HAL open science

Mixture model estimation with soft labels

Etienne Côme, Latifa Oukhellou, Thierry Denoeux, Patrice Aknin

To cite this version:

Etienne Côme, Latifa Oukhellou, Thierry Denoeux, Patrice Aknin. Mixture model estimation with soft labels. Fourth International Workshop on Soft Methods in Probabilities and Statistics, Sep 2008, Toulouse, France. pp.165-174. hal-00446806

HAL Id: hal-00446806

https://hal.science/hal-00446806v1

Submitted on 13 Jan 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Author manuscript, published in "Fourth International Workshop on Soft Methods in Ptobabilities and Statistics, Toulouse : France (2008)"

hal-00446806, version 1 - 13 Jan 2010

Chapter 1
Mixture model estimation with soft labels

E. Côme, L. Oukhellou, T. Denœux, and P. Aknin

Abstract This paper addresses classification problems in which the class membership of training data is only partially known. Each learning sample is assumed to consist in a feature vector and an imprecise and/or uncertain "soft" label m_i defined as a Dempster-Shafer basic belief assignment over the set of classes. This framework thus generalizes many kinds of learning problems including supervised, unsupervised and semi-supervised learning. Here, it is assumed that the feature vectors are generated from a mixture model. Using the General Bayesian Theorem, we derive a criterion generalizing the likelihood function. A variant of the EM algorithm dedicated to the optimization of this criterion is proposed, allowing us to compute estimates of model parameters. Experimental results demonstrate the ability of this approach to exploit partial information about class labels.

Key words: Dempster-Shafer theory, Transferable Belief Model, Mixture models, EM algorithm, Classification, Clustering, Partially supervised learning, Semi-supervised learning.

1.1 Introduction

Machine learning classically deals with two different problems: supervised learning (classification) and unsupervised learning (clustering). However, other paradigms exist such as *semi-supervised learning* [10], and *partially-supervised learning* [5, 1, 9, 11]. In the former approach, one use a mix of unlabelled and labelled examples, whereas in the latter, one can define constraints on the possible classes of the examples. The importance for such problems comes from the fact that labelled data are often difficult to obtain, while unlabelled or partially labelled data are easily available.

The investigations reported in this paper follow this path, in the context of belief functions. In this way, both the uncertainty and the imprecision of class labels may be handled. The considered training sets are of the form $\mathbf{X}^{iu} = \{(\mathbf{x}_1, m_1), \ldots, (\mathbf{x}_N, m_N)\}$, where m_i is a basic belief assignment, or Dempster-Shafer mass function [14] encoding our knowledge about the class of example i. The m_is (hereafter referred to as "soft labels") may represent different kinds of knowledge, from precise to imprecise and from certain to uncertain. Thus, previous problems are special cases of this general formulation. Other studies have already proposed solutions in which class labels are expressed by

E. Côme
French National Institute for Transport and Safety Research (INRETS) - LTN 2 av. Malleret-Joinville, 94114 Arcueil Cedex, France, Tel.: +33 1 47 40 73 49, Fax: +33 1 45 47 56 06, e-mail: come@inrets.fr

L. Oukhellou
Université Paris XII - CERTES, 61 av. du Général de Gaulle, 94100 Créteil, France, e-mail: oukhellou@univ-paris12.fr

T. Denœux
HEUDIASYC, Université de Technologie de Compiègne, CNRS - Centre de Recherches de Royallieu, B.P. 20529, 60205 Compiègne Cedex, France, e-mail: ijar@hds.utc.fr

P. Aknin
French National Institute for Transport and Safety Research (INRETS) - LTN, e-mail: aknin@inrets.fr

possibility distributions or belief functions [6, 8]. In this article, we present a new approach to solve learning problems of this type, which completes a preliminary study by Vannoorenberghe and Smets [21]. This solution is based on mixture models, and therefore assumes a generative model for the data.

This article is organized as follows. Background material on belief functions and estimation of parameters in mixture models using the EM algorithm will first be recalled in Sections 1.2 and 1.3, respectively. The problem of learning from data with soft labels will then be addressed in Section 1.4, through the definition of a learning criterion, and of an EM type algorithm dedicated to its optimization. Finally we will presented some simulations results in Section 1.5.

1.2 Background on Belief Functions

1.2.1 Belief Functions on a Finite Frame

The theory of belief functions was introduced by Dempster [3] and Shafer [14]. The interpretation adopted throughout this paper will be that of the Transferable Belief Model (TBM) introduced by Smets [20]. The first building block of belief function theory is the *basic belief assignment* (bba), which models the beliefs held by an agent regarding the actual value of a given variable taking values in a finite domain (or *frame of discernment*) Ω, based on some body of evidence. A bba m^Ω is a mapping from 2^Ω to $[0,1]$ verifying $\sum_{\omega\subseteq\Omega} m^\Omega(\omega) = 1$. The subsets ω for which $m^\Omega(\omega) > 0$ are called the *focal sets*. Several kind of belief functions are defined according to the structure of focal sets. In particular, a bba is *Bayesian* if its focal sets are singletons, it is *consonant* if its focal sets are nested and a it is *categorical* if it has only one focal set. Bbas are in one to one correspondence with other representations of the agent's belief, including the plausibility function defined as:

$$pl^\Omega(\omega) \triangleq \sum_{\alpha\cap\omega\neq\emptyset} m^\Omega(\alpha), \quad \forall\omega\subseteq\Omega. \tag{1.1}$$

The quantity $pl^\Omega(\omega)$ is thus equal to the sum of the basic belief masses assigned to propositions that are not in contradiction with ω. The plausibility function associated to a Bayesian bba is a probability measure. If m^Ω is consonant, then pl^Ω is a possibility measure: it verifies $pl^\Omega(\alpha\cup\beta) = \max(pl^\Omega(\alpha), pl^\Omega(\beta))$, for all $\alpha,\beta\subseteq\Omega$.

1.2.2 Conditioning and Combination

Given two bbas m_1^Ω and m_2^Ω supported by two distinct bodies of evidence, we may build a new bba $m_{1\textcircled{\cap}2}^\Omega = m_1^\Omega \textcircled{\cap} m_2^\Omega$ that corresponds to the conjunction of these two bodies of evidence:

$$m_{1\textcircled{\cap}2}^\Omega(\omega) \triangleq \sum_{\alpha_1\cap\alpha_2=\omega} m_1^\Omega(\alpha_1) m_2^\Omega(\alpha_2), \quad \forall\omega\subseteq\Omega. \tag{1.2}$$

This operation is usually referred to as the *unnormalized Dempster's rule* or the TBM conjunctive rule. If the frame of discernment is supposed to be exhaustive, the mass of the empty set is usually reallocated to other subsets, leading to the definition of the normalized Demspter's rule $\oplus$ defined as:

$$m_{1\oplus2}^\Omega(\omega) = \begin{cases} 0 & \text{if } \omega = \emptyset \\ \frac{m_{1\textcircled{\cap}2}^\Omega(\omega)}{1-m_{1\textcircled{\cap}2}^\Omega(\emptyset)} & \text{if } \omega\subseteq\Omega, \omega\neq\emptyset, \end{cases} \tag{1.3}$$

which is well defined provided $m_{1\textcircled{\cap}2}^\Omega(\emptyset) \neq 1$. Note that, if m_1^Ω (or m_2^Ω) is Bayesian, then $m_{1\oplus2}^\Omega(\omega)$ is also Bayesian. The combination of a bba m^Ω with a categorical bba focused on $\alpha\subseteq\Omega$ using the TBM conjunctive rule is called (unnormalized) *conditioning*. The resulting

hal-00446806, version 1 - 13 Jan 2010

bba is denoted $m^{\Omega}(\omega|\alpha)$. Probabilistic conditioning is recovered when m^{Ω} is Bayesian, and normalization is performed. Using this definition, we may rewrite the conjunctive combination rule: $m^{\Omega}_{1\bigcap 2}(\omega) = \sum_{\alpha\subseteq\Omega} m^{\Omega}_1(\alpha)m^{\Omega}_2(\omega|\alpha), \forall\omega \subseteq \Omega$, which is a counterpart of the total probability theorem in probability theory [7, 17]. This expression provides a shortcut to perform marginal calculations on a product space when conditional bbas are available [17]. Consider two frames Ω and Θ, and a set of conditional belief functions $m^{\Theta|\Omega}(\cdot|\omega)$ for all $\omega \subseteq \Omega$. Each conditional bba $m^{\Theta|\Omega}(\cdot|\omega)$ represents the agent's belief on Θ in a context where ω holds. The combination of these conditional bbas with a bba m^{Ω} on Ω yields the following plausibility on Θ:

$$pl^{\Theta}(\theta) = \sum_{\omega\subseteq\Omega} m^{\Omega}(\omega)pl^{\Theta|\Omega}(\theta|\omega), \quad \forall\theta \subseteq \Theta. \tag{1.4}$$

This property bears some resemblance with the total probability theorem, except that the sum is taken over the power set of Ω and not over Ω. We will name it the *total plausibility theorem.*

1.2.3 Independence, Continuous Belief functions and Bayes Theorem

The usual independence concept of probability theory does not easily find a counterpart in belief function theory, where different notions must be used instead. The simplest form of independence defined in the context of belief functions is *cognitive independence* [14, p. 149]. Frames Ω and Θ are said to be cognitively independent with respect to $pl^{\Omega\times\Theta}$ iff we have $pl^{\Omega\times\Theta}(\omega \times \theta) = pl^{\Omega}(\omega)\, pl^{\Theta}(\theta), \forall\omega \subseteq \Omega, \forall\theta \subseteq \Theta$. Cognitive independence boils down to probabilistic independence when $pl^{\Omega\times\Theta}$ is a probability measure.

The TBM can be extended to continuous belief functions on the real line, assuming focal sets to be real intervals [19]. In this context, the concept of bba is replaced by that of *basic belief density* (bbd), defined as a mapping $m^{\mathbb{R}}$ from the set of closed real intervals to $[0, +\infty)$ such that $\int_{-\infty}^{+\infty}\int_{x}^{+\infty} m^{\mathbb{R}}([x, y])dydx \leq 1$. By convention, the one's complement of this integral is allocated to $\emptyset$. As in the discrete case, $pl^{\mathbb{R}}([a, b])$ is defined as an integral over all intervals whose intersection with $[a, b]$ is non-empty. Further extension of these definitions to $\mathbb{R}^d, d > 1$ is possible and it is also possible to define belief functions on mixed product spaces involving discrete and continuous frames.

The Bayes' theorem of probability theory is replaced in the framework of belief functions by the Generalized Bayesian Theorem (GBT), [18]. This theorem provides a way to reverse conditional belief functions without any prior knowledge. Let us suppose two spaces, $\mathcal{X}$ the observation space and Θ the parameter space. Assume that our knowledge is encoded by a set of conditional bbas $m^{\mathcal{X}|\Theta}(.|\theta_i), \theta_i \in \Theta$, which express our belief in future observations conditionally on each θ_i, and we observe a realization $x \subseteq \mathcal{X}$. The question is: given this observation and the set of conditional bbas, what is our belief on the value of Θ? The answer is given by the GBT and states that the resulting plausibility function on Θ has the following form:

$$pl^{\Theta|\mathcal{X}}(\theta|x) = pl^{\mathcal{X}|\Theta}(x|\theta) = 1 - \prod_{\theta_i\in\theta}(1 - pl^{\mathcal{X}|\Theta}(x|\theta_i)). \tag{1.5}$$

When a prior bba m^{Θ}_0 on Θ is available, it should be combined conjunctively with the bba defined by (1.5). The classical Bayes' theorem is recovered when the conditional bbas $m^{\mathcal{X}|\Theta}(.|\theta_i)$ and the prior bba m^{Θ}_0 are Bayesian.

1.3 Mixture Models and the EM Algorithm

After this review of some tools from belief functions theory, the next part is dedicated to the probabilistic formulation of the clustering problem in terms of mixture model. We will therefore

hal-00446806, version 1 - 13 Jan 2010

present the data generation scheme underlying mixture models and the solution to parameter estimation in the unsupervised case.

1.3.1 Mixture Models

Mixture models suppose the following data generation scheme:

- The true class labels $\{y_1, \ldots, y_N\}$ of data points are realizations of independent and identically distributed (i.i.d) random variables $Y_1, \ldots, Y_N \sim Y$ taking their values in the set of all K classes $\mathcal{Y} = \{c_1, \ldots, c_K\}$ and distributed according to a multinomial distribution $\mathcal{M}(1, \pi_1, \ldots, \pi_K)$. The π_k are thus the class proportions and they verify $\sum_{k=1}^{K} \pi_k = 1$. The information on the true class labels of samples coming from such variables can also be expressed by a binary variable $\mathbf{z}_i \in \{0,1\}^K$, such that $z_{ik} = 1$ if $y_i = c_k$, and $z_{ik} = 0$ otherwise.
- The observed values $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are drawn using the class conditional density in relation with the class label. More formally, $X_1, \ldots, X_N \sim X$ are continuous random variables taking values in $\mathcal{X}$, with conditional probability density functions $f(\mathbf{x}|Y = c_k) = f(\mathbf{x}; \boldsymbol{\theta}_k), \quad \forall k \in \{1, \ldots, K\}$.

The parameters that need to be estimated are therefore the proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and the parameters of the class conditional densities $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K$. To simplify the notations, the vector of all model parameters is denoted $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$. In unsupervised learning problems, the available data are only the i.i.d realizations of X, $\mathbf{X}^u = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, provided by the generative model. To learn the parameters and the associated clustering, the log-likelihood must be computed according to the marginal density $\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)$ of X_i. This leads to the unsupervised log-likelihood criterion :

$$L(\boldsymbol{\Psi}; \mathbf{X}^u) = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \right). \tag{1.6}$$

1.3.2 EM Algorithm

The log-likelihood function defined by (1.6) is difficult to optimize and may lead to a set of different local maxima. The EM algorithm [4] is nowadays the classical solution to this problem. The missing data of the clustering problem are the true class labels y_i of learning examples. The basis of the EM algorithm can be found in the decomposition of the likelihood function in two terms :

$$L(\boldsymbol{\Psi}; \mathbf{X}^u) = \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik}^{(q)} \ln \left(\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k) \right)}_{Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})} - \underbrace{\sum_{i=1}^{N} \sum_{k=1}^{K} t_{ik}^{(q)} \ln \left(\frac{\pi_k f(\mathbf{x}_i; \boldsymbol{\theta}_k)}{\sum_{k'=1}^{K} \pi_{k'} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'})} \right)}_{H(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})}, \tag{1.7}$$

with:

$$t_{ik}^{(q)} = \mathbb{E}_{\boldsymbol{\Psi}^{(q)}}[z_{ik}|\mathbf{x}_i] = \mathbb{P}(z_{ik} = 1|\boldsymbol{\Psi}^{(q)}, \mathbf{x}_i) = \frac{\pi_k^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^{K} \pi_{k'}^{(q)} f(\mathbf{x}_i; \boldsymbol{\theta}_{k'}^{(q)})}. \tag{1.8}$$

Such a decomposition is useful to define an iterative ascent strategy thanks to the form of H. As a consequence of Jensen's inequality we may write $H(\boldsymbol{\Psi}^{(q)}, \boldsymbol{\Psi}^{(q)}) - H(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)}) \geq 0, \forall \boldsymbol{\Psi}$. Consequently, the maximization of the auxiliary function $\boldsymbol{\Psi}^{(q+1)} = \arg\max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}, \boldsymbol{\Psi}^{(q)})$ is sufficient to improve the likelihood. Furthermore, because the sum over the classes is outside the logarithm in the Q function, the optimization problems are decoupled and the maximization is simpler. The EM algorithm can be described as follows. It starts with initial estimates $\boldsymbol{\Psi}^{(0)}$ and alternates two steps : the E step where the t_{ik} are computed according to the current parameters estimates, defining a new Q function maximized during the M step. Thanks to (1.7), this defines a sequence of parameter estimates with increasing likelihood values. Finally, the mixture model setting and

hal-00446806, version 1 - 13 Jan 2010

the EM algorithm can be adapted to handle specific learning problems such as the semi-supervised [10] and the partially supervised cases [1].

1.4 Extension to Imprecise and Uncertain Labels

1.4.1 Derivation of a Generalized Likelihood Criterion

Our method extends the approach described above to handle *imprecise* and *uncertain* class labels defined by belief functions. In this section, we shall assume the learning set to be of the form $\mathbf{X}^{iu} = \{(\mathbf{x}_1, m_1^{\mathcal{Y}}), \ldots, (\mathbf{x}_N, m_N^{\mathcal{Y}})\}$, where each $m_i^{\mathcal{Y}}$ is a bba on the set $\mathcal{Y}$ of classes, encoding all available information about the class of example i. As before, the $\mathbf{x}_i$ will be assumed to have been generated according to the mixture model defined in Section 1.3.1. Our goal is to extend the previous method to estimate the model parameters from such dataset. For that purpose, an objective function generalizing the likelihood function needs to be defined.

The concept of likelihood function has strong relations with that of possibility and, more generally, plausibility, as already noted by several authors [16, 15, 13]. Furthermore, selecting the simple hypothesis with highest plausibility given the observations $\mathbf{X}^{iu}$ is a natural decision strategy in the belief function framework [2]. We thus propose as an estimation principle to search for the value of parameter $\boldsymbol{\psi}$ with maximal conditional plausibility given the data: $\hat{\boldsymbol{\psi}} = \arg\max_{\boldsymbol{\psi}} pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})$. The correctness of the intuition leading to this choice of criterion as an estimation principle seems to be confirmed by the fact that the logarithm of $pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})$ is an immediate generalization of criterion (1.6), and the other likelihood criteria used for semi-supervised learning and partially supervised learning of mixture model, as shown by the following proposition.

Proposition 1. *If the samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are drawn independently according to the generative mixture model setting and if the soft labels $\{m_1, \ldots, m_N\}$ are independent from the parameters values, then the logarithm of the conditional plausibility of $\boldsymbol{\Psi}$ given $\mathbf{X}^{iu}$ is given by*

$$\ln\left(pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu})\right) = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} pl_{ik}.\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) + \nu, \tag{1.9}$$

where the pl_{ik} are the plausibilities of each class k for each sample i according to soft labels m_i and ν is a constant independent of $\boldsymbol{\psi}$.

Proof. Using the GBT (1.5), the plausibility of parameters can be expressed from the plausibility of the observed values. By making the conditional independence assumption, this plausibility can be decomposed as a product over samples. Using the Total Plausibility Theorem (1.4), we may express the plausibility of an observed value as:

$$pl^{\mathcal{X}_i}(\mathbf{x}_i|\boldsymbol{\psi}) = \sum_{C \subseteq \mathcal{Y}} m^{\mathcal{Y}_i}(C|\boldsymbol{\psi}) pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|C, \boldsymbol{\psi}), \tag{1.10}$$

where $m^{\mathcal{Y}_i}(.|\boldsymbol{\psi})$ is a bba representing our beliefs regarding the class of example i. This bba comes from the combination of two information sources: the "soft" label $m_i^{\mathcal{Y}}$ and the proportions $\boldsymbol{\pi}$, which induce a Bayesian bba $m^{\mathcal{Y}}(\cdot|\boldsymbol{\pi})$. As these two sources are supposed to be distinct, they can be combined using the conjunctive rule (1.2). As $m^{\mathcal{Y}}(\cdot|\boldsymbol{\pi})$ is Bayesian, the same property holds for the result of the combination $m^{\mathcal{Y}_i}(.|\boldsymbol{\psi})$ and we have $m^{\mathcal{Y}_i}(\{c_k\}|\boldsymbol{\psi}) = pl_{ik}\,\pi_k$. Therefore, in the right-hand side of (1.10), the only terms in the sum that need to be considered are those corresponding to the singletons. Consequently, we only need to express $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k, \boldsymbol{\psi})$ for all k. There is a difficulty at this stage, since $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\cdot|c_k, \boldsymbol{\psi})$ is the continuous probability measure with density function $f(\mathbf{x};\boldsymbol{\theta}_k)$: consequently, the plausibility of any single value would be null if observations $\mathbf{x}_i$ had an infinite precision. However, observations always have a finite precision, so that what we denote by $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k, \boldsymbol{\psi})$ is in fact the plausibility of a infinitesimal region around $\mathbf{x}_i$ with volume $\mathrm{d}x_{i1} \ldots \mathrm{d}x_{ip}$ (where p is the feature space dimen-

hal-00446806, version 1 - 13 Jan 2010

sion). We thus have $pl^{\mathcal{X}_i|\mathcal{Y}_i}(\mathbf{x}_i|c_k,\boldsymbol{\psi}) = f(\mathbf{x}_i;\boldsymbol{\theta}_k)\mathrm{d}x_{i1}\dots\mathrm{d}x_{ip}$. Using all this results we obtain $pl^{\boldsymbol{\Psi}}(\boldsymbol{\psi}|\mathbf{X}^{iu}) = \prod_{i=1}^N \left[\left(\sum_{k=1}^K pl_{ik}\pi_k f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right)\mathrm{d}x_{i1}\dots\mathrm{d}x_{ip}\right]$. The terms $\mathrm{d}x_{ij}$ can be considered as multiplicative constants that do not affect the optimization problem. By taking the logarithm we get (1.9), which completes the proof. □

Remark 1. Our approach can be shown to extend unsupervised, partially supervised and semi-supervised learning when the labels are, respectively, vacuous, categorical, and either vacuous or certain. This justifies denoting criterion, 1.9 as $L(\boldsymbol{\Psi},\mathbf{X}^{iu})$, as it generalizes the classical log-likelihood function.

1.4.2 EM algorithm for Imprecise and Uncertain Labels

Once the criterion is defined, the remaining work concerns its optimization. This section presents a variant of the EM algorithm dedicated to this task. To build an EM algorithm able to optimize $L(\boldsymbol{\Psi};\mathbf{X}^{iu})$, we follow a path that parallels the one recalled in Section 1.3.2. At iteration q, our knowledge of the class of example i given the current parameter estimates comes from three sources: the class label $m_i^{\mathcal{Y}}$ of example i; the current estimates $\boldsymbol{\pi}^{(q)}$ of the proportions; the vector $\mathbf{x}_i$ and the current parameter estimate $\boldsymbol{\theta}^{(q)}$, which, using the GBT (1.5), gives a plausibility over $\mathcal{Y}$. By combining these three items of evidence using Dempster's rule (1.3), we get a Bayesian bba. Let us denote by $t_{ik}^{(q)}$ the mass assigned to $\{c_k\}$ after combination. We have

$$t_{ik}^{(q)} = \frac{pl_{ik}\pi_k^{(q)} f(\mathbf{x}_i;\boldsymbol{\theta}_k^{(q)})}{\sum_{k'=1}^K pl_{ik'}\pi_{k'}^{(q)} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'}^{(q)})}, \tag{1.11}$$

Using this expression, we may decompose the log-likelihood in two parts, as in (1.7).

$$L(\boldsymbol{\Psi};\mathbf{X}^{iu}) = \sum_{i=1}^N\sum_{k=1}^K t_{ik}^{(q)} \ln\left(\pi_k pl_{ik} f(\mathbf{x}_i;\boldsymbol{\theta}_k)\right) - \sum_{i=1}^N\sum_{k=1}^K t_{ik}^{(q)} \ln\left(\frac{\pi_k pl_{ik} f(\mathbf{x}_i;\boldsymbol{\theta}_k)}{\sum_{k'=1}^K \pi_{k'} pl_{ik'} f(\mathbf{x}_i;\boldsymbol{\theta}_{k'})}\right) \tag{1.12}$$

This decomposition can be established thanks to basic properties of logarithmic functions and the fact that $\sum_{i=1}^K t_{ik}^{(q)} = 1$. Therefore, using the same argument as for the classical EM algorithm (Section 1.3.2), an algorithm which alternates between computing t_{ik} using (1.11) and maximization of the first term in the right hand side of (1.12) will increase our criterion. This algorithm is therefore the classical EM algorithm, except for the E step, where the posterior distributions t_{ik} are weighted by the plausibility of each class. During the M step the proportions are updated classically using $\pi_k^{(q+1)} = \frac{1}{N}\sum_{i=1}^N t_{ik}^{(q)}$. If multivariate normal densities functions are considered, $f(x;\boldsymbol{\theta}_k) = \mathcal{N}(x;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$, their parameters are updated using the following equations :

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{\sum_{i=1}^N t_{ik}^{(q)}}\sum_{i=1}^N t_{ik}^{(q)}\mathbf{x}_i, \quad \boldsymbol{\Sigma}_k^{(q+1)} = \frac{1}{\sum_{i=1}^N t_{ik}^{(q)}}\sum_{i=1}^N t_{ik}^{(q)}(\mathbf{x}_i-\boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i-\boldsymbol{\mu}_k^{(q+1)})'. \tag{1.13}$$

1.4.3 Comparison with Previous Work

The idea of adapting the EM algorithm to handle soft labels can be traced back to the work of Vannoorenberghe and Smets [21], which was recently extended to categorical data by Jraidi et al. [12]. These authors proposed a variant of the EM algorithm called CrEM (Credal EM), based on a modification of the auxiliary function $Q(\boldsymbol{\Psi},\boldsymbol{\Psi}^{(q)})$. However, our method differs from this previous approach in several respects. First, the CrEM algorithm was not derived as optimizing a generalized likelihood criterion such as (1.9); consequently, its interpretation was unclear, the relationship with related work (see Remark 1) could not be highlighted and, most importantly,

hal-00446806, version 1 - 13 Jan 2010

the convergence of the algorithm was not proven. Furthermore, in our approach, the soft labels $m_i^{\mathcal{Y}}$ appear in the criterion and in the update formulas for posterior probabilities (1.11) only in the form of the plausibilities pl_{ik} of the singletons. In constrast, the CrEM algorithm uses the $2^{|\mathcal{Y}|}$ values in each bba $m_i^{\mathcal{Y}}$. This fact has an important consequence, as the computations involved in the E step of the CrEM algorithm have a complexity in $O(2^{|\mathcal{Y}|})$ whereas our solution only involves calculations which scale with the cardinality of the set of classes.

1.5 Simulations

The experiment presented in this section aimed at using information on class labels simulating expert opinions. As a reasonable setting, we assumed that the expert supplies, for each sample i, his/her more likely label c_k and a measure of doubt p_i. This doubt is represented by a number in $[0, 1]$, which can be seen as the probability that the expert knows nothing about the true label. To handle this additional information in the belief function framework, it is natural to *discount* the categorical bba associated to the guessed label with a discount rate p_i [14, Page 251]. Thus, the imperfect labels built from expert opinions are simple bbas such that $m_i^{\mathcal{Y}}(\{c_{k^*}\}) = 1 - p_i$ for some k^*, and $m_i^{\mathcal{Y}}(\mathcal{Y}) = p_i$. The corresponding plausibilities are $pl_{ik^*} = 1$ and $pl_{ik} = p_i$ for all $k \neq k^*$.

Simulated data sets were build as follows. Two data sets of size $N \in \{2000, 4000\}$ were generated in a ten-dimensional feature space from a two component normal mixture with common identity covariance matrix and balanced proportions. The distance between the two centers was kept fixed at $\delta = 2$. For each training sample i, a number p_i was drawn from a specific probability distribution to define the doubt expressed by a hypothetical expert on the class of that sample. With probability $(1 - p_i)$, the true label of sample i was kept and with probability p_i the expert's label was drawn uniformly in the set of all class. The probability distribution used to draw the p_i specifies the expert's labelling error rate. For our experiments we used Beta distributions with expected value equal to $\{0.1, \ldots, 0.8\}$ and variance kept equal to 0.2.

The results of our approach were compared to *supervised learning* using the potentially wrong expert's labels; *unsupervised learning*, which does not use any information on class label coming from experts, and a strategy based on *semi-supervised learning* which takes into account the reliability of labels supplied by the p_i. This strategy considers each sample as labelled if the expert's doubt is moderate ($p_i \leq 0.5$) and as unlabelled otherwise ($p_i > 0.5$). Figure 1.1 shows the averaged

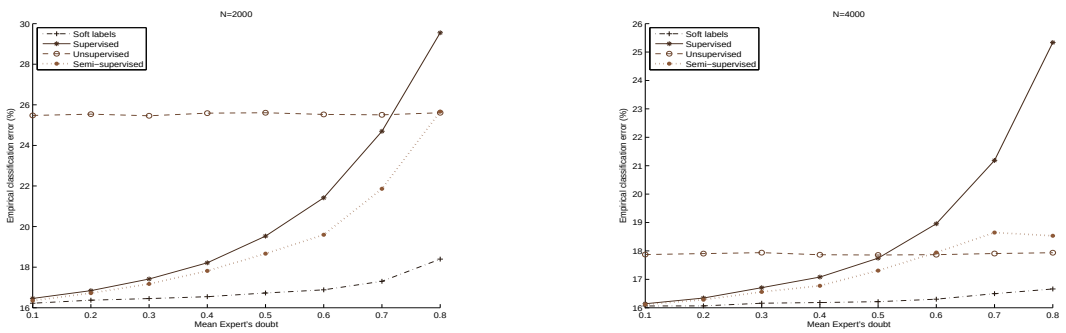

Fig. 1.1 Empirical classification error (%, estimated on a test set of 5000 observations) averaged over one hundred independent training sets, as a function of expert's mean doubt and for different sample size. For all methods, the EM algorithm was initialized with the true parameter values.

performances of the different classifiers trained with one hundred independent training sets. As expected, when the expert's doubt increases, the error rate of supervised learning also increases. Our solution based on soft labels does not suffer as much as supervised learning and adaptive semi-supervised learning from label noise. Whatever the dataset size, our solution takes advantage of additional information on the reliability of labels to keep good performances. Finally, our approach clearly outperforms unsupervised learning, when the number of samples is low ($N = 2000$).

hal-00446806, version 1 - 13 Jan 2010

1.6 Conclusions

The approach presented in this paper, based on concepts coming from maximum likelihood estimation and belief function theory, offers an interesting way to deal with imperfect and imprecise labels. The proposed criterion has a natural expression that is closely related to previous solutions found in the context of probabilistic models, and has also a clear and justified origin in the context of belief functions. Moreover, the practical interest of imprecise and imperfect labels, as a solution to deal with label noise, has been highlighted by an experimental study using simulated data.

References

[1] C. Ambroise and G. Govaert. EM algorithm for partially known labels. In IFCS' 00, pages 161–166, Namur, Belgium, 2000. Springer.
[2] B. R. Cobb and P. P. Shenoy. On the plausibility transformation method for translating belief function models to probability models. Int. Jour. of Appr. Reasoning, 41(3):314–330, 2006.
[3] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. Ann. of Math. Stat., 38:325–339, 1967.
[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Jour. of the Royal Stat. Soc., B 39:1–38, 1977.
[5] T. Denœux. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. on Systems, Man and Cybernetics, 25(05):804–813, 1995.
[6] T. Denœux and L. M. Zouhal. Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets and Sys., 122(3):47–62, 2001.
[7] D. Dubois and H. Prade. On the unicity of Dempster's rule of combination. Int. Jour. of Intel. Sys., 1:133–142, 1986.
[8] Z. Elouedi, K. Mellouli, and Ph. Smets. Belief decision trees: Theoretical foundations. Int. Jour. of Appr. Reasoning, 28:91–124, 2001.
[9] Y. Grandvallet. Logistic regression for partial labels. In IPMU' 02, volume III, pages 1935–1941, Annecy, France, 2002.
[10] D. W. Hosmer. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. Biometrics, 29:761–770, 1973.
[11] E. Hüllermeier and J. Beringer. Learning from ambiguously labeled examples. In IDA' 05, Madrid, Spain, September 2005.
[12] I. Jraidi and Z. Elouedi. Belief classification approach based on generalized credal EM. In K. Mellouli, editor, ECSQARU '07, pages 524–535, October 2007. Springer.
[13] P.-A. Monney. A Mathematical Theory of Arguments for Statistical Evidence. Contributions to Statistics. Physica-Verlag, Heidelberg, 2003.
[14] G. Shafer. A mathematical theory of evidence. Princeton University Press, 1976.
[15] P. P. Shenoy and P. H. Giang. Decision making on the sole basis of statistical likelihood. Artif. Intel., 165(2):137–163, 2005.
[16] Ph. Smets. Possibilistic inference from statistical data. In WCMSM' 82, pages 611–613, Las-Palmas, Spain, 1982.
[17] Ph. Smets. Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. Int. Jour. of Appr. Reasoning, 9:1–35, 1993.
[18] Ph. Smets. Quantifying beliefs by belief functions: An axiomatic justification. In IJCAI' 93, volume 1, pages 598–603, Chambéry, 1993.
[19] Ph. Smets. Belief functions on real numbers. Int. Jour. of Appr. Reasoning, 40(3):181–223, 2005.
[20] Ph. Smets and R. Kennes. The Transferable Belief Model. Artif. Intel., 66:191–243, 1994.
[21] P. Vannoorenberghe and P. Smets. Partially supervised learning by a credal EM approach. In ECSQARU' 05, pages 956–967, Barcelona, Spain, 2005. Springer.

hal-00446806, version 1 - 13 Jan 2010