Dempster-Shafer reasoning in large partially ordered sets

Applications in Machine Learning

Thierry Denœux[1], Marie-Hélène Masson[2]

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

[2]Université de Picardie Jules Verne
HEUDIASYC, UMR CNRS 6599

IUM 2010
9-11 April 2010, Ishikawa, Japan

Motivation

Generality of belief functions

- The theory of Belief Functions (Dempster-Shafer theory) is a rich framework for representing and reasoning with uncertainty.
- The expressive power of BF theory comes from the fact that it generalizes both set-valued (logical) and probabilistic representations of uncertainty.
- As a consequence, it allows us to express various kinds of uncertainty such as aleatory and epistemic uncertainty.

Motivation

Complexity of belief functions

- The generality and representation power of belief functions comes at a cost: a higher complexity than probabilistic reasoning.
- In the worst case, representing beliefs on a finite domain (frame of discernment) of size K requires the storage of $2^K - 1$ numbers, and operations on belief functions have exponential complexity.
- The application of belief functions to problems involving very large frames of discernment poses severe difficulties.
- What do we mean by “very large frames”?

Problems with "very large frames"

Multi-label classification

- Problems where 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 $\Theta = \{\theta_1, \ldots, \theta_K\}$ denotes the set of classes, the class label of an instance may be represented by a variable X taking values in $\Omega = 2^\Theta$.
- Expressing partial knowledge of X in the Dempster-Shafer framework may imply storing 2^{2^K} numbers.

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

Problems with "very large frames"

Clustering

- Problem: find a partition of a set E of n objects.
- Let p^* denote the "true" partition (assumed to exist).
- Variable p^* takes values in the set $\mathcal{P}(E)$ of partitions of E, with size s_n.
- A clustering algorithm can be seen as providing an item of evidence about p^*.
- Expressing such evidence in the Dempster-Shafer framework implies working with sets of partitions.
- There are 2^{s_n} such sets.

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

Approach

- Problem: How can the Dempster-Shafer framework be applied to such problems involving huge frames of discernment?
- Basic idea: exploit a special structure of the frame of discernment so as to restrict the form of belief functions, without losing too much flexibility.
- 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. Apply the Dempster-Shafer calculus in the lattice $(\mathcal{I}, \subseteq)$.

Outline

Outline

Mass functions
Definition

- A (normalized) mass function on a finite set Ω is a function $m : 2^{\Omega} \to [0,1]$ such that $m(\emptyset) = 0$ and
$$\sum_{A \subseteq \Omega} m(A) = 1.$$
- The subsets A of Ω such that $m(A) > 0$ are called the focal elements of m.
- A mass function m is often used to model a piece of evidence about a variable X.
- The quantity $m(A)$ can be interpreted as a measure of the belief that is committed exactly to the proposition $X \in A$.

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.
- 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 function

- Definition:

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

- Interpretation: $bel(A)$ = total degree of justified belief in A.
- Conversely,

$$m(A) = \sum_{B \subseteq A} (-1)^{|A \setminus B|} bel(B)$$

(m is called the Möbius transform of bel).

- m and bel are thus two equivalent representations of a belief state about a variable X. (There are others: plausibility, commonality, ...)

Characterization of belief functions

- For every function f from 2^Ω to $[0,1]$ such that $f(\emptyset)=0$ and $f(\Omega)=1$, the following conditions are known to be equivalent (Shafer, 1976):
  1. The Möbius transform m of f is a positive;
  2. f is totally monotone, i.e., for any $k \geq 2$ and for any family $A_1,\ldots,A_k$ in 2^Ω,

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

- A belief function can be characterized by any one of these two properties.

Dempster's rule

- Let m_1 and m_2 be two mass functions on Ω induced by two distinct items of evidence. How should they be combined?
- Dempster's rule:

$$(m_1 \oplus m_2)(A) = \begin{cases} \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases}$$

with $\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$: degree of conflict.

- This rule is commutative, associative, and admits the vacuous mass function ($m(\Omega) = 1$) as neutral element.

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 $\rightarrow m_2(\{John, Mary\}) = 0.6$, $m_2(\Omega) = 0.4$.

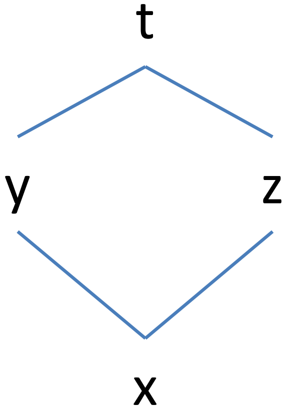
	$\{Peter, John\}$ 0.8	Ω 0.2
$\{John, Mary\}$ 0.6	$\{John\}$ 0.48	$\{John, Mary\}$ 0.12
Ω 0.4	$\{Peter, John\}$ 0.32	Ω 0.08

Outline

Lattices
Definitions

- Let L be a finite set and $\leq$ a partial ordering on L. $(L, \leq)$ is called a poset.
- We say that $(L, \leq)$ is a lattice if, for every $x, y \in L$, there is a unique greatest lower bound (denoted $x \wedge y$) and a unique least upper bound (denoted $x \vee y$).
- Operations $\wedge$ and $\vee$ are called the meet and join operations, respectively.
- For finite lattices, the greatest element (denoted $\top$) and the least element (denoted $\bot$) always exist.

Example



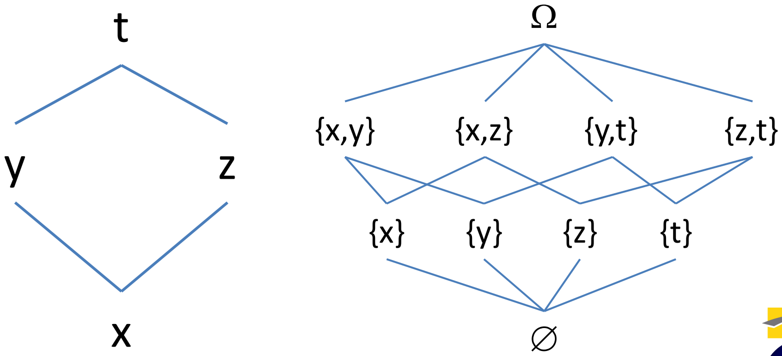
Lattice intervals

- Let $(L, \leq)$ be lattice.
- A (lattice) interval of L is defined as

$$[a, b] = \{x \in L | a \leq x \leq b\}$$

  for some a and b in L.
- Let $\mathcal{I} \subseteq 2^L$ be the set of intervals, including the empty set of L.
- The poset $(\mathcal{I}, \subseteq)$ is a lattice with
  - meet = intersection;
  - joint defined by $[a, b] \sqcup [c, d] = [a \wedge c, b \vee d]$;
  - least element = $\emptyset_L$
  - greatest element = L.

Example



Belief functions on lattices

- Belief functions are usually defined on the Boolean lattice $(2^\Omega, \subseteq)$.
- However, they can be defined on any lattice, not necessarily Boolean (Grabisch, 2009).
- Most of the above definitions and formula can be translated into this very general setting.

Mass and belief functions

- Let $(L, \leq)$ be a finite lattice.
- A (normalized) mass function on L is a function $L \rightarrow [0, 1]$ such that $m(\perp) = 0$ and

$$\sum_{x \in L} m(x) = 1.$$

- Corresponding belief function:

$$bel(x) = \sum_{y \leq x} m(y), \quad \forall x \in L.$$

Equivalence of representations

- m can be recovered from bel:

$$m(x) = \sum_{y \leq x} \mu(y, x) bel(y),$$

  where $\mu(x, y) : L^2 \rightarrow \mathbb{R}$ is the Möbius function, which is uniquely defined for each poset $(L, \leq)$.
- A belief function is totally monotone, but the converse is not true in general.

Dempster's rule

- Dempster's rule can be defined as in the Boolean case:

$$(m_1 \oplus m_2)(x) = \begin{cases} \frac{1}{1-\kappa} \sum_{y \wedge z = x} m_1(y) m_2(z) & \text{if } x \neq \bot \\ 0 & \text{if } x = \bot \end{cases}$$

with $\kappa = \sum_{y \wedge z = \bot} m_1(y) m_2(z)$.

Belief functions in $(\mathcal{I}, \subseteq)$

- Let Ω be the domain of a variable X, with $|\Omega| = K$.
- If Ω has a lattice structure for some partial ordering $\leq$, then uncertain knowledge about X may be encoded as a belief function on the lattice $(\mathcal{I}, \subseteq)$ of intervals of $(\Omega, \leq)$.
- As the cardinality of $\mathcal{I}$ is at most proportional to K^2, all the operations of Demspter-Shafer theory can be performed in polynomial time (instead of exponential when working in $(2^\Omega, \subseteq)$).

Outline

Disjunctive vs conjunctive variables

- Let Θ be a finite domain. A variable X may take
  - One and only one value in Θ: disjunctive variable (usual case);
  - Several values in Θ simultaneously: conjunctive variable.
- For instance, Θ may be a set of faults, and X the faults actually occurring at a given time (under the assumption that multiple faults can occur).
- The Dempster-Shafer framework is usually applied to express partial knowledge about disjunctive variables.
- How to extend it to conjunctive variables?

Proposed framework

- X takes values in $\Omega = 2^{\Theta}$.
- Standard approach: define belief functions on $(2^{\Omega}, \subseteq)$ (intractable).
- Proposed approach: exploit the lattice structure induced by the ordering $\subseteq$ in Ω and apply the above general framework.
- The intervals of the lattice $(\Omega, \subseteq)$ are sets of subsets of Θ of the form:

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

for some subsets A and B of Θ.

Interpretation

- A certain piece of information $X \in [A, B]$ tell us that the unknown set X surely contains all elements of A, and possibly contains elements of B.
- An uncertain piece of information about the unknown set X can be modeled by a mass function with focal elements of the form $[A_i, B_i]$, $i = 1, \ldots, n$.

Example

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

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

- Item of evidence 2: c is present and a, b may also be present, with confidence 0.8. This is represented by

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

Example
Combination

- Conjunctive combination

	$[\{a\},\{a,b,c\}]$ 0.7	$[\emptyset_\Theta,\Theta]$ 0.3
$[\{c\},\{a,b,c\}]$ 0.8	$[\{a,c\},\{a,b,c\}]$ 0.48	$[\{c\},\{a,b,c\}]$ 0.12
$[\emptyset_\Theta,\Theta]$ 0.2	$[\{a\},\{a,b,c\}]$ 0.32	$[\emptyset_\Theta,\Theta]$ 0.08

- Based on this evidence, what is our belief that fault a is present?

$$bel([\{a\},\Theta]) = 0.48 + 0.32 + 0.08 = 0.88$$

Multi-label classification

Example (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 emotion corresponds to a class.
- Each song was
  - described by 72 features;
  - 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.
- How to learn a classifier from such data?

Multi-label classification
Learning data

- In multi-label classification problems, data are usually to have the following form:

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

  where
  - $\mathbf{x}_i \in \mathbb{R}^d$ is a feature vector for instance i
  - A_i is the set of classes that apply to instance i.
- When data are labeled by one or several experts, this format does not allow us to express uncertainty on class labels due to doubt or disagreement between experts.

Multi-label classification
Imprecise labels

- More general form considered here:

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

  where
  - 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

- Generalization of the evidential k-NN rule (Denœux,1995).
- 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 $(\mathcal{I}, \subseteq)$:

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

  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

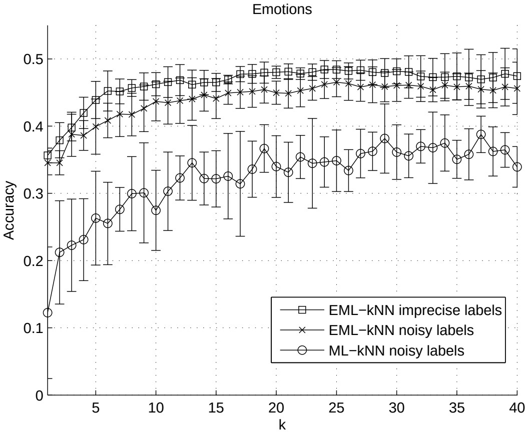
Decision

- Let $\widehat{Y}$ be the predicted label set for instance $\mathbf{x}$.
- To decide whether to include 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

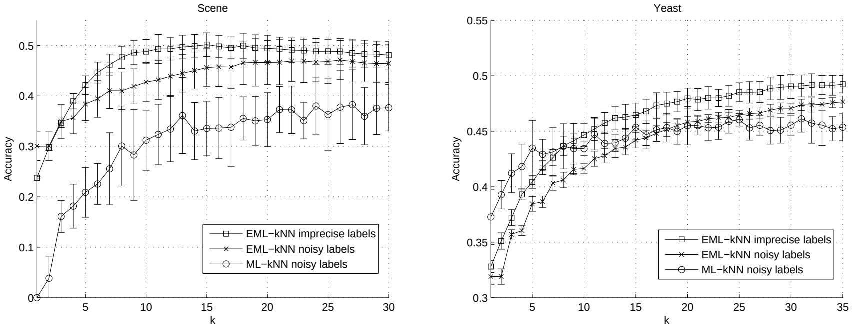
$$\widehat{Y} = \{\theta \in \Theta \mid bel([\{\theta\}, \Theta]) \geq bel([\emptyset, \overline{\{\theta\}}])\}.$$

Emotions data
Results

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



Results on other data sets



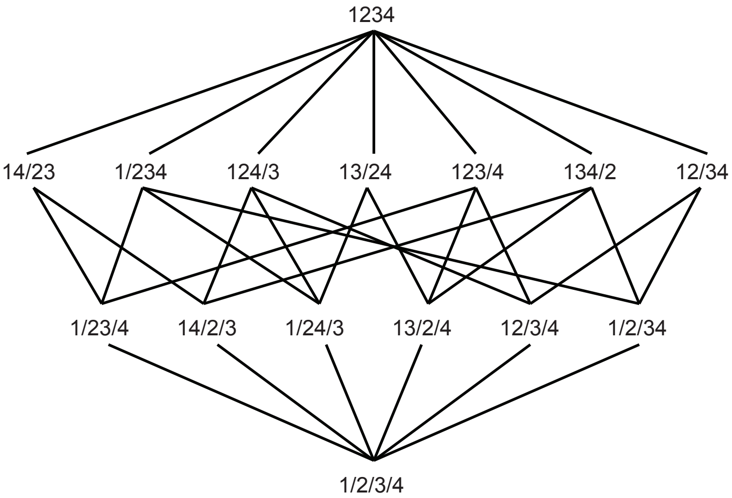
Outline

Partitions of a finite set

Ordering relation

- In clustering, the frame of discernment is the set of all partitions of a finite set E, denoted $\mathcal{P}(E)$.
- This set can be partially ordered using the following relation:
- 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.

Example: lattice of partitions of a four-element set



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 feasable 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

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

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

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

Method

Exploitation of the results

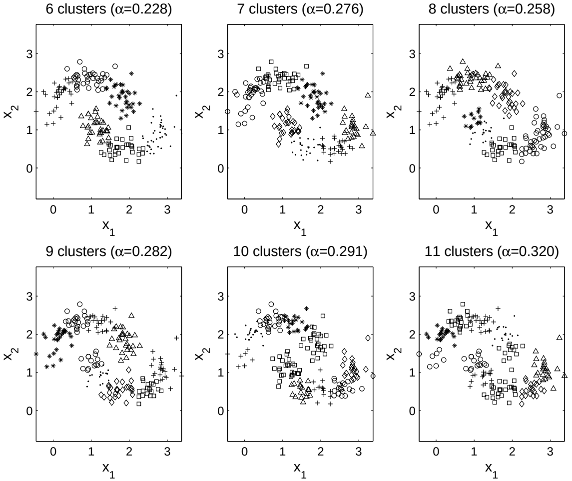
- Let p_{ij} denote the partition with $(n-1)$ clusters, in which objects i and j are clustered together.
- The interval $[p_{ij}, p_E]$ is the set of all partitions in which objects i and j are clustered together.
- The degree of belief in the hypothesis that i and j belong to the same cluster is then:

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

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

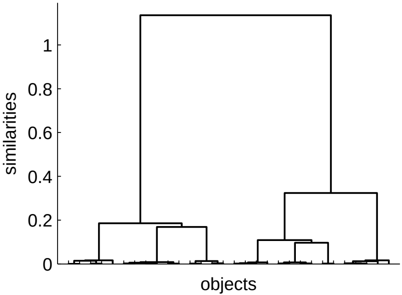
Results

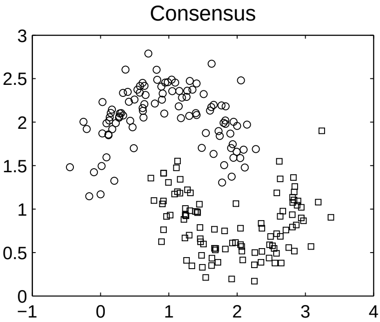
Individual partitions



Results

Synthesis

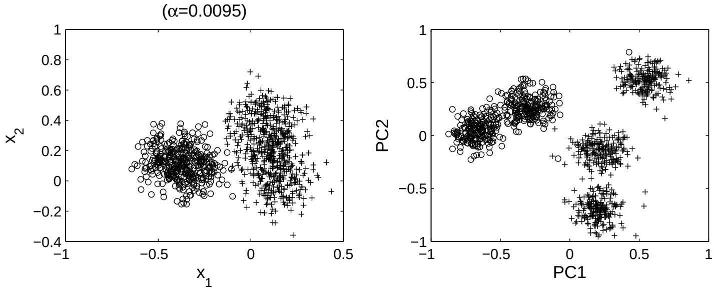




Distributed clustering

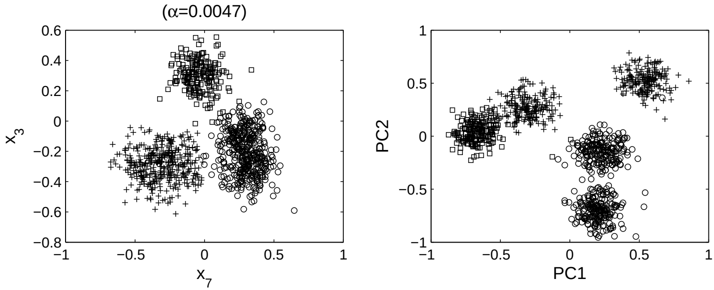
8D5K data (Strehl and Gosh, 2002)

Gaussian data, 8 features, 5 clusters



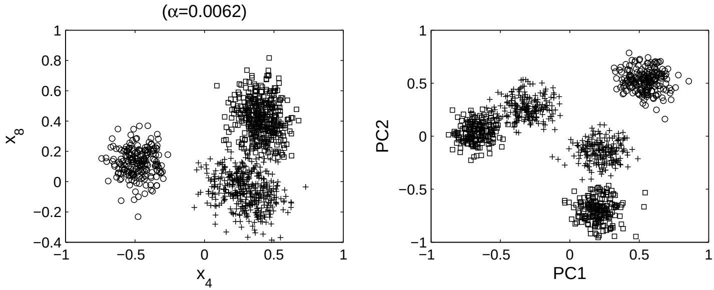
Distributed clustering

8D5K data (Strehl and Gosh, 2002)



Distributed clustering

8D5K data (Strehl and Gosh, 2002)



Distributed clustering

Method

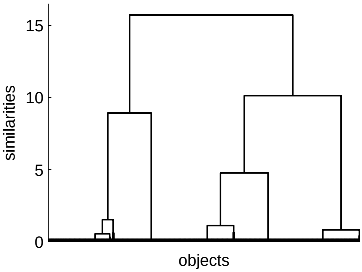
- Here, each clusterer provides a partition p_k that tends to be coarser than the true partition p_k.
- The output from clusterer k can be represented as a mass function

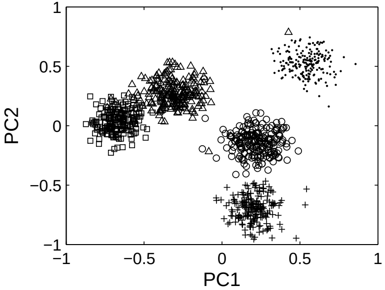
$$\begin{cases} m_k([p_0, p_k]) = \alpha_k \\ m_k([p_0, p_E]) = 1 - \alpha_k. \end{cases}$$

- 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, To appear, 2010.
doi:10.1016/j.artint.2010.02.002

M.-H. Masson and T. Denœux.
Belief Functions and Cluster Ensembles.
In C. Sossai and G. Chemello (Eds.): ECSQARU 2009, LNAI 5590, pp. 323-334, 2009. Springer-Verlag.