Handling imprecise and uncertain class labels in classification and clustering

Thierry Denœux[1]

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

COST Action IC 0702
Working group C,
Mallorca, March 16, 2009

Classification and clustering
Classical framework

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

Classification and clustering

Problems

- Problem 1: predict the class membership of objects drawn from the same population as $\mathcal{L}$ (classification).
- Problem 2: Estimate parameters of the population from which $\mathcal{L}$ is drawn (mixture model estimation).
- Problem 3: Determine the class membership of objects in $\mathcal{L}$ (clustering);

	supervised	unsupervised	semi-supervised
Problem 1	x		x
Problem 2	x	x	x
Problem 3		x	x

Motivations

- In real situations, we may have only partial knowledge of class labels: intermediate situation between supervised and unsupervised learning → partially supervised learning.
- The class membership of objects can usually be predicted with some remaining uncertainty: the outputs from classification and clustering algorithms should reflect this uncertainty.
- The theory of belief functions is suitable for representing uncertain and imprecise class information:
  - as input to classification and mixture model estimation algorithms;
  - as output of classification and clustering algorithms.

Outline

Mass function

- 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.
- Interpretation: m represents
  - An item of evidence regarding the value of X.
  - A state of knowledge (belief state) induced by this item of evidence.

Full notation

$$\boxed{\mathrm{m}^{\Omega}_{Ag,t}\{X\}[EC]}$$

denotes the mass function

- Representing the beliefs of agent Ag;
- At time t;
- Regarding variable X;
- Expressed on frame Ω;
- Based on evidential corpus EC.

Special cases

- m may be seen as:
  - A family of weighted sets $\{(A_i, m(A_i)), i = 1, \ldots, r\}$.
  - A generalized probability distribution (masses are distributed in 2^Ω instead of Ω).
- Special cases:
  - $r = 1$: categorical mass function ($\sim$ set). We denote by m_A the categorical mass function with focal set A.
  - $|A_i| = 1$, $i = 1, \ldots, r$: Bayesian mass function ($\sim$ probability distribution).
  - $A_1 \subset \ldots \subset A_r$: consonant mass function ($\sim$ possibility distribution).

Belief and plausibility functions

- Belief function:

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

(degree of belief (support) in hypothesis "$X \in A$")

- Plausibility function:

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

(upper bound on the degree of belief that could be assigned to A after taking into account new information)

- $bel \leq pl$.

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

- Relations:

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

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

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

Dempster's rule

Definition (Dempster's rule of combination)

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

- Properties:
  - Commutativity, associativity.
  - Neutral element: vacuous m_Ω such that $m_\Omega(\Omega) = 1$ (represents total ignorance).
  - $(m_1 \textcircled{\cap} m_2)(\emptyset) \geq 0$: degree of conflict.
- Justified axiomatically.
- Other rules exist (disjunctive rule, cautious rule, etc...).

Discounting

- Discounting allows us to take into account meta-knowledge about the reliability of a source of information.
- Let
  - m be a mass function provided by a source of information.
  - $\alpha \in [0, 1]$ be the plausibility that the source is not reliable.
- Discounting m with discount rate α yields the following mass function:

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

- Properties: $^{0}m = m$ and $^{1}m = m_{\Omega}$.

Pignistic transformation

- Assume that our knowledge about X is represented by a mass function m, and we have to choose one element of Ω.
- Several strategies:
  1. Select the element with greatest plausibility.
  2. Select the element with greatest pignistic probability:

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

(m assumed to be normal)

Outline

Problem

- Let Ω denote the set of classes, et $\mathcal{L}$ the learning set

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

  where
  - $\mathbf{x}_i$ is the attribute vector for object o_i, and
  - $m_i = m^{\Omega}\{y_i\}$ is a mass function on the class y_i of object o_i.
- Special cases:
  - $m_i(\{\omega_k\}) = 1$: precise labeling;
  - $m_i(A) = 1$ for $A \subseteq \Omega$: imprecise (set-valued) labeling;
  - m_i is a Bayesian mass function: probabilistic labeling;
  - m_i is a consonant mass function: possibilistic labeling, etc...
- Problem: Build a mass function $m^{\Omega}\{y\}[\mathbf{x}, \mathcal{L}]$ regarding the class y of a new object o described by $\mathbf{x}$.

Solution
(Denoeux, 1995)

- Each example $e_i = (\mathbf{x}_i, m_i)$ in $\mathcal{L}$ is an item of evidence regarding y.
- The reliability of this information decreases with the distance between $\mathbf{x}$ and $\mathbf{x}_i$. It should be discounted with a discount rate

$$\alpha_i = \phi\left(d(\mathbf{x}, \mathbf{x}_i)\right),$$

where ϕ is a decreasing function from $\mathbb{R}^+$ to $[0, 1]$:

$$m\{y\}[\mathbf{x}, e_i] = {}^{\alpha_i}m_i.$$

- The n mass functions should then be combined conjunctively:

$$m\{y\}[x, \mathcal{L}] = m\{y\}[\mathbf{x}, e_i] \circledcap \ldots \circledcap m\{y\}[\mathbf{x}, e_n].$$

Implementation

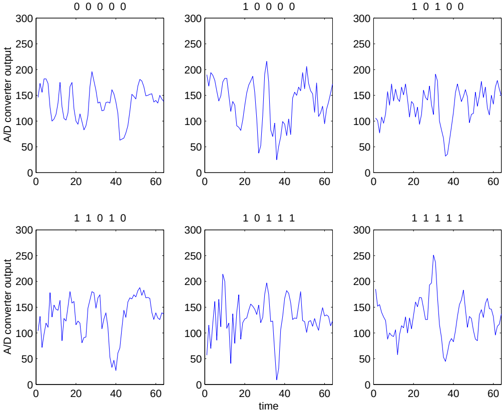
- Take into account only the k nearest neighbors of $\mathbf{x}$ dans $\mathcal{L}$ $\rightarrow$ evidential k-NN rule (generalizes the voting k-NN rule).
- Definition of ϕ: for instance,

$$\phi(d) = \beta \exp(-\gamma d^2).$$

- Determination of hyperparameters β and γ heuristically or by minimizing an error function (Zouhal and Denoeux, 1997).
- Summarize $\mathcal{L}$ as r prototypes learnt by minimizing an error function $\rightarrow$ RBF-like neural network approach (Denoeux, 2000).

Example: EEG data

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



Results on EEG data
(Denoeux and Zouhal, 2001)

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

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

Outline

Mixture model

- The feature vectors and class labels are assumed to be drawn from a joint probability distribution:

$$P(Y = k) = \pi_k, \quad k = 1, \ldots, K,$$

$$f(\mathbf{x}|Y = k) = f(x, \boldsymbol{\theta}_k), \quad k = 1, \ldots, K.$$

- Let $\psi = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$.

Data

- We consider a realization of an iid random sample from $(\mathbf{X}, Y)$ of size n:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n).$$

- The class labels are assumed to be imperfectly observed and partially specified by mass functions. The learning set has the following form:

$$\mathcal{L} = \{(\mathbf{x}_1, m_1), \ldots, (\mathbf{x}_n, m_n)\}.$$

- Problem 2: estimate ψ using $\mathcal{L}$.
- Remark: this problem encompasses supervised, unsupervised and semi-supervised learning as special cases.

Generalized likelihood criterion
(Côme et al., 2009)

Approach:

$$\hat{\psi} = \arg\max_{\psi} pl^{\boldsymbol{\Psi}}(\psi|\mathcal{L}).$$

Theorem

The logarithm of the conditional plausibility of ψ given $\mathcal{L}$ is given by

$$\ln\left(pl^{\boldsymbol{\Psi}}(\psi\mathcal{L})\right) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) + \nu,$$

where $pl_{ik} = pl(y_i = k)$ and ν is a constant.

Generalized EM algorithm
(Côme et al., 2009)

- An EM algorithm (with guaranteed convergence) can be derived to maximize the previous criterion.
- This algorithm becomes identical to the classical EM algorithm in the case of completely unsupervised or semi-supervised data.
- The complexity of this algorithm is identical to that of the classical EM algorithm.

Experimental settings

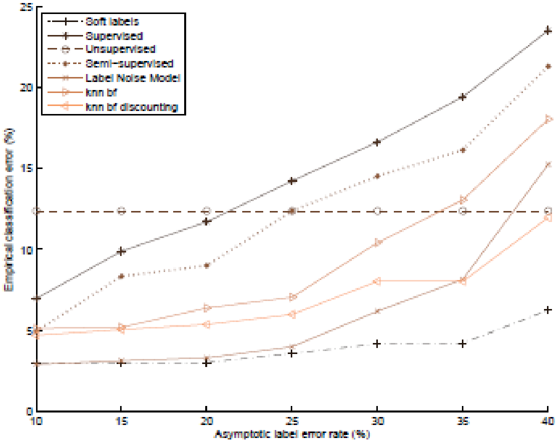
- Simulated and real data sets.
- Each example i was assumed to be labelled by an expert who provides his/her most likely label $\widehat{y}_i$ and a measure of doubt p_i.
- This information is represented by a simple mass function:

$$\begin{aligned} m_i(\{\widehat{y}_i\}) &= 1 - p_i \\ m_i(\Omega) &= p_i. \end{aligned}$$

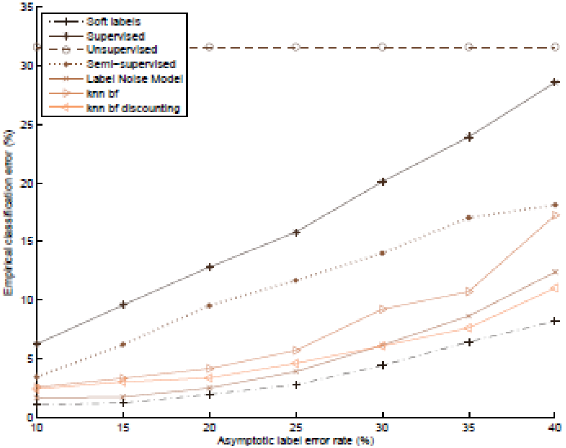
- Simulations: p_i drawn randomly form a Beta distribution, true label changed to any other label with probability p_i.

Results

Iris data



Results
Wine data



Outline

Credal partition

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

Example

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

Special cases

- Each m_i is a *certain bba* $\rightarrow$ crisp partition of Ω.
- Each m_i is a *Bayesian bba* $\rightarrow$ fuzzy partition of Ω

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

- Each m_i is a *consonant bba* $\rightarrow$ possibilistic partition of Ω

$$u_{ik} = pl_i^{\Omega}(\{\omega_k\})$$

Algorithms

- EVCLUS (Denoeux and Masson, 2004):
  - proximity (possibly non metric) data,
  - multidimensional scaling approach.
- Evidential c-means (ECM): (Masson and Denoeux, 2008):
  - attribute data,
  - alternate optimization of an FCM-like cost function.

Basic ideas

- Let $\mathbf{v}_k$ be the prototype associated to class ω_k $(k = 1, \ldots, K)$.
- Let A_j a non empty subset of Ω (a set of classes).
- We associate to A_j a prototype $\bar{\mathbf{v}}_j$ defined as the center of mass of the $\mathbf{v}_k$ for all $\omega_k \in A_j$.
- Basic ideas:
  - for each non empty $A_j \in \Omega$, $m_{ij} = m_i(A_j)$ should be high if $\mathbf{x}_i$ is close to $\bar{\mathbf{v}}_j$.
  - The distance to the empty set is defined as a fixed value δ.

Optimization problem

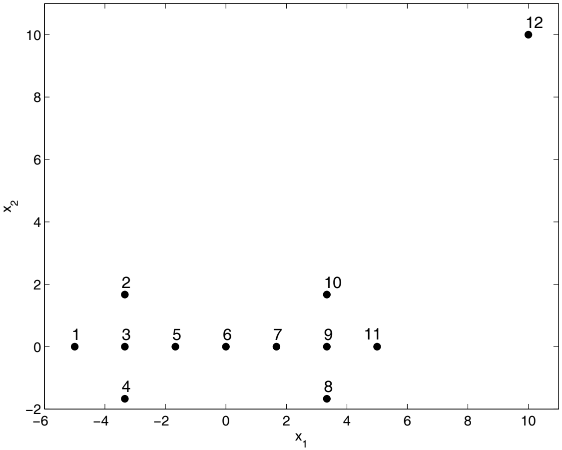
- Minimize

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

subject to

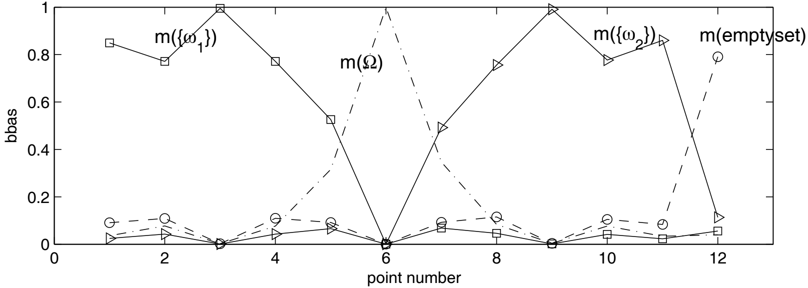
$$\sum_{\{j / A_j \subseteq \Omega, A_j \neq \emptyset\}} m_{ij} + m_{i\emptyset} = 1 \quad \forall i = 1, n,$$

Butterfly dataset

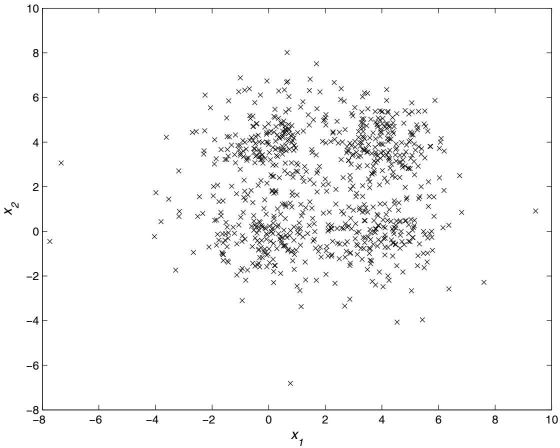


Butterfly dataset

Results

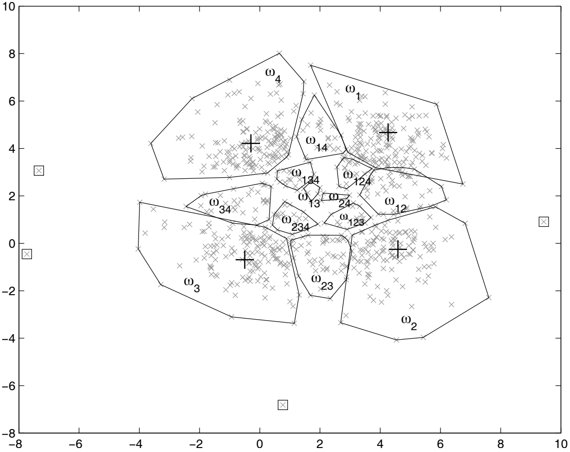


4-class data set



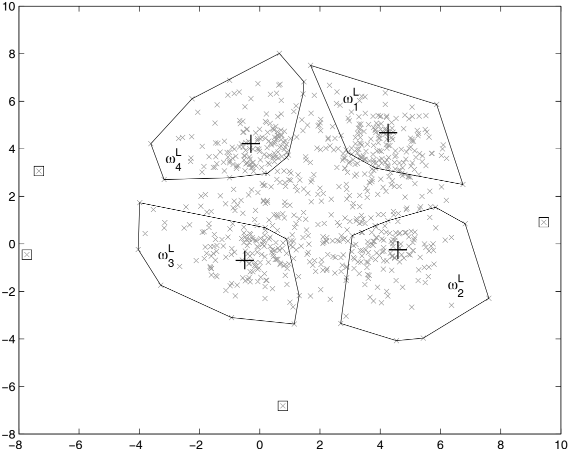
4-class data set

Hard credal partition



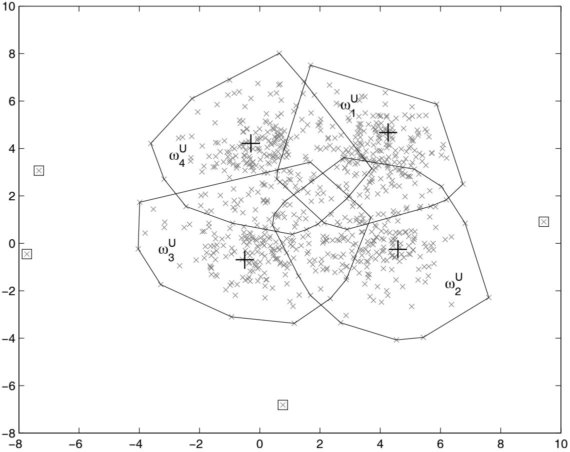
4-class data set

Lower approximation



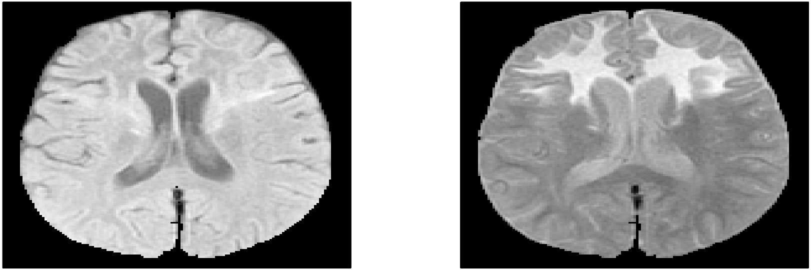
4-class data set

Upper approximation



Brain data

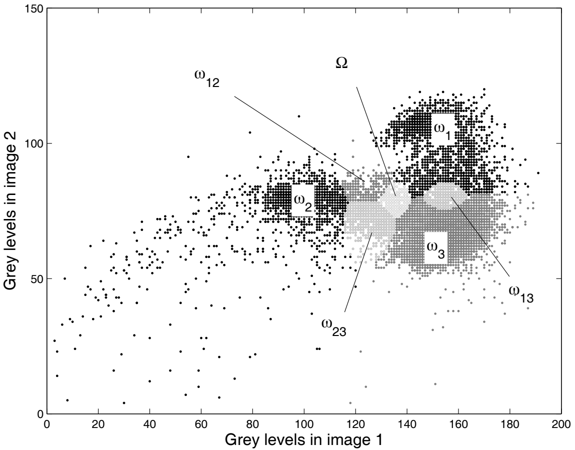
- Magnetic resonance imaging of pathological brain, 2 sets of parameters.
- Image 1 shows normal tissue (bright) and ventricals + cerebrospinal fluid (dark). Image 2 shows pathology (bright).



(a) (b)

Brain data

Results in gray level space



Brain data

Lower and upper approximations

(a) (b) (c)

Conclusion

- The theory of belief functions extends both set theory and probability theory:
  - It allows for the representation of imprecision and uncertainty.
  - It is more general than possibility theory.
- Belief functions may be used to represent imprecise and/or uncertain knowledge of class labels $\rightarrow$ soft labels.
- Many classification and clustering algorithms can be adapted to
  - handle such class labels (partially supervised learning)
  - generate them from data (credal partition)

References I

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

T. Denœux.

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

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

L. M. Zouhal and T. Denoeux.

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

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

References II

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

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

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

References III

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

T. Denœux and P. Smets.

Classification using Belief Functions: the Relationship between the Case-based and Model-based Approaches.

IEEE Transactions on Systems, Man and Cybernetics B, 36(6), 1395-1406, 2006.

M.-H. Masson and T. Denoeux.

ECM: An evidential version of the fuzzy c-means algorithm.

Pattern Recognition, 41(4), 1384-1397, 2008.

References IV

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

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.