Partially supervised learning by a credal EM approach

Patrick Vannoorenberghe[1] and Philippe Smets[2]

[1] PSI, FRE 2645 CNRS, Université de Rouen
Place Emile Blondel, 76821 Mont Saint Aignan cedex, France
Patrick.Vannoorenberghe@univ-rouen.fr
[2] IRIDIA, Université Libre de Bruxelles
50, av. Roosevelt, 1050 Bruxelles, Belgique
psmets@ulb.ac.be

Abstract. In this paper, we propose a credal EM (CEM) approach for partially supervised learning. The uncertainty is represented by belief functions as understood in the transferable belief model (TBM). This model relies on a non probabilistic formalism for representing and manipulating imprecise and uncertain information. We show how the EM algorithm can be applied within the TBM framework when applied for the classification of objects and when the learning set is imprecise (the actual class of each object is only known as belonging to a subset of classes), and/or uncertain (the knowledge about the actual class is represented by a probability function or by a belief function).
Keywords Learning, belief functions, EM, transferable belief model

1 Introduction

Supervised learning consists in assigning an input pattern $\boldsymbol{x}$ to a class, given a learning set $\mathcal{L}$ composed of N patterns $\boldsymbol{x}_i$ with known classification. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ be the set of K possible classes. Each pattern in $\mathcal{L}$ is represented by a p-dimensional feature vector $\boldsymbol{x}_i$ and its corresponding class label y_i. When the model generating the data is known, the classical methods of discriminant analysis (DA) permits the estimation of the parameters of the model.

Still these methods assumed in practice that the actual class y_i of each case in the learning set is well known. Instead suppose the data of the learning set are only partially observed, i.e., the actual class of a given object is only known to be one of those in a given subset C of Ω. Classical methods for parametric learning encounter then serious problems. One of the solution was based on the EM algorithm (Dempster, Laird, & Rubin, 1977; McLaclan & Krishnan, 1997).

Parametric learning requires a model of the generation of the data and an algorithm for estimating the parameters of this model using the available information contained in the learning set. A major drawback of many parametric

methods is their lack of flexibility when compared with nonparametric methods. However, this problem can be circumvented using mixture models which combine much of the flexibility of nonparametric methods with certain of the analytic advantages of parametric methods. In this approach, we assume that the data $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ are generated independently from a mixture density model which probability density function (pdf) is given by:

$$f(\boldsymbol{x}_i; y_i = \omega_k, \theta) = \sum_{g=1}^{G_k} \pi_{kg} f_{kg}(\boldsymbol{x}_i; \alpha_{kg}) \tag{1}$$

where G_k is the number of components in the mixture for the cases in class ω_k, π_{kg} are the mixing proportions, f_{kg} denotes a component, i.e. a probability distribution function parametrized by α_{kg}, and $\theta = \{(\pi_{kg}, \alpha_{kg}) : g = 1, \ldots, G_k; k = 1, \ldots, K\}$ are the model parameters to be estimated. For mixture of Gaussian pdfs, the function $f_{kg}(\boldsymbol{x}_i; \alpha_{kg})$ is a Gaussian pdf and α_{kg} is a set of parameters $\alpha_{kg} = (\boldsymbol{\mu}_{kg}, \boldsymbol{\Sigma}_{kg})$ where $\boldsymbol{\mu}_{kg}$ is the mean and $\boldsymbol{\Sigma}_{kg}$ the variance-covariance matrix of the Gaussian pdf f_{kg}.

Generally, the maximum likelihood estimation of the parameters of this model cannot be obtained analytically, but learning θ could be easily achieved if the particular component f_{kg} responsible for the existence of each observation $\boldsymbol{x}_i$ was known. In reality, this ideal situation is hardly encountered.

Several real world contexts can be described.

1. **The precise teacher case.** For each learning case, we know the actual class to which it belongs. The missing information is the g value for each case. The classical approach to solve this problem is the EM algorithm.
2. **The imprecise teacher case.** For each learning case, we only know that the actual class belongs to a subset of Ω. The missing information is the k and the g values for each case, where k is constrained to a subset of $1, \ldots, K$. The EM algorithm can be extended to such a case (Hastie & Tibshirani, 1996; Ambroise & Govaert, 2000).
3. **The precise and uncertain teacher case.** For each learning case, we only have some beliefs about what is the actual class to which the case belongs. The uncertainty is represented by a probability function on Ω. The uncertainty concerns the k value, and the g values are still completely unknown.
4. **The imprecise and uncertain teacher case.** For each learning case, we only have some beliefs about what is the actual class to which the case belongs. The uncertainty is represented by a belief function on Ω. The uncertainty and imprecision concern the k value, and the g values are still completely unknown. The EM algorithm can be further extended to such a case as done here.

In this paper, we consider the imprecise teacher case and the imprecise and uncertain teacher case, the first case being covered by the second one. Uncertainty is represented by belief functions as understood in the TBM (Smets &

Kennes, 1994; Smets, 1998). We propose to use the advantages of both the EM algorithm and the belief functions to learn the parameter of a TBM classifier. This algorithm is called the 'Credal EM' (CEM) and its related classifier is called the 'CEM classifier'.

Previous work on comparing a TBM classifier with an EM based classifier was performed in (Ambroise, Denoeux, Govaert, & Smets, 2001). Performance were analogous, but the TBM classifier was much simpler to use. The TBM classifier used in that comparison was based on non parametric methods as developed by (Denœux, 1995; Zouhal & Denœux, 1998). Here the TBM is used for parameter estimation and the final TBM classifier is based on a parametric method. This paper is organized as follows. The basic concepts of belief functions theory are briefly introduced in Section 2. The notion of likelihood is extended into the TBM in Section 3. The principle of parameters estimation via the EM algorithm is recalled in Section 4. The proposed algorithm is presented in Section 5. Finally, Section 6 gives some experimental results using synthetic data.

2 Background materials on belief functions

Let Ω be a finite space, and let 2^Ω be its power set. A belief function defined on Ω can be mathematically defined by introducing a set function, called the basic belief assignment (bba) $m^\Omega : 2^\Omega \rightarrow [0, 1]$ which satisfies:

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{2}$$

Each subset $A \subseteq \Omega$ such as $m^\Omega(A) > 0$ is called a focal element of m^Ω. Given this bba, a belief function bel^Ω and a plausibility function pl^Ω can be defined, respectively, as:

$$bel^\Omega(A) = \sum_{\emptyset \neq B \subseteq \Omega} m^\Omega(B), \ \ \forall\, A \subseteq \Omega. \tag{3}$$

$$pl^\Omega(A) = \sum_{A \cap B \neq \emptyset} m^\Omega(B), \ \ \forall\, A \subseteq \Omega. \tag{4}$$

The three functions bel^Ω, pl^Ω and m^Ω are in one-to-one correspondence and represent three facets of the same piece of information. We can retrieve each function from the others using the fast Möbius transform (Kennes, 1992). Let m_1^Ω and m_2^Ω be two bbas defined on the same frame Ω. Suppose that the two bbas are induced by two distinct pieces of evidence. Then the joint impact of the two pieces of evidence can be expressed by the conjunctive rule of combination which results in the bba:

$$m_{12}^\Omega(A) = (m_1^\Omega \textcircled{\cap} m_2^\Omega)(A) = \sum_{B \cap C = A} m_1^\Omega(B).m_2^\Omega(C). \tag{5}$$

In the TBM, we distinguish the credal level where beliefs are entertained (formalized, revised and combined) and the pignistic level used for decision making. Based on rationality arguments developed in the TBM, Smets proposes to transform m^{Ω} into a probability function $BetP$ on Ω (called the *pignistic* probability function) defined for all $\omega_k \in \Omega$ as:

$$BetP(\omega_k) = \sum_{A \ni \omega_k} \frac{m^{\Omega}(A)}{|A|} \frac{1}{1 - m^{\Omega}(\emptyset)} \tag{6}$$

where $|A|$ denotes the cardinality of $A \subseteq \Omega$ and $BetP(A) = \sum_{\omega \in A} BetP(\omega)$, $\forall A \subseteq \Omega$. In this transformation, the mass of belief $m(A)$ is distributed equally among the elements of A (Smets & Kennes, 1994; Smets, 2005).

Let us suppose the two finite spaces X, the observation space, and Θ, the unordered parameter space. The Generalized Bayesian Theorem (GBT), an extension of Bayes theorem within the TBM, consists in defining a belief function on Θ given an observation $x \subseteq X$, the set of conditional bbas $m^X[\theta_i]$ over X, one for each $\theta_i \in \Theta$[3] and a vacuous a priori on Θ. Given this set of bbas (which can be associated to their related belief or plausibility functions), then for $x \subseteq X$ and $\forall A \subseteq \Theta$, we have:

$$pl^{\Theta}[x](A) = 1 - \prod_{\theta_i \in A} (1 - pl^X[\theta_i](x)). \tag{7}$$

3 Explaining the likelihood maximization within the TBM

Suppose a random sample of a distribution with parameters $\theta \in \Theta$ and let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N : \boldsymbol{x}_i \in I\!R^p\}$ be the set of observations. In probability theory many estimation procedures for θ are based on the maximization of the likelihood, i.e. $P^{I\!R^p}(\boldsymbol{X}|\theta)$ considered as a function of θ. How do we generalize this procedure within the TBM? We reconsider the issue.

For each $\theta \in \Theta$, we have a conditional belief function on $I\!R$, denoted $m^{I\!R}[\theta]$. We observe $\boldsymbol{x} \subseteq I\!R$. This induce a bba on Θ by the application of the GBT. So we get the bba $m^{\Theta}[\boldsymbol{x}]$. How to estimate θ_0, the actual value of Θ? We could select the θ that maximizes $BetP^{\Theta}[\boldsymbol{x}]$, thus the most 'probable' value of Θ. This last solution means finding the modal value of $BetP^{\Theta}[\boldsymbol{x}]$. We feel this principle fits with the idea underlying the maximum likelihood estimators.

So we must find the $\theta \in \Theta$ such that $BetP^{\Theta}[\boldsymbol{x}](\theta) \geq BetP^{\Theta}[\boldsymbol{x}](\theta_i), \forall\ \theta_i \in \Theta$. This maximization seems hard to solve, but we can use theorem III.1. in

[3] We use the next notational convention for the indices and [ ]: $m^D[u](A)$ denotes the mass given to the subset A of the domain D by the conditional bba $m^D[u]$ defined on D given u is accepted as true.

(Delmotte & Smets, 2004) which states that the θ that maximizes $BetP^{\Theta}[\boldsymbol{x}]$ is the same as the one that maximizes the plausibility function $pl^{\Theta}[\boldsymbol{x}](\theta)$, provided the *a priori* belief on Θ is vacuous, as it is the case here.

Theorem 1. *Given $\boldsymbol{x} \subseteq X$ and $pl^{X}[\theta](\boldsymbol{x})$ for all $\theta \in \Theta$, let $pl^{\Theta}[\boldsymbol{x}]$ be the plausibility function defined on Θ and computed by the GBT, and $BetP^{\Theta}[\boldsymbol{x}]$ be the pignistic probability function constructed on Θ from $pl^{\Theta}[\boldsymbol{x}]$, then:*

$$BetP^{\Theta}[\boldsymbol{x}](\theta_i) > BetP^{\Theta}[\boldsymbol{x}](\theta_j) \qquad \textit{iff} \qquad pl^{\Theta}[\theta_i](\boldsymbol{x}) > pl^{\Theta}[\theta_j](\boldsymbol{x}). \tag{8}$$

In the TBM, $pl^{\Theta}[\boldsymbol{x}](\theta)$ is equal to $pl^{X}[\theta](\boldsymbol{x})$. Furthermore when N i.i.d. data $\boldsymbol{x}_i, i = 1, \ldots, N$, are observed, we get $pl^{X^N}[\theta](\boldsymbol{x}_1, ..., \boldsymbol{x}_N) = \prod_{i=1}^{N} pl^{X}[\theta](\boldsymbol{x}_i)$. This last term is easy to compute and leads thus to applicable algorithms. Maximizing the likelihood over θ turns out to mean maximizing over θ the conditional plausibilities of the data given θ.

4 Parameter estimation by EM algorithm

We introduce the classical EM approach to find the parameters of a mixture models from a data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ made of cases which belong to a same class. The aim is to estimate the posterior distribution of the variable $\boldsymbol{y}$ which indicates the component of the mixture that generated $\boldsymbol{x}_i$ taking into account the available information $\mathcal{L}$. For simplicity sake, we do not indicate the class index k. For that estimation, we need to know π_g, f_g and α_g for $g = 1, \ldots, G$. For their estimation, we use the EM algorithm to maximize according to θ the log likelihood:

$$L(\theta; \boldsymbol{X}) = \log(\prod_{i=1}^{N} f(\boldsymbol{x}_i; \theta)) = \sum_{i=1}^{N} \log(\sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}_i; \alpha_g)). \tag{9}$$

In order to solve this problem, the idea is that if one had access to a hidden random variable $\boldsymbol{z}$ that indicates which data point was generated by which component, then the maximization problem would decouple into a set of simple maximizations. Using this indicator variable $\boldsymbol{z}$, relation (9) can be written as the next complete-data log likelihood function:

$$L_c(\theta; \boldsymbol{X}, \boldsymbol{z}) = \sum_{i=1}^{N} \sum_{g=1}^{G} z_{ig} \log(\pi_g f_g(\boldsymbol{x}_i; \alpha_g)) \tag{10}$$

where $z_{ig} = 1$ if the Gaussian pdf having generated the observation $\boldsymbol{x}_i$ is f_g, and 0 otherwise. Since $\boldsymbol{z}$ is unknown, L_c cannot be used directly, so we usually work with its expectation denoted $Q(\theta|\theta^l)$ where l is used as the iteration index. As shown in (Dempster et al., 1977), $L(\theta; \boldsymbol{X})$ can be maximized by iterating the following two steps:

– E step: $Q(\theta|\theta^l) = E[L_c(\theta; \boldsymbol{X}, \boldsymbol{z})|\boldsymbol{X}, \theta^l]$
– M step: $\theta^{l+1} = \arg\max_\theta Q(\theta|\theta^l)$

The E (Expectation) step computes the expected complete data log likelihood and the M (Maximization) step finds the parameters that maximize that likelihood. $Q(\theta|\theta^l)$ can be rewritten as

$$Q(\theta|\theta^l) = \sum_{i=1}^{N}\sum_{g=1}^{G} E[z_{ig}|\boldsymbol{X}, \theta^l] \log(\pi_g f_g(\boldsymbol{x}_i; \alpha_g)) \tag{11}$$

In a probabilistic framework, $E[z_{ig}|\boldsymbol{X}, \theta^l]$ is nothing more than $P(z_{ig} = 1|\boldsymbol{X}, \theta^l)$, the posterior distribution easily computed from the observed data.

5 CEM : the credal solution

In this section, we introduce a credal EM approach for partially supervised learning. The imprecision or/and uncertainty on the observed labels are represented by belief functions (cf. section 5.1). We consider the imprecise and uncertain teacher case (section 5.2).

5.1 Partially observed labels

Thanks to its flexibility, a belief function can represent different forms of labels including hard labels (HL), imprecise labels (IL), probabilistic labels (PrL), possibilistic (PoL) labels and credal labels (CrL). Table 1 illustrates an example of the bbas that characterize the knowledge about the labels on a three-class frame. Note that a possibility measure is known to be formally equivalent to a consonant belief function, i.e., a belief function with nested focal elements (Denœux & Zouhal, 2001). Unlabeled samples (UL) can be encoded using the vacuous belief

$A \subseteq \Omega$	HL	IL	PrL	PoL	CrL	UL
$\{\omega_1\}$	0	0	0.2	0	.1	0
$\{\omega_2\}$	1	0	0.6	0	0	0
$\{\omega_1, \omega_2\}$	0	1	0	0	.2	0
$\{\omega_3\}$	0	0	0.2	0.7	.3	0
$\{\omega_1, \omega_3\}$	0	0	0	0.2	.3	0
$\{\omega_2, \omega_3\}$	0	0	0	0	0	0
Ω	0	0	0	0.1	.1	1

Table 1. Example of imprecise and uncertain labeling with belief functions

function m_v defined as $m_v(\Omega) = 1$. This show that handling the general case based on belief functions covers all cases of imperfect teacher (imprecise and/or uncertain). Of course, the TBM covers the HL, IL, PrL and CrL cases. For the PoL, the CEM algorithm presented here has to be adapted as we use the GBT and other combination rules that differ from their possibilistic counterparts.

5.2 The imprecise and uncertain teacher case

Let $\Omega = \{\omega_1, \ldots, \omega_K\}$ be a set of K mutually exclusive classes[4]. Let $\mathcal{L}$ be a set of N observed cases and called the learning set. For $i = 1, \ldots, N$, let c_i denotes the i-th case. For case c_i, we collect a feature vector $\boldsymbol{x}_i$ taking values in $\mathbb{R}^p$, and a bba m_i^Ω that represents all we know about the actual class $y_i \in \Omega$ to which case c_i belongs. We then assume that the probability density function (pdf) of $\boldsymbol{x}_i$ is given by the next mixture of pdfs :

$$f(\boldsymbol{x}_i; y_i = \omega_k, \theta_k) = \sum_{g=1}^{G_k} \pi_{kg} f_{kg}(\boldsymbol{x}_i; \alpha_{kg}) \tag{12}$$

where f_{kg} is the p-dimensional Gaussian pdf with parameters $\alpha_{kg} = (\boldsymbol{\mu}_{kg}, \boldsymbol{\Sigma}_{kg})$.

Let the available data be $\{(\boldsymbol{x}_1, m_1^\Omega)..., (\boldsymbol{x}_N, m_N^\Omega)\}$ where $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ is an i.i.d sample. Let $\boldsymbol{Y} = (y_1, ..., y_N)$ be the unobserved labels and $\boldsymbol{m}^\Omega = (m_1^\Omega, \ldots, m_N^\Omega)$ are the bbas representing our beliefs about the actual values of the y_i's. For the estimation of the parameters $\theta = (\{\alpha_{kg} : j = 1, \ldots G_k, k = 1, \ldots, K\}, \boldsymbol{Y})$, we use the EM algorithm to maximize the log likelihood given by:

$$L(\theta; \mathcal{L}) = \log(\prod_{i=1}^{N} f(\boldsymbol{x}_i; y_i = \omega_k, \theta_k)) = \sum_{i=1}^{N} \log(\sum_{g=1}^{G_k} \pi_{kg} f_{kg}(\boldsymbol{x}_i; \alpha_{kg})). \tag{13}$$

We can rephrase the relation by considering all the Gaussian pdfs. There are $G = \sum_{k=1}^{K} G_k$ Gaussian pdfs. Let J_k be the indexes in the new ordering of the components of the class ω_k. So $J_k = \{j : \sum_{\nu=1}^{k-1} G_\nu < j \leq \sum_{\nu=1}^{k} G_\nu\}$ where $\sum_{\nu=1}^{0} G_\nu = 0$. This reindexing is analogous to a refinement $\mathcal{R}$ of the classes in $\Omega = \{\omega_k : k = 1, \ldots, K\}$ into a set of new 'classes' $\Omega^* = \{\omega_j^* : j = 1, \ldots, G\}$ where ω_k is mapped onto $\{\omega_j^* : j \in J_k\}$. The bba m_i^Ω can be refined on Ω^* as $m_i^{\Omega^*}$ where

$$\begin{aligned} m_i^{\Omega^*}(\mathcal{R}(A)) &= m_i^\Omega(\mathcal{A}) \qquad \forall A \subseteq \Omega \\ &= 0 \qquad\qquad \text{otherwise} \end{aligned} \tag{14}$$

[4] In the TBM, we do not require Ω to be exhaustive, but one could add this requirement innocuously.

For each case c_i, we must find out which of the G pdfs generated their $\boldsymbol{x}_i$ data. So, equation (13) can be written as:

$$L(\theta;\mathcal{L}) = \sum_{i=1}^{N} \log(\sum_{j=1}^{G} \pi_j f_j(\boldsymbol{x}_i;\alpha_j)) \tag{15}$$

where the sum of the π_j taken on the j indexes corresponding to the possible classes of c_i must add to 1, all others being 0.

We reconsider the EM algorithm when the teacher is imperfect. We need for each case c_i the plausibility of $\boldsymbol{x}_i$ given the bba $m_i^{\Omega^*}$ about its class in Ω^*. If the actual class is ω_j^*, then $pl^{\mathbb{R}^p}[\omega_j^*](\boldsymbol{x}_i)$ is given by $f_j(\boldsymbol{x}_i,\alpha_j)$. If x_i is a singleton (as usual and assumed hereafter) then $pl^{\mathbb{R}^p}[\omega_j^*](\boldsymbol{x}_i) = f_j(\boldsymbol{x}_i,\alpha_j)d\boldsymbol{x}$ where we put $d\boldsymbol{x}$ to mention that a plausibility is a set function whereas f itself is a density. This $d\boldsymbol{x}$ term will cancel when normalizing. Let $A \subseteq \Omega^*$, then from the disjunctive rule of combination associated to the GBT we get:

$$pl^{\mathbb{R}^p}[A](\boldsymbol{x}_i) = 1 - \prod_{j:\omega_j^* \in A} (1 - pl^{\mathbb{R}^p}[\omega_j^*](\boldsymbol{x}_i)). \tag{16}$$

We then assess the bba on Ω^* given θ^l and $\boldsymbol{x}_i$. From the GBT, we get $m^{\Omega^*}[\boldsymbol{x}_i,\theta^l]$. We combine this bba with the prior bba given by $m_i^{\Omega^*}$ by the conjunctive combination rule. The term to maximize is then:

$$Q(\theta|\theta^l) = \sum_{i=1}^{N} \sum_{A\subseteq\Omega^*} (m^{\Omega^*}[x_i,\theta^l] \textcircled{\cap} m_i^{\Omega^*})(A) \log(pl^{\mathbb{R}^p}[A](\boldsymbol{x}_i)) \tag{17}$$

where $pl^{\mathbb{R}^p}[A](\boldsymbol{x}_i)$ is given by relation (16).

6 Simulations results

In this section, we propose to illustrate the performance of the CEM algorithm described in the previous sections using two learning tasks.

6.1 Learning task 1: Isosceles triangles

In this task, we have three classes: $\Omega = \{\omega_1,\omega_2,\omega_3\}$ and two-dimensional data. In each class, there are 2 components ($G_k = 2, k = 1,2,3$). For a given subset, each vector $\boldsymbol{x}$ is generated from a Gaussian $f(\boldsymbol{x}|\omega_g) \sim \mathcal{N}(\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$ where $\boldsymbol{\Sigma}_g = \sigma\boldsymbol{I}$. The parameters for the 6 pdfs are presented in table 2. The pdf corresponds to 3 largely spread data ($\sigma = 2$) located at the 3 corners of an isosceles triangle,

and to 3 clustered data ($\sigma = 0.5$) located at the 3 corners of another isosceles triangle. The pair of pdf corresponding to one class are thus located at one corner and half way on the line between the other 2 corners. In figure 1, we illustrate an example of such a learning set with its respective isosceles triangles (fine lines).

	$\omega_1(+)$ subset1	$\omega_1(+)$ subset2	$\omega_2(\times)$ subset1	$\omega_2(\times)$ subset2	$\omega_3(\cdot)$ subset1	$\omega_3(\cdot)$ subset2
μ_a	10	17.5	15	15	20	12.5
μ_b	10	14.3	18.6	10	10	14.3
σ	2	0.5	2	0.5	2	0.5
IL cases	50 ω_1	25 ω_1, ω_2 25 ω_1, ω_3	50 ω_2	25 ω_1, ω_2 25 ω_2, ω_3	50 ω_3	25 ω_1, ω_3 25 ω_2, ω_3
m_a	9.13	17.54	15.60	14.92	20.36	12.42
m_b	10.35	14.32	18.95	10.12	9.86	14.35
s	2.57	0.38	1.85	0.37	3.24	0.35
r	0.185	0.152	0.178	0.148	0.179	0.154

Table 2. Parameters of the learning set for task 1 with imprecise labels (IL) and the estimations obtained with the CEM for one run.

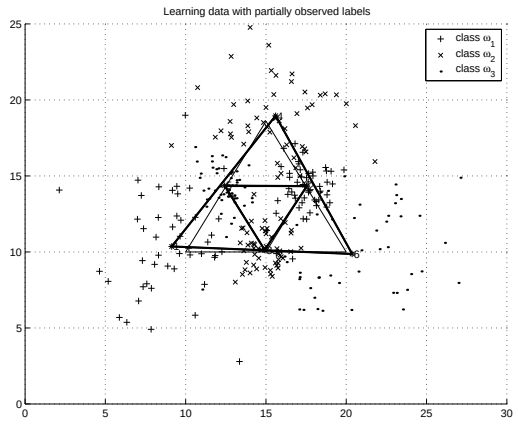

Fig. 1. Learning set in the feature space

We generate a sample of 50 cases from each of the 6 pdfs. Labels for each case can be of two types, either imprecise (IL) or credal (CrL). In the IL case, the labels for the 50 cases from the largely spread data (those at the corners) are precise. The other 50 cases are randomly split into two groups of 25 cases. Their labels are imprecise and made of 2 classes, the actual class being one of them. So for the 50 cases in subset 2 of class ω_1, 25 are labeled $\{\omega_1, \omega_2\}$ and 25 are labeled $\{\omega_1, \omega_3\}$. In the CrL case, the labels are subsets of Ω randomly

generated and each one receives a random mass. We thus generate imprecise and uncertain learning sets as they can be encountered in real world applications.

We run 10 simulations. For each of them, we generate the labels for the IL and CrL cases. In figure 1, we present the data for one simulation. The bold line triangle illustrates the result of the application of the CEM for the IL case. As can be seen, the means (the corners of the triangles) are well located. The estimated parameters are listed at the bottom of table 2. On the IL data, we apply both a classical EM algorithm and the CEM. On the CrL, we apply only the CEM algorithm as the classical does not seem fitted for such type of data. In table 3, we present the Percentage of Correct Classification (PCC) obtained for each of the 10 independent training sets. Each method produces very similar

Triangles	1	2	3	4	5	6	7	8	9	10	mean	std
EM	85.3	84.3	86.3	88.0	86.7	87.0	83.3	85.7	90.7	88.0	86.5	2.1
CEM IL	86.3	85.3	88.0	90.3	88.0	87.3	84.0	88.0	91.0	88.0	87.6	2.0
CEM CrL	87.0	86.6	87.6	90.0	87.6	88.0	85.3	88.3	91.3	86.7	87.8	1.7

Table 3. Percentage of correct classification for classical EM and CEM algorithms.

results but only the CEM algorithm is able to use credal labels, a much more flexible information than the one encountered in the IL case.

6.2 Learning task 2: Qualitative example

This learning set is drawn using three bi-dimensional Gaussian classes of standard deviation 1.5 respectively centered on $(3, 0)$, $(0, 5)$ and $(0, 0)$. Figure 2 illustrates this learning task associated to the decision regions computed using parameters of the CEM algorithm learnt from credal labels (CrL). A very important, but classical feature using EM and mixture models algorithms, is the ability to cope with unlabeled samples. The first intuition is that these unlabeled data don't bring any information for learning the parameters of the generated data. Contrary to this idea, we can show on this illustrative example that unlabeled data give clearly a more precise idea of the real distributions. To highlight this issue, two training sets were considered: a training set (set 1) which contains all the data except that we randomly remove 40 cases (80%) of class ω_2, and a training set (set 2) with all the data (150 cases). In this second learning set, we replace the credal labels generated for the 40 previous cases with vacuous belief functions (UL) before applying the CEM classifier. Table 4 shows the estimated parameters for these two learning tasks. Additionally, estimated means are illustrated with gray levels disks in figure 2. This last capacity makes CEM a very suitable algorithm for cluster analysis which is under study. In all these simulations, the estimation of the number of components G_k is a difficult model

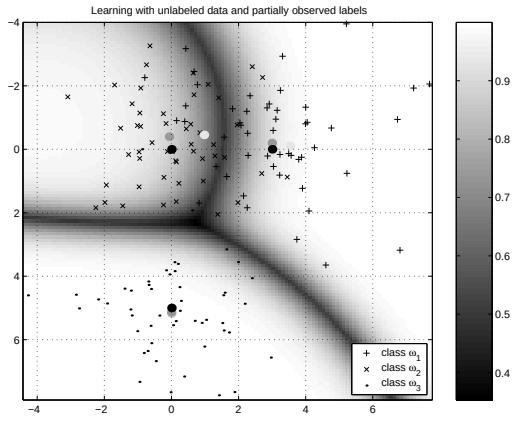

Fig. 2. Maximum pignistic probabilities as grey level values

	$\omega_1(+)$ μ_a	$\omega_1(+)$ μ_b	$\omega_2(\times)$ μ_a	$\omega_2(\times)$ μ_b	$\omega_3(\cdot)$ μ_a	$\omega_3(\cdot)$ μ_b
Real values	3.00	0.00	0.00	0.00	0.00	5.00
Training set 1	3.52	-0.10	0.96	-0.45	-0.00	5.18
Training set 2	2.99	-0.19	-0.07	-0.40	-0.00	5.14

Table 4. Estimated parameters of the learning task 2.

choice problem for which there is a number of possible solutions (Figueiredo & Jain, 2002). This problem is left for future works.

7 Conclusion

In this paper, a credal approach for partially supervised learning has been presented. The proposed methodology uses a variant of EM algorithm to estimate parameters of mixture models and can cope with learning set where the knowledge about the actual class is represented by a belief function. Several simulations have proved the good performance of this CEM algorithm compared to classical EM estimation in learning mixture of Gaussians.

Numerous applications of this approach can be mentioned. As example, let us consider Bayesian networks which use EM algorithms to estimate parameters of unknown distributions. Using CEM algorithm can be a good alternative for belief networks. Future work is concerned with model selection issue which includes the choice of the number of components, shape of each component... Another important issue is the detection of outliers which can be solved by adding an extra component (uniform for example) in the mixture.

References

Ambroise, C., Denoeux, T., Govaert, G., & Smets, P. (2001). Learning from an imprecise teacher: probabilistic and evidential approaches. In *Proceedings of asmda'2001* (Vol. 1, pp. 100–105). Compiègne, France.

Ambroise, C., & Govaert, G. (2000). EM algorithm for partially known labels. In *Proceeding of IFCS'2000* (Vol. 1). Namur, Belgium.

Delmotte, F., & Smets, P. (2004). Target identification based on the transferable belief model interpretation of Dempster-Shafer model. *IEEE Transactions on Systems, Man and Cybernetics*, *A 34*, 457–471.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, series B*, *39*, 1-38.

Denœux, T. (1995). A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, *25*(5), 804–813.

Denœux, T., & Zouhal, L. M. (2001). Handling possibilistic labels in pattern classification using evidential reasoning. *Fuzzy Sets and Systems*, *122*, 47–62.

Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, *24*(3), 381–396.

Hastie, T., & Tibshirani, R. J. (1996). Discriminant analysis by gaussian mixtures. *Journal of the Royal Statistical Society B*, *58*, 155–176.

Kennes, R. (1992). Computational aspects of the Möbius transform of a graph. *IEEE-SMC*, *22*, 201–223.

McLaclan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions.* New York: John Wiley.

Smets, P. (1998). The transferable belief model for quantified belief representation. In D. M. Gabbay & P. Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems* (Vol. 1, pp. 267–301). Kluwer, Doordrecht, The Netherlands.

Smets, P. (2005). Decision making in the TBM: the necessity of the pignistic transformation. *Int. J. Approximate Reasoning*, *38*, 133–147.

Smets, P., & Kennes, R. (1994). The transferable belief model. *Artificial Intelligence*, *66*, 191–234.

Zouhal, L. M., & Denœux, T. (1998). An evidence theoretic k-nn rule with parameter optimisation. *IEEE Transactions on Systems, Man and Cybernetics - Part C*, *28*, 263-271.