Theory of belief functions for data analysis and machine learning applications

Thierry Denœux

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

KSEM 2010
1-3 September 2010,
Belfast, Northern Ireland, UK

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) in the 1980's and 1990's.
- A belief function may be viewed both as
  - a generalized set and
  - a non additive measure,

  and the theory includes extensions of probabilistic notions (conditioning, marginalization) and set-theoretic notions (intersection, union, inclusion, etc.).
- The theory of belief functions thus generalizes both the set-membership and probabilistic approaches to uncertain reasoning.

Applications of Dempster-Shafer Theory

- Initially introduced as a formal framework for statistical inference (Dempster, 1968), but the theory has not (yet) become very popular among statisticians.
- Increasing number of applications in engineering:
  - Expert systems (1980's);
  - Information fusion (since the 1990's);
  - Statistical pattern recognition and Machine Learning (since the 2000's).

Outline

Outline

Mass functions
Definition

- Let Ω be a finite set of possible answers to some question: frame of discernment.
- A mass function on Ω is a function $m : 2^\Omega \to [0,1]$ such that

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

- 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 taking values in Ω.
- The quantity $m(A)$ can be interpreted as a measure of the belief that is committed exactly to the proposition $X \in A$.

Mass functions
Special cases

- Only one focal element:

$$m(A) = 1 \text{ for some } A \subseteq \Omega$$

  $\rightarrow$ categorical mass function ($\sim$ set). Special case: $A = \Omega$, vacuous mass function, represents total ignorance.

- All focal elements are singletons:

$$m(A) > 0 \Rightarrow |A| = 1$$

  $\rightarrow$ Bayesian mass function ($\sim$ probability mass function).

- A Dempster-Shafer mass function can thus be seen as
  - a generalized set;
  - a generalized probability distribution.

Example

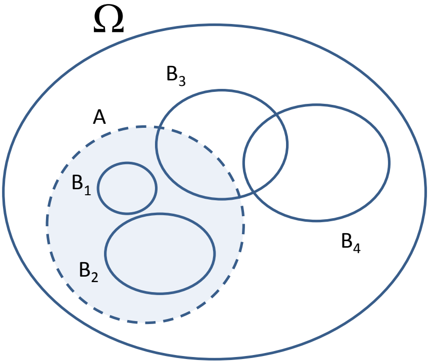
- A murder has been committed. There are three suspects: $\Omega = \{Peter, John, Mary\}$.
- A witness saw the murderer going away, but he is short-sighted and he only saw that it was a man. We know that the witness likes Irish pubs and is drunk 20 % of the time.
- This piece of evidence can be represented 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 and plausibility functions
Definitions



$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$$

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B).$$

Belief and plausibility functions
Interpretation and special cases

- 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 if more specific information became available ($\geq bel(A)$).
- Special cases:
  - If m is Bayesian, $bel = pl$ (probability measure).
  - If the focal elements are nested, pl is a possibility measure, and bel is the dual necessity measure.

Outline

1. Theory of belief functions
   - Belief representation
   - Combination
2. Application to classification and clustering
   - Supervised Classification
   - Clustering
3. Working in very large frames
   - General approach
   - Multi-label classification
   - Ensemble Clustering

Dempster's rule

- Let m_1 and m_2 be two mass functions on Ω obtained from independent sources of information. How should they be combined?
- Dempster's rule:

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

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

- This rule is commutative, associative, and admits the vacuous mass function as neutral element.

Dempster's rule
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

Dempster's rule
Properties

- 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.

Outline

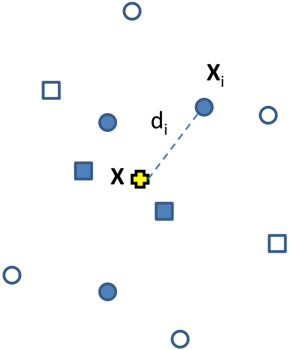
Problem statement

- A population is assumed to be partitioned in c groups or classes.
- Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ denote the set of classes.
- Each instance is described by
  - A feature vector $\mathbf{x} \in \mathbb{R}^p$;
  - A class label $y \in \Omega$.
- Problem:
  - Given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$,
  - Predict the class label of a new instance described by $\mathbf{x}$.

Main belief function approaches

1. Approach 1: Convert the outputs from standard classifiers into belief functions and combine them using Dempster's rule or any other alternative rule (e.g., Bi et al., *Art. Intell.*, 2008);
2. Approach 2: Develop evidence-theoretic classifiers directly providing belief functions as outputs:
   - Generalized Bayes theorem, extends the Bayesian classifier when class densities and priors are ill-known (Denœux and Smets, *IEEE SMC*, 2008);
   - Distance-based approach: evidential k-NN rule (Denœux, *IEEE SMC*, 1995), evidential neural network classifier (Denœux, *IEEE SMC*, 2000).

Evidential k-NN rule (1/2)



- Let $\mathcal{N}_k(\mathbf{x}) \subset \mathcal{L}$ denote the set of the k nearest neighbors of $\mathbf{x}$ in $\mathcal{L}$, based on some distance measure.
- Each $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ can be considered as a piece of evidence regarding the class of $\mathbf{x}$.
- The strength of this evidence decreases with the distance d_i between $\mathbf{x}$ and $\mathbf{x}_i$.

Evidential k-NN rule (2/2)

- The evidence of $(\mathbf{x}_i, y_i)$ can be represented by

$$m_i(\{y_i\}) = \varphi\left(d_i\right)$$

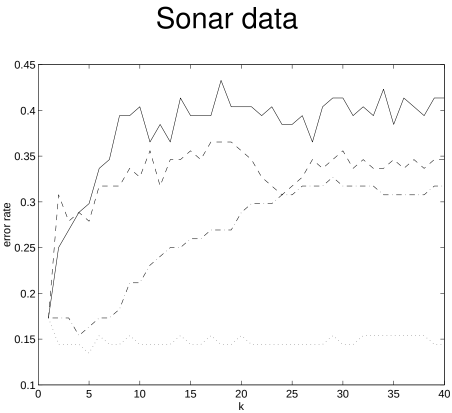
$$m_i(\Omega) = 1 - \varphi\left(d_i\right),$$

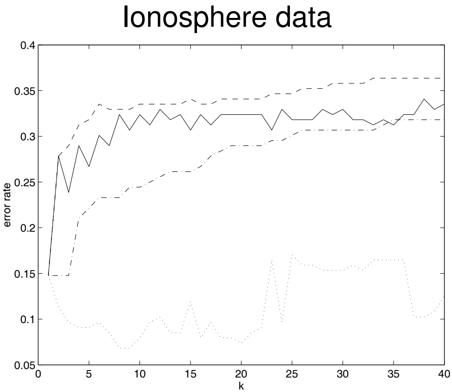
  where φ is a decreasing function from $[0, +\infty)$ to $[0, 1]$ such that $\lim_{d \to +\infty} \varphi(d) = 0$.
- The evidence of the k nearest neighbors of $\mathbf{x}$ is pooled using Dempster's rule of combination:

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

- Function φ can be fixed heuristically or selected among a family $\{\varphi_\theta | \theta \in \Theta\}$ using, e.g., cross-validation.

Performance comparison (UCI database)





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

Partially supervised data

- We now consider a learning set of the form

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

  where
  - $\mathbf{x}_i$ is the attribute vector for instance i, and
  - m_i is a mass function representing uncertain expert knowledge about the class y_i of instance i.
- Special cases:
  - $m_i(\{\omega_k\}) = 1$ for all i: supervised learning;
  - $m_i(\Omega) = 1$ for all i: unsupervised learning;

Evidential k-NN rule for partially supervised data

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

$$m_i'(A) = \varphi(d_i)\, m_i(A), \quad \forall A \subset \Omega.$$

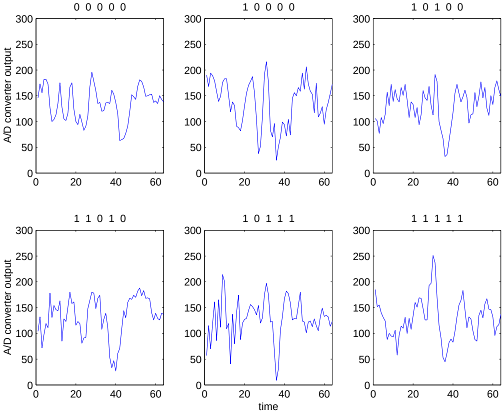
$$m_i'(\Omega) = 1 - \sum_{A \subset \Omega} m_i'(A).$$

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

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

Example: EEG data

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



Results on EEG data
(Denoeux and Zouhal, 2001)

- $c = 2$ classes, $p = 64$
- For each learning instance $\mathbf{x}_i$, the expert opinions were modeled as a mass function m_i.
- $n = 200$ learning patterns, 300 test patterns

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

Evidential neural network classifier

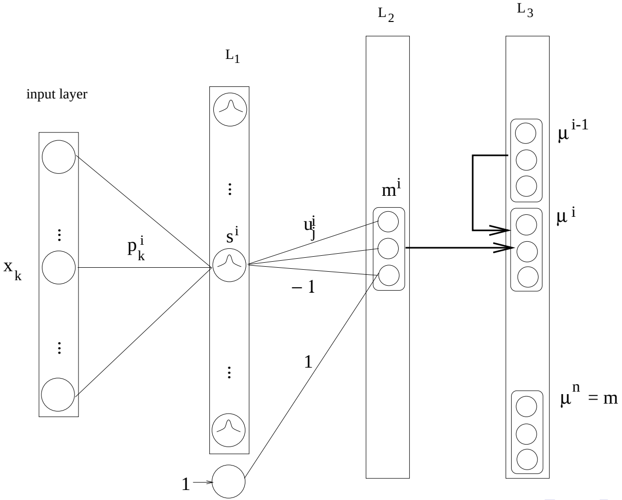
- Implementation in a RBF-like neural network architecture with r prototypes: $\mathbf{p}_1, \ldots, \mathbf{p}_r$.
- Each prototype $\mathbf{p}_i$ has membership degree u_{ik} to each class ω_k with $\sum_{k=1}^{c} u_{ik} = 1$
- The distance between $\mathbf{x}$ and $\mathbf{p}_i$ induces a mass function:

$$\begin{aligned} m_i(\{\omega_k\}) &= \alpha_i u_{ik} \exp(-\gamma_i \|\mathbf{x} - \mathbf{p}_i\|^2) \quad \forall k \\ m_i(\Omega) &= 1 - \alpha_i \exp(-\gamma_i \|\mathbf{x} - \mathbf{p}_i\|^2) \end{aligned}$$

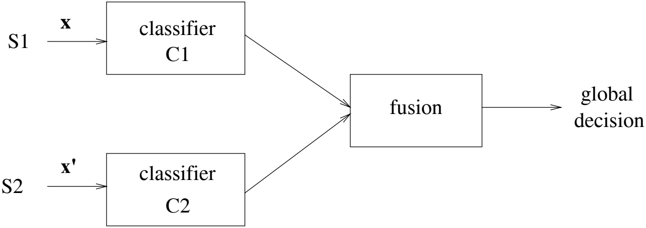
$$m = \bigoplus_{i=1}^{r} m_i$$

- Learning of parameters $\mathbf{p}_i$, u_{ik}, γ_i, α_i from data by minimizing an error function

Neural network architecture



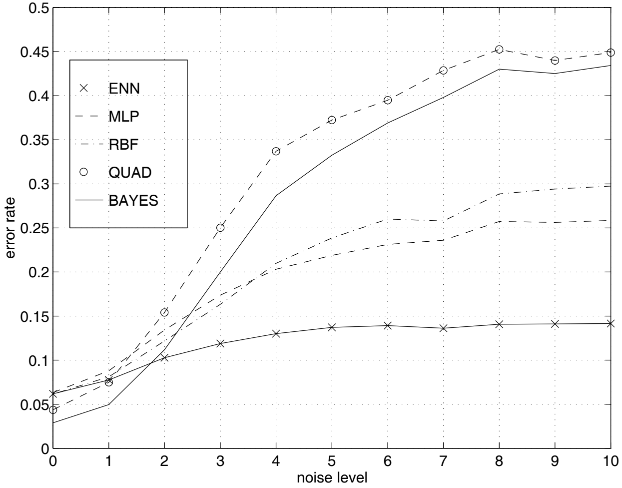
Data fusion example



- $c = 2$ classes
- Learning set ($n = 60$): $\mathbf{x} \in \mathbb{R}^5, \mathbf{x}' \in \mathbb{R}^3$, Gaussian distributions, conditionally independent
- Test set (real operating conditions): $\mathbf{x} \leftarrow \mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

Results

Test error rates: $\mathbf{x}+\epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$



Outline

Relational clustering

- 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?

Credal partition

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

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

Hard and fuzzy partitions are recovered as special cases when all mass functions are certain or Bayesian, respectively.

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 m_i and m_j be mass functions regarding the class membership of objects o_i and o_j.
- The plausibility of the proposition S_{ij}: “objects o_i and o_j belong to the same class” can be shown to be equal to:

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

  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

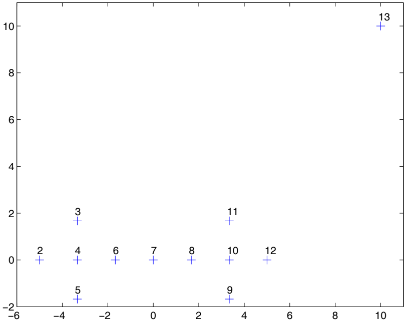
- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict K_{ij}.
- Example of a cost function:

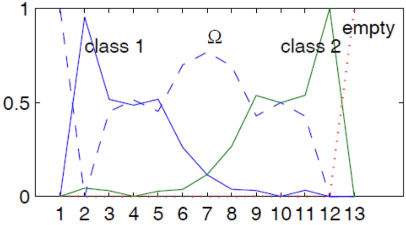
$$J(M) = \sum_{i<j} \left(K_{ij} - d_{ij}\right)^2$$

(assuming the d_{ij} have been scaled to $[0,1]$).

- M can be determined by minimizing J using an alternate directions method, solving a QP problem at each step.
- To reduce the complexity, focal sets can be reduced to $\{\omega_k\}_{k=1}^c$, $\emptyset$, and Ω.

Butterfly example



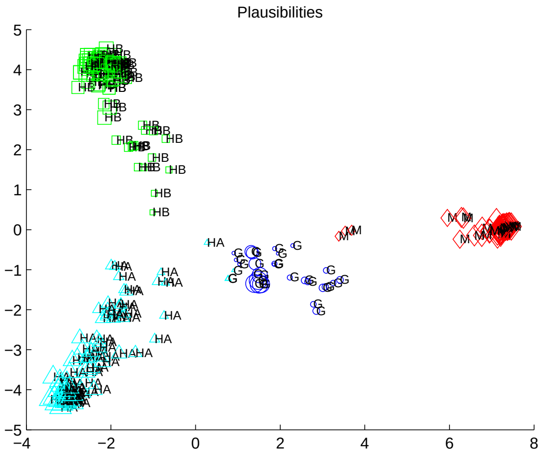


one additional object (#1) similar to all other objects (“inlier”)

Protein dataset

- Proximity matrix derived from the structural comparison of 213 protein sequences.
- Each of these proteins is known to belong to one of four classes of globins: hemoglobin-α (HA), hemoglobin-β (HB), myoglobin (M) and heterogeneous globins (G).
- Non-metric dissimilarities: most relational fuzzy clustering algorithms fail on this data (they converge to a trivial solution).
- EVCLUS recovers the true partition with only one error.

Protein dataset: result



Advantages and drawbacks

- Advantages
  - Applicable to proximity data (not necessarily Euclidean, or even numeric).
  - Robust against atypical observations (similar or dissimilar to all other objects).
  - Usually performs better than relational fuzzy clustering procedures.
- Drawback: computational complexity (iterative optimization, limited to datasets of a few thousand objects and less than 20 classes).
- More computationally efficient procedures: ECM (Masson and Denoeux, 2008), RECM (Masson and Denoeux, 2009), CECM (Antoine et al., 2010).

Outline

1. Theory of belief functions
   - Belief representation
   - Combination
2. Application to classification and clustering
   - Supervised Classification
   - Clustering
3. Working in very large frames
   - General approach
   - Multi-label classification
   - Ensemble Clustering

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 classification and clustering, the frame of discernment (set of classes) is usually of moderate size (less than 100). Can we address more complex problems 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*, 2010).

Multi-label classification

- Classification problems in which 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

Ensemble Clustering

- Clustering may be defined as the search for a partition of a set E of n objects.
- In ensemble clustering, we need to combine the outputs of several clustering algorithms (clusterers), regarded as item of evidence about the true partition p^*.
- 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

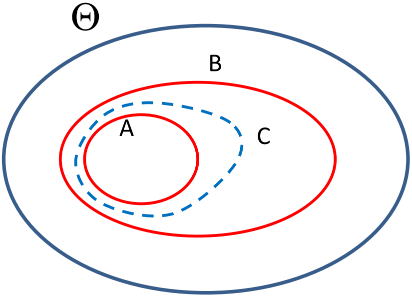
General Approach

- 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

- 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 $(\mathcal{I}, \subseteq)$:

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

  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

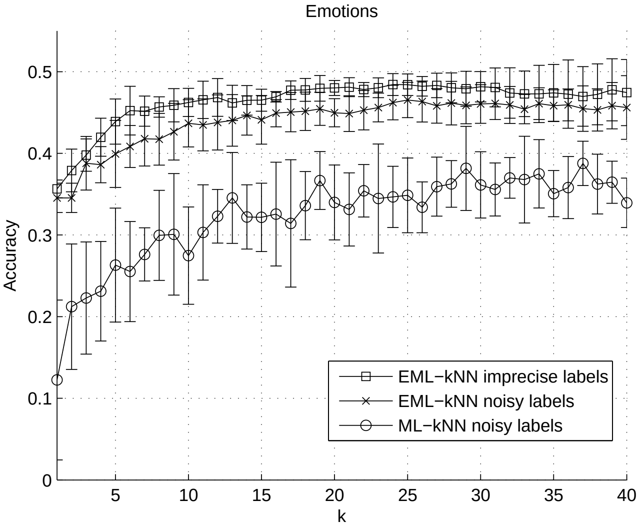
$$\widehat{Y} = \{\theta \in \Theta \mid bel([\{\theta\}, \Theta]) \geq bel([\emptyset, \overline{\{\theta\}}])\}.$$

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

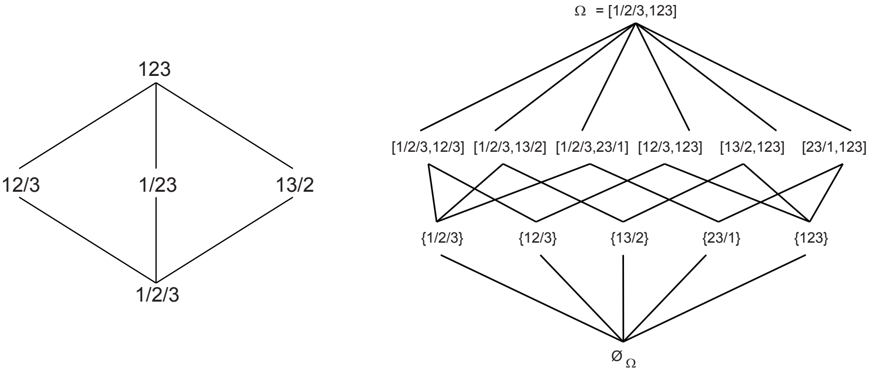

Emotions
Accuracy
k
0
0.1
0.2
0.3
0.4
0.5
5
10
15
20
25
30
35
40
EML−kNN imprecise labels
EML−kNN noisy labels
ML−kNN noisy labels

Outline

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 $\mathcal{P}(E)$.
- 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 partitions and 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

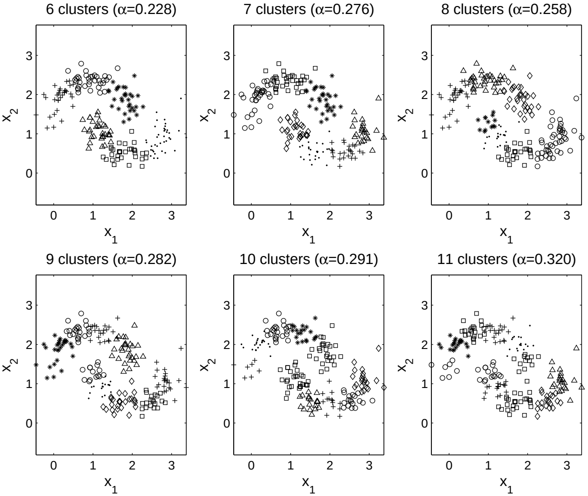
Exploitation of the results

- 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:

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

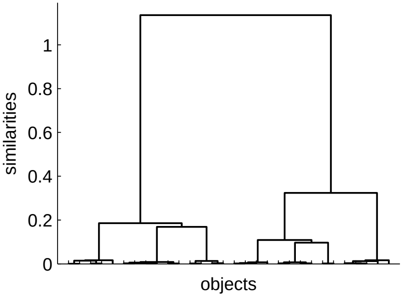
- Matrix $Bel = (Bel_{ij})$ can be considered as a new similarity matrix and can be processed by, e.g., a hierarchical clustering algorithm.

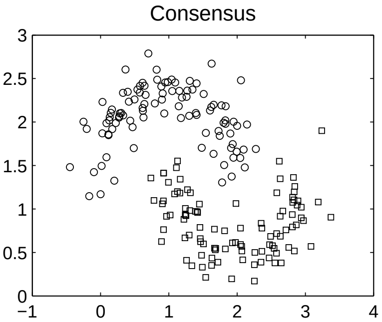
Results
Individual partitions



Results

Synthesis





Summary

- The theory of belief functions provides a very general framework for representing and reasoning with partial information.
- This framework has great potential to help solve complex machine learning problems, particularly those involving:
  - Weak information (partially labeled data, unreliable sensor data, etc.);
  - Both data and expert knowledge (constrained clustering);
  - Multiple sources of information (classifier or clustering ensembles). See also, e.g., Quost et al. (2007), Bi et al. (2008).

Research Challenges/Ongoing work

1. Developing more sophisticated classifer fusion schemes using
   - new combination rules allowing us to pool information from dependent and/or very conflicting sources (e.g., cautious rule and extensions, Denœux, *Art. Intell.*, 2008);
   - meta-knowledge about the quality (reliability) of information sources.
2. Addressing new challenging problems in Machine Learning:
   - Preference learning (using belief functions on sets of preference relations) ;
   - Learning from uncertain data (e.g., attributes or class labels).

References

Papers and Matlab software available at:

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

THANK YOU!