according to the emotions they generate.
- The emotions were: amazed-surprise, happy-pleased, relaxing-calm, quiet-still, sad-lonely and angry-fearful.
- Each song was described by 72 features and labeled with one or several emotions (classes).
- The dataset was split in a training set of 391 instances and a test set of 202 instances.
- Evaluation of results:

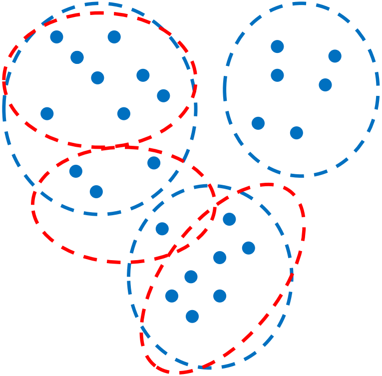
$$Acc = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|}$$

Results


Emotions
Accuracy
k
0
0.1
0.2
0.3
0.4
0.5
5
10
15
20
25
30
35
40
EML−kNN imprecise labels
EML−kNN noisy labels
ML−kNN noisy labels

Outline

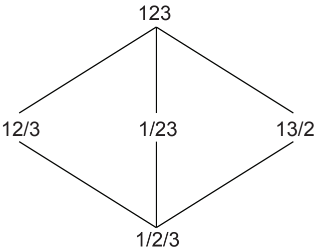
Problem statement



- Clustering may be defined as the search for a partition of a set E of n objects.
- The natural frame of discernment for this problem is the set $\mathcal{P}(E)$ of partitions of E, with size s_n.
- Expressing such evidence in the Dempster-Shafer framework implies working with sets of partitions.

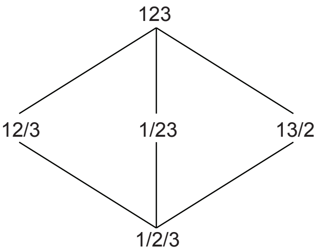
n	3	4	5	6	7
s_n	5	15	52	203	876
2^{s_n}	23	32768	4.5e15	1.3e61	5.0e263

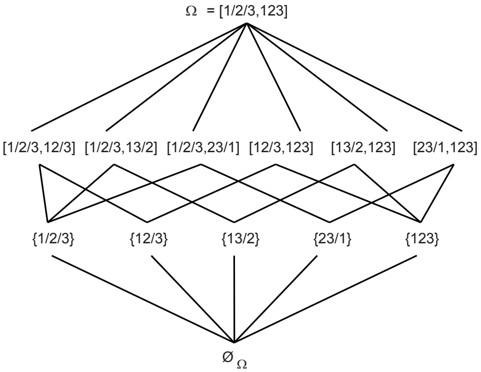
Lattice of partitions of a finite set



- A partition p is said to be finer than a partition p' (or, equivalently p' is coarser than p) if the clusters of p can be obtained by splitting those of p'; we write $p \preceq p'$.
- The poset $(\mathcal{P}(E), \preceq)$ is a lattice.

Lattices of partition intervals ($n = 3$)





13 partition intervals $< 2^5 = 32$ sets of partitions.

Ensemble clustering

- Ensemble clustering aims at combining the outputs of several clustering algorithms ("clusterers") to form a single clustering structure (crisp or fuzzy partition, hierarchy).
- This problem can be addressed using evidential reasoning by assuming that:
  - There exists a "true" partition p^*;
  - Each clusterer provides evidence about p^*;
  - The evidence from multiple clusterers can be combined to draw plausible conclusions about p^*.
- To implement this scheme, we need to manipulate Dempster-Shafer mass functions, the focal elements of which are sets of partitions.
- This is feasible by restricting ourselves to intervals of the lattice $(\mathcal{P}(E), \preceq)$.

Method

Mass construction and combination

- Compute r partitions $p_1, \ldots, p_r$ with large numbers of clusters using, e.g., the FCM algorithm.
- For each partition p_k, compute a validity index α_k.
- The evidence from clusterer k can be represented as a mass function

$$\begin{cases} m_k([p_k, p_E]) = \alpha_k \\ m_k([p_0, p_E]) = 1 - \alpha_k, \end{cases}$$

where p_E is the coarsest partition.
- The r mass functions are combined using Dempster's rule:

$$m = m_1 \oplus \ldots \oplus m_r$$

Method

Exploitation of the results

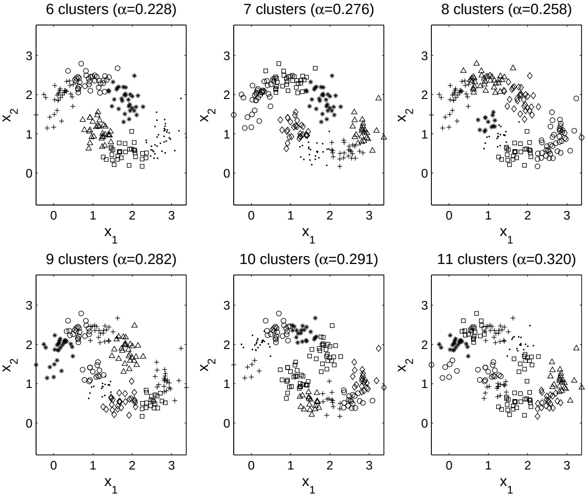
- Let p_{ij} denote the partition with $(n-1)$ clusters, in which objects i and j are clustered together.
- The interval $[p_{ij}, p_E]$ is the set of all partitions in which objects i and j are clustered together.
- The degree of belief in the hypothesis that i and j belong to the same cluster is then:

$$Bel_{ij} = bel([p_{ij}, p_E]) = \sum_{[\underline{p}_k, \overline{p}_k] \subseteq [p_{ij}, p_E]} m([\underline{p}_k, \overline{p}_k])$$

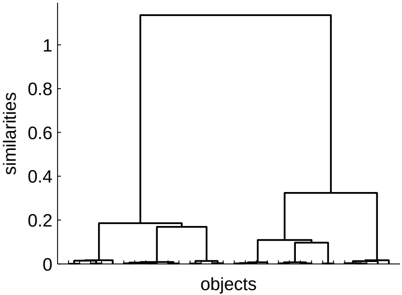
- Matrix $Bel = (Bel_{ij})$ can be considered as a new similarity matrix and can be processed by, e.g., a hierarchical clustering algorithm.

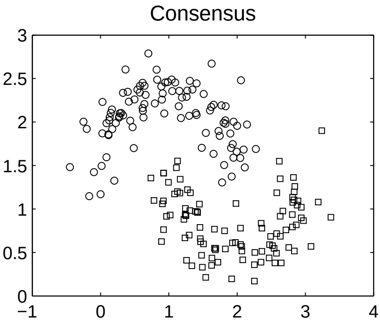
Results

Individual partitions



Results
Synthesis

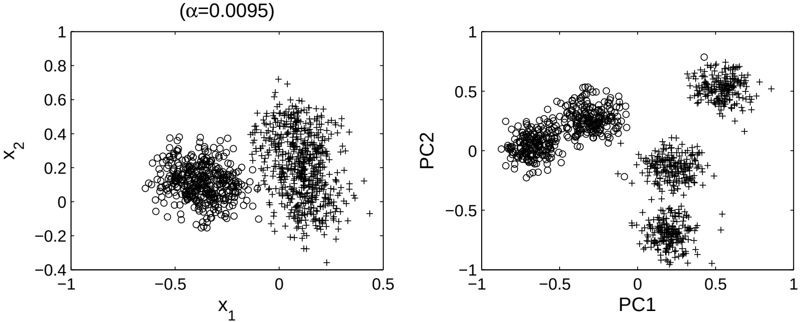




Distributed clustering

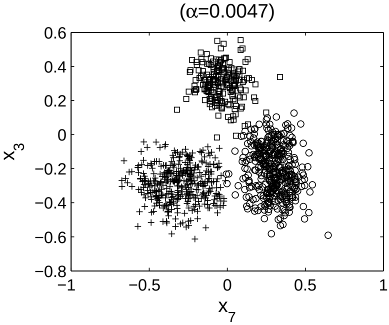
8D5K data (Strehl and Gosh, 2002)

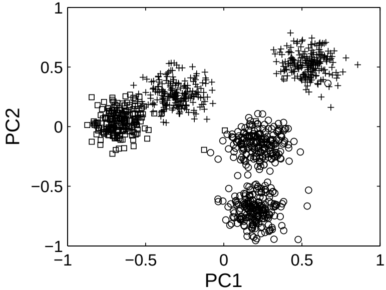
Gaussian data, 8 features, 5 clusters



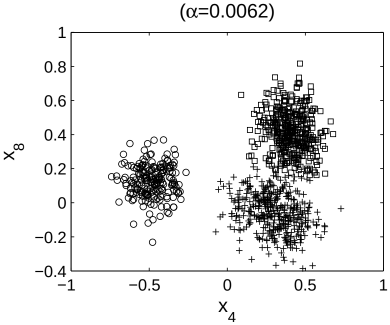
Distributed clustering

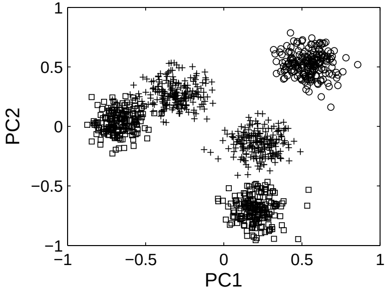
8D5K data (Strehl and Gosh, 2002)





Distributed clustering
8D5K data (Strehl and Gosh, 2002)





Distributed clustering
Method

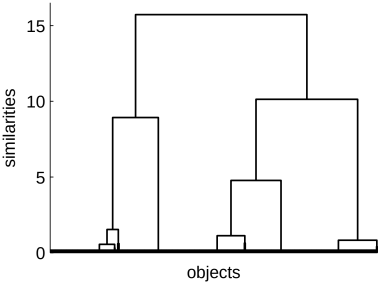
- Here, each clusterer provides a partition p_k that tends to be coarser than the true partition p^*.
- The output from clusterer k can be represented as a mass function

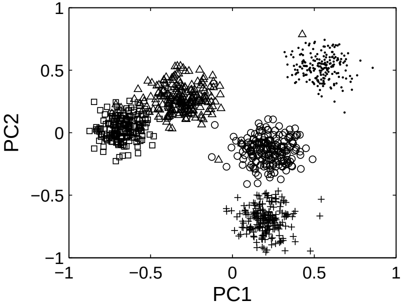
$$\begin{cases} m_k([p_0, p_k]) = \alpha_k \\ m_k([p_0, p_E]) = 1 - \alpha_k. \end{cases}$$

- As before, the mass functions are combined and synthesized in the form of a similarity matrix.

Distributed clustering

Consensus





Conclusion

- The exponential complexity of operations in the theory of belief functions has long been prevented its application to very large frames of discernment.
- When the frame of discernment has a lattice structure, it is possible to restrict the set of events to intervals in that lattice.
- This approach drastically reduces the complexity of the Dempster-Shafer calculus and makes it possible to define and manipulate belief functions in very large frames.
- This approach opens the way to the application of Dempster-Shafer theory to computationally demanding Machine Learning tasks such as multi-label classification and ensemble clustering.

References

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

T. Denœux, Z. Younes and F. Abdallah.
Representing uncertainty on set-valued variables using belief functions.
Artificial Intelligence, 174(7-8):479-499, 2010.

M.-H. Masson and T. Denœux.
Ensemble clustering in the belief functions framework.
International Journal of Approximate Reasoning, 52(1):92-109, 2011.

T. Denœux and M.-H. Masson.
Evidential reasoning in large partially ordered sets.
Annals of Operations Research, Accepted for publication, 2011.