Adv Data Anal Classif (2017) 11:659–690
https://doi.org/10.1007/s11634-017-0301-2

REGULAR ARTICLE

Parametric classification with soft labels using the evidential EM algorithm: linear discriminant analysis versus logistic regression

Benjamin Quost[1] · **Thierry Denœux**[1,2] · **Shoumei Li**[2]

Received: 12 April 2017 / Revised: 19 October 2017 / Accepted: 6 November 2017 /
Published online: 11 November 2017

Abstract Partially supervised learning extends both supervised and unsupervised learning, by considering situations in which only partial information about the response variable is available. In this paper, we consider partially supervised classification and we assume the learning instances to be labeled by Dempster–Shafer mass functions, called soft labels. Linear discriminant analysis and logistic regression are considered as special cases of generative and discriminative parametric models. We show that the evidential EM algorithm can be particularized to fit the parameters in each of these models. We describe experimental results with simulated data sets as well as with two real applications: K-complex detection in sleep EEGs signals and facial expression recognition. These results confirm the interest of using soft labels for classification as compared to potentially erroneous crisp labels, when the true class membership is partially unknown or ill-defined.

The author thanks Cédric Richard and Régis Lengellé for providing the EEG data, as well as the Foundation for Applied Neuroscience Research in Psychiatry (CHS de Rouffach, 68250 Rouffach, France) for authorizing their use. They also thank Franck Davoine for providing the facial expression data. This research was supported by a grant from the Beijing Government as part of the Overseas Talents program.

✉ Thierry Denœux
thierry.denoeux@utc.fr

Benjamin Quost
benjamin.quost@utc.fr

Shoumei Li
lisma@bjut.edu.cn

1 CNRS, Heudiasyc (UMR 7253), Sorbonne Universités, Université de Technologie de Compiègne, Compiègne, France

2 College of Applied Sciences, Beijing University of Technology, Beijing, China

Keywords Partially supervised learning · Belief functions · Dempster–Shafer theory · Machine learning · Uncertain data · Discriminant analysis · Logistic regression

Mathematics Subject Classification 62H30 · 62F86 · 68T10 · 68T37

1 Introduction

Classically, a distinction is made in data analysis and machine learning between *supervised* and *unsupervised* learning. In supervised learning, the data set is composed of observations of the response variable and a list of input variables for n individuals of some population. The problem is then to predict the value of the response variable for a new individual. In unsupervised learning, no response variable is observed, and the task is to find some underlying structure in the data (such as a partition or a manifold).

In recent years, though, the distinction between supervised and unsupervised learning has been blurred by the introduction of new paradigms that lie between these two extremes. One of these paradigms is *semi-supervised* learning (Chapelle et al. 2006), in which the response variable is perfectly known for some individuals, and totally unknown for others. An even more general paradigm is *partially supervised learning* (Denœux 1995; Denœux and Zouhal 2001; Hüllermeier and Beringer 2005; Nguyen and Caruana 2008; Côme et al. 2009; Cour et al. 2011), in which the values of the response variable for learning instances is only assumed to be partially known or uncertain, i.e., is subject to some hard or soft constraints. In Denœux (1995), the author first considered a type of classification problem in which partial knowledge about class labels is expressed in the Dempster–Shafer (DS) framework by belief functions (Dempster 1967; Shafer 1976): we can then speak of *soft labels* (Côme et al. 2009). The interest of DS theory in this context relies on the generality of belief functions, which encompass probabilities, sets and possibility measures as special cases: using belief functions to label training data thus allows us to encode various pieces of evidence about class labels. In particular, soft labels typically occur when no ground truth about class labels is available and data have to be labeled by experts or using some indirect method, or when the presence of noise casts doubt on the observations of the response variable. In these cases, using soft labels makes it possible to model the available side knowledge regarding the class information to be predicted. Due to the generality of the DS framework, the concept of soft label allows us to express uncertainty about class labels in a variety of ways, including sets, possibility distributions or probability distributions.

To learn from data with soft labels, nonparametric techniques such as the evidential k-nearest neighbor rule (Denœux 1995) and decision trees (Denœux and Skarstein-Bjanger 2000; Elouedi et al. 2001; Trabelsi et al. 2007) have first been proposed. A general mechanism for parametric inference in the presence of uncertain data (of which soft labels are a special case) was later introduced in Denœux (2013). In this approach, uncertain class information is represented by mass functions. The notion of likelihood can be extended to such soft class labels: it can be shown to depend on the contour functions associated with the uncertain observations. Maximizing this generalized likelihood generally requires using an iterative procedure. Such a procedure, called

the *Evidential EM* (E^2M) *algorithm*, was introduced in Denœux (2013); it extends the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) by allowing for maximum-likelihood estimation from *partially missing* data. The E^2M algorithm has been applied to a variety of tasks and models, including partially supervised Independent Factor Analysis (Cherfi et al. 2012), Hidden Markov Models with partially hidden states (Ramasso and Denœux 2013), fuzzy data clustering using Gaussian mixture models (Quost and Denoeux 2016), decision trees (Sutton-Charani et al. 2013, 2014; Ma et al. 2016), and mixture models with progressively censored data (Zhou et al. 2014).

Two types of parametric models are commonly used in classification: *generative models* describe the joint distribution of the input vector W and the class label Z, while *discriminative models* represent the conditional distribution of Z given W. Generative models, such as Gaussian mixture models, are suitable both for unsupervised learning (clustering) and for supervised classification. In contrast, discriminative models (such as logistic regression) cannot be used in an unsupervised setting, because they need observations of the response variable Y. However, they rely on fewer adaptive parameters and less restrictive assumptions, which may result in better predictive performance in supervised learning tasks, especially when class-conditional distributions are misspecified Press and Wilson (1978) (Bishop 2006, p. 204). The question then arises of the relative performances of generative and discriminative models in the presence of soft labels. To the best of our knowledge, this question has not been addressed until now. It will be investigated in this paper, with emphasis on linear classification. As representatives of generative and discriminative classifiers, we will consider, respectively, partially supervised Linear Discriminant Analysis (LDA), and Logistic Regression (LR) based on the E^2M algorithm. These two techniques will be introduced and compared using simulated and real data with soft labels obtained both by simulation, and from real experts. The objective of this paper is to study the influence of label uncertainty on the performances of these two classification models and, if possible, to formulate prescriptions regarding the suitability of each of these two approaches in situations where class labels are uncertain. Another problem addressed in this paper concern the assessment of classifier performances based on partially labeled data, a problem which had not received attention before. We propose a solution based on the notions of lower and upper expected losses. This method is useful, in particular, for model (e.g., classifier or feature) selection in presence of data with soft labels.

The rest of this paper is organized as follows. Background information about the theory of belief functions and the E^2M algorithm will first be recalled in Sect. 2. Partially supervised LDA and LR will then be introduced in Sect. 3. The issues of performance evaluation and model selection will be addressed in Sect. 4, and experimental results will be presented in Sect. 5. Section 6 will conclude the paper.

2 Background

In order to make the paper self-contained, the basic concepts of Dempster–Shafer theory will first be recalled in Sect. 2.1. The notion of evidential likelihood and the E^2M algorithm will then be described in Sects. 2.2 and 2.3, respectively.

2.1 Dempster–Shafer theory

In Dempster–Shafer theory (Shafer 1976), uncertain knowledge about some variable X taking values in a finite set $\mathcal{X}$ is described by a *mass function*, which is defined as a mapping m from the power set $2^{\mathcal{X}}$ to $[0, 1]$ verifying

$$\sum_{A \subseteq \mathcal{X}} m(A) = 1$$

and $m(\emptyset) = 0$. The subsets A of $\mathcal{X}$ such that $m(A) > 0$ are called the *focal sets* of m. A mass function m with a single focal set is said to be *logical*. Logical mass functions are in one-to-one correspondence with the subsets of $\mathcal{X}$. The mass function $m_?$ such that $m_?(\mathcal{X}) = 1$ is said to be *vacuous*; it represents total ignorance.

From a mass function m, we can compute a belief function and a plausibility function, defined, respectively, as

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

and

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\overline{A}),$$

for all $A \subseteq \mathcal{X}$. Functions Bel and Pl are, respectively, completely monotone and completely alternating capacities (Shafer 1976). When all the focal sets of m are singletons, m is said to be *Bayesian*; Bel and Pl then boil down to the same probability measure. When m is *consonant*, i.e., when its focal sets are nested, then Pl is a possibility measure (Zadeh 1978): it verifies

$$Pl(A \cup B) = \max(Pl(A), Pl(B))$$

for all $A, B \subseteq \mathcal{X}$. The mapping $pl : \mathcal{X} \to [0, 1]$ defined by $pl(x) = Pl(\{x\})$ for any $x \in \mathcal{X}$ is called the *contour function* of m. When m is consonant, the following equalities hold,

$$Pl(A) = \max_{x \in A} pl(x)$$

for all $A \subseteq \mathcal{X}$, and $\max_{x \in} pl(x) = 1$. When m is Bayesian, pl is a probability mass function and $\sum_{x \in \mathcal{X}} pl(x) = 1$.

The *discounting* operation (Shafer 1976) allows us to take into account the reliability of a source of information. Assume that a source provides us with a mass function m, and we have a degree of confidence $1 - \alpha$ in the reliability of the source. We can then construct a new mass function ${}^{\alpha}m$ defined as

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

Parameter α is called the *discount rate*. If $\alpha = 0$, the mass function is unchanged. If $\alpha = 0$, the discounted mass function is vacuous.

Given two mass function m_1 and m_2, their *orthogonal sum* is defined as the mass function $m_1 \oplus m_2$ such that $(m_1 \oplus m_2)(\emptyset) = 0$ and

$$(m_1 \oplus m_2)(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1(B) m_2(C) \tag{2}$$

for all $A \subseteq \mathcal{X}$, $A \neq \emptyset$, where

$$\kappa = \sum_{B \cap C = \emptyset} m_1(B) m_2(C) \tag{3}$$

is called the *degree of conflict* between m_1 and m_2. The orthogonal sum of m_1 and m_2 is well-defined as long as they are not totally conflicting, i.e., their degree of conflict is strictly less than 1. Operation $\oplus$ is called *Dempster's rule of combination*. The contour function $pl_1 \oplus pl_2$ of $m_1 \oplus m_2$ is proportional to the product of the contour functions pl_1 and pl_2 associated with m_1 and m_2: we have

$$(pl_1 \oplus pl_2)(x) = \frac{pl_1(x) pl_2(x)}{1-\kappa} \tag{4}$$

for all $x \in \mathcal{X}$. The orthogonal sum of a Bayesian mass function m_1 and an arbitrary mass function m_2 is Bayesian; the degree of conflict between m_1 and m_2 can then be obtained from the contour functions as

$$\kappa = 1 - \sum_{x \in \mathcal{X}} pl_1(x) pl_2(x). \tag{5}$$

2.2 Evidential likelihood

Let X be a discrete random vector with finite sample space $\mathcal{X}$ and probability mass function $p_X(x;\theta)$ assumed to be known up to a parameter $\theta \in \Theta$. After a realization x of X has been observed, the (complete-data) likelihood function is the mapping from Θ to [0, 1] defined by

$$L_c(\theta) = p_X(x;\theta), \quad \forall \theta \in \Theta. \tag{6}$$

Let us now assume that x is not observed precisely, but we collect some evidence about x. This evidence induces partial knowledge of x described by a mass function m on $\mathcal{X}$. The likelihood function (6) can then be generalized (Denœux 2013) to

$$L(\theta) = \sum_{A \subseteq \mathcal{X}} m(A) \sum_{x \in A} p_X(x;\theta), \quad \forall \theta \in \Theta, \tag{7}$$

Function $L(\theta)$ defined by (7) is called the *evidential likelihood function* induced by the uncertain data m. It must be emphasized that the notion of evidential likelihood is a new concept. This concept is distinct from the classical one of likelihood, because mass function m is *not* a realization of a random element. See Denœux (2013) for a detailed

discussion on this issue, and Couso and Dubois (2017) for alternative definitions of likelihood in the case of imprecise observations. When m is logical with $m(A) = 1$ for some $A \subseteq \mathcal{X}$, then $L(\theta)$ is simply the probability of the event $X \in A$. The evidential likelihood function then becomes identical to the likelihood function under the Coarsening at Random (CAR) assumption (Heitjan and Rubin 1991). However, we do not assume in our model that A is generated at random. Instead, the mass function characterizes *epistemic* uncertainty about x. This partial knowledge can be due to a indirect observation of the quantity of interest, to the subjectivity of the information source, or to the presence of noise with ill-known distribution. Whenever mass function m in (7) is certain, i.e., when $m(\{x\}) = 1$, the evidential likelihood (7) coincides with the classical likelihood (6), which only depends on the pdf p_X modeling the random data generating process.

By permuting the two summations in (7), we get another expression for $L(\theta)$ as

$$L(\theta) = \sum_{x \in \mathcal{X}} p_X(x;\theta) \sum_{A \ni x} m(A) = \sum_{x \in \mathcal{X}} p_X(x;\theta)\, pl(x), \tag{8}$$

where pl is the contour function associated to m. Comparing (8) and (5), we can see that $1 - L(\theta)$ equals the degree of conflict between the uncertain data m and the probability mass function $p(x;\theta)$. Maximizing $L(\theta; m)$ thus amounts to minimizing the conflict between the data and the model.

Equation (8) also reveals that $L(\theta;\, pl)$ can alternatively be viewed as the expectation of $pl(X)$,

$$L(\theta) = \mathbb{E}_\theta[pl(X)]. \tag{9}$$

In the special case where $X = (X_1, \ldots, X_n)$ is an independent sample, and assuming that the contour function pl can be decomposed as

$$pl(x) = pl_1(x_1) \ldots pl_n(x_n), \tag{10}$$

a property called *cognitive independence* by Shafer (1976), (9) simplifies to

$$L(\theta) = \prod_{i=1}^{n} \mathbb{E}_\theta[pl_i(X_i)]. \tag{11}$$

Finally, we can remark that the normalized likelihood $pl(\theta) = L(\theta)/L(\widehat{\theta})$, where $\widehat{\theta}$ is a maximizer of $L(\theta)$, and it is assumed that $L(\widehat{\theta}) < \infty$, can be interpreted as the contour function of a consonant belief function on Θ (or, equivalently, as a possibility distribution on Θ) (Denœux 2014). Equations (9) and (11) can directly be extended to the case of continuous data.

2.3 E^{2}M algorithm

The E^{2}M algorithm introduced in (Denœux 2013) is a generalization of the EM algorithm (Dempster et al. 1977), which allows one to maximize the evidential likelihood

(7)–(9). Similarly to the EM algorithm, each iteration q of E^{2}M is composed of two steps:

1. The E-step requires the combination of $p_X(x;\theta^{(q)})$, the probability mass function of X for the current estimate $\theta^{(q)}$ of θ, with the contour function pl. The result of this combination is a probability mass function $p_X(\cdot|pl;\theta^{(q)}) = p_X(\cdot;\theta^{(q)}) \oplus pl$, which can be computed as

$$p_X\left(x|pl;\theta^{(q)}\right) = \frac{p_X\left(x;\theta^{(q)}\right) pl(x)}{L\left(\theta^{(q)};\, pl\right)},$$

   for all $x \in \mathcal{X}$. Then, the expectation of the complete-data log-likelihood $\ell_c(\theta) = \log p_X(X;\theta)$ with respect to $p(x|pl;\theta^{(q)})$ is calculated,

$$Q\left(\theta,\theta^{(q)}\right) = \frac{\sum_{x\in\mathcal{X}} \ell_c(\theta;x)\, p_X\left(x;\theta^{(q)}\right) pl(x)}{L\left(\theta^{(q)};\, pl\right)}. \tag{12}$$

2. The M-step then consists in maximizing function $Q(\theta,\theta^{(q)})$ with respect to θ, finding the new estimate $\theta^{(q+1)}$ such that $Q(\theta^{(q+1)},\theta^{(q)}) \geq Q(\theta,\theta^{(q)})$ for all $\theta \in \Theta$.

The E- and M-steps are iterated until $L(\theta^{(q+1)}) - L(\theta^{(q)}) \leq \epsilon$ for some arbitrarily small ϵ.

The E^{2}M algorithm was shown in Denœux (2013) to increase the evidential likelihood as each iteration, i.e., to ensure that $L(\theta^{(q+1)}) \geq L(\theta^{(q)})$ for all q. Consequently, it converges to a local maximum of the evidential likelihood if this function is bounded from above. When the contour function pl corresponds to a logical mass function, i.e., when $pl(x) = I(x \in A)$ for some $A \subset \mathcal{X}$, where $I(\cdot)$ is the indicator function, the evidential likelihood—and thus the expectation (12)—coincide with their certain counterparts: the E^{2}M algorithm then boils down to the classical EM algorithm.

The E^{2}M algorithm has been applied to several Machine Learning problems. A first category of application concerns problems in which we want to exploit some information about variables that are usually considered as latent: we can then speak of "partially latent" variables. For instance, in Cherfi et al. (2012), the authors consider an Independent Factor Analysis model in which the discrete latent variables represent the states of railway track circuits, about which partial knowledge is elicited from experts. In Ramasso and Denœux (2013), the authors apply the E^{2}M algorithm to estimate the parameter of a Hidden Markov Model in which the hidden state variable is not totally "hidden", but partially observed; they describe an application to machine condition monitoring. A second category of applications concerns the problem of learning from uncertain data. For instance, Sutton-Charani et al. (2013, 2014) and Ma et al. (2016) applied the E^{2}M algorithm to decision tree inference from data with uncertain attributes, while the problem of clustering data with fuzzy attributes was considered by Quost and Denœux in (2016). In Zhou et al. (2014), the authors considered the modeling of lifetime data using mixture models and progressively censored observations. The task of learning a classifier from partially labeled data also belongs to this second category. This problem will be addressed in the next section.

3 Application to linear classification models

In this section, the complete data are assumed to consist in an i.i.d. sample $X = \{(W_i, Z_i)\}_{i=1}^{n}$ from (W, Z), where W is a d-dimensional random input vector and Z is the class variable, taking values in a finite set $\mathcal{Z} = \{1, \ldots, K\}$. The complete dataset $x = \{(w_i, z_i)\}_{i=1}^{n}$ is a realization from X. The notion of soft label will first be introduced in Sect. 3.1. The application of the E^{2}M algorithm to estimate the parameter of LDA and LR using uncertain class information will then be described, respectively, in Sects. 3.2 and 3.3. The complexity of partially supervised LDA and LR will then be discussed in Sect. 3.4.

3.1 Soft labels

In some applications, class information is uncertain. This is the case, in particular, when the class labels cannot be observed directly and have to be inferred by an unsupervised learning algorithm (see, e.g., Denœux et al. 2016; Liu et al. 2015, 2017) or by any other indirect method. Sometimes, class labels are assessed subjectively by an expert, a group of experts (as in the two examples described in Sect. 5.2), or even by a large number of individuals through crowdsourcing (see, e.g. Abassi and Boukhris 2016; Rjab et al. 2016). In the multiple-expert case, the more discordant the opinions are, the more doubt can be cast on the class information that would be obtained, for instance, by majority voting.

Rather than using crisp but possibly erroneous class information z_i, *soft* class labels may be used, so as to reflect the degree of confidence in each of the possible classes. Figure 1 illustrates the advantage of soft labels over crisp, but potentially erroneous labels. In this toy example, two doubtful instances have a strong influence on the decision boundary. By expressing lack of confidence in the class membership of these instances using soft labels, we decrease their influence on the classifier. When information about the class labels of some instances is not reliable, it is preferable to take this uncertainty into account, in order to decrease the influence of the most doubtful patterns on the decision rule. In this paper, we consider the representation of uncertainty in the Dempster–Shafer framework. A *soft label* for instance i is defined as a mass function m_i on the set $\mathcal{Z}$ of classes. An n-tuple $(m_1, \ldots, m_n)$ of mass functions for n learning instances is called a *credal partition* (Denœux and Masson 2004). As noted in Denoeux and Kanjanatarakul (2016), the notion of credal partition subsumes most other *soft clustering* notions. For instance, if mass functions m_i are consonant, they define possibility distributions and, equivalently, fuzzy subsets of $\mathcal{Z}$. If mass functions are logical (i.e., if they verify $m_i(A_i) = 1$ for some $A_i \subseteq \mathcal{Z}$), then we can define the lower and upper approximations of each class k as, respectively, the set of instances for which $A_i = \{k\}$, and the set of instances for which $k \in A_i$. We then recover notions from rough classification and clustering (Peters et al. 2013).

Because the E^{2}M algorithm described in Sect. 2.3 uses only the contour functions, we can equivalently represent soft labels by the plausibilities $pl_{ik} = pl_i(k)$ of each class k, where pl_i is the contour function of m_i. *Partially supervised learning* is then defined as the task of learning a classifier from a learning set $\{(w_i, pl_i)\}_{i=1}^{n}$ with soft

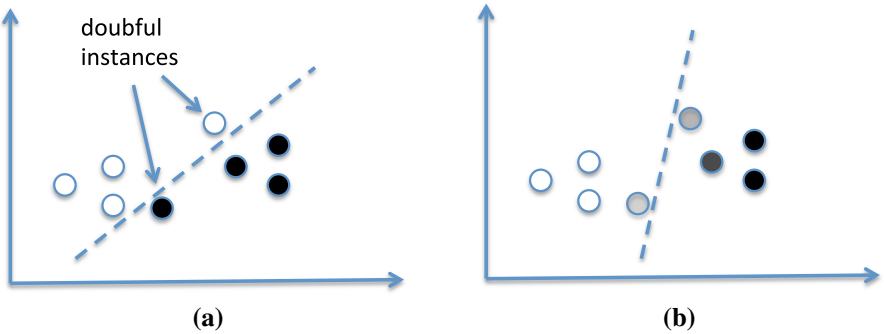

Fig. 1 Effect of soft labels. Left: crisp labels; two wrongly labeled instances result in an incorrect decision boundary. Right: soft labels; the influence of doubtful instances is decreased

labels. We can remark that the partially supervised paradigm encompasses most of the classical learning frameworks. Fully supervised learning is recovered when there is no uncertainty on class labels, i.e., when we have, for all i, $pl_{ik} = 1$ for some $k \in \{1, \ldots, K\}$ and $pl_{i\ell} = 0$ for all $\ell \neq k$. Unsupervised learning corresponds to the situation where the soft labels are vacuous, i.e., $pl_{ik} = 1$ for all i and all k. When some labels are crisp and some are vacuous, we get the semi-supervised learning paradigm. It is clear, however, that soft labelling goes much further than these classical formalisms, by allowing the user to specify numerical degrees of plausibility for each class and each learning instance. Induction strategies may then be adapted to such soft labels in order to avoid learning a biased model by attaching too much importance to doubtful information.

In the sequel, we will consider the application of the E^2M algorithm to estimate the parameter of two models: LDA and LR using data with soft labels.

3.2 Linear discriminant analysis

LDA is based on the assumption that the conditional distribution of W given $Z = k$ is multivariate normal with mean μ_k and covariance matrix Σ independent on k:

$$W|(Z = k) \sim \mathcal{N}(\mu_k, \Sigma), \quad k = 1, \ldots, K.$$

Let π_k be the marginal probability that $Z = k$, and

$$\theta = (\mu_1, \ldots, \mu_K, \Sigma, \pi_1, \ldots, \pi_{K-1})$$

the parameter vector. The complete-data likelihood is

$$L_c(\theta) = \prod_{i=1}^{n} p(w_i|Z_i = z_i)p(z_i) \tag{13a}$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{K} \phi(w_i; \mu_k, \Sigma)^{z_{ik}} \pi_k^{z_{ik}}, \tag{13b}$$

where $\phi(\cdot; \mu_k, \Sigma)$ is the multivariate normal density,

$$\phi(w; \mu_k, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\right\},$$

and z_{ik} is a binary class indicator variable, such that $z_{ik} = 1$ if $z_i = k$ and $z_{ik} = 0$ otherwise.

Under the assumption of cognitive independence (10), the contour function on $\mathcal{X}$ is $pl(x) = \prod_{i=1}^n pl_i(x_i)$, with

$$pl_i(x_i) = \begin{cases} pl_{ik} & \text{if } x_i = (w_i, k) \text{ for some } k = 1, \ldots, K \\ 0 & \text{otherwise.} \end{cases}$$

The evidential likelihood (11) is thus

$$L(\theta) = \prod_{i=1}^n \sum_{k=1}^K pl_{ik}\phi(w_i; \mu_k, \Sigma)\pi_k. \tag{14}$$

We can remark that, when there is no uncertainty, i.e., when $pl_{ik} = z_{ik}$ for all (i, k), we have

$$\sum_{k=1}^K pl_{ik}\phi(w_i; \mu_k, \Sigma)\pi_k = \prod_{k=1}^K \phi(w_i; \mu_k, \Sigma)^{pl_{ik}} \pi_k^{pl_{ik}},$$

and the evidential likelihood (14) becomes identical to the complete-data likelihood (13b). When, on the other hand, uncertainty is maximal, i.e., class labels are completely unknown, then $pl_{ik} = 1$ for all (i, k), and the evidential likelihood (14) becomes

$$L(\theta) = \prod_{i=1}^n \sum_{k=1}^K \phi(w_i; \mu_k, \Sigma)\pi_k,$$

which is the likelihood function corresponding to the unsupervised case.

In the E-step of the E^{2}M algorithm for this model, we compute the expectation of the complete-data log-likelihood

$$\ell_c(\theta) = \sum_{i=1}^n \sum_{k=1}^K z_{ik} \left[\log \phi(w_i; \mu_k, \Sigma) + \log \pi_k\right]$$

with respect to the combined probability mass function

$$p_X\left(x | pl; \theta^{(q)}\right) = \prod_{i=1}^n p\left(x_i | pl_i; \theta^{(q)}\right),$$

with

$$p(x_i|pl_i;\theta^{(q)}) = \begin{cases} \dfrac{pl_{ik}\pi_k^{(q)}\phi\left(w_i;\mu_k^{(q)},\Sigma^{(q)}\right)}{\sum_\ell pl_{i\ell}\pi_\ell^{(q)}\phi\left(w_i;\mu_\ell^{(q)},\Sigma^{(q)}\right)} & \text{if } x_i=(w_i,k) \text{ for some } k \\ 0 & \text{otherwise.} \end{cases}$$

We get

$$Q(\theta,\theta^{(q)}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\zeta_{ik}^{(q)}\left[\log\phi(w_i;\mu_k,\Sigma)\pi_k + \log\pi_k\right], \tag{15}$$

with

$$\zeta_{ik}^{(q)} = \mathbb{E}\left(Z_{ik}|pl;\theta^{(q)}\right) = \frac{pl_{ik}\pi_k^{(q)}\phi\left(w_i;\mu_k^{(q)},\Sigma^{(q)}\right)}{\sum_\ell pl_{i\ell}\pi_\ell^{(q)}\phi\left(w_i;\mu_\ell^{(q)},\Sigma^{(q)}\right)}. \tag{16}$$

We can remark that the function $Q(\theta,\theta^{(q)})$ defined by (15) has exactly the same form as the function computed in the E-step of the EM algorithm applied to the normal mixture model in the unsupervised case (see, e.g., McLachlan and Peel 2000, pp. 81–83). When class labels are completely unknown, i.e., $pl_{ik}=1$ for all i and k, then $\zeta_{ik}^{(q)}$ defined by (16) becomes equal to the conditional expectation of Z_{ik} given $W_i=w_i$, and E^{2}M boils down to the classical EM algorithm.

Because of the formal similarity with the EM algorithm, the parameter values maximizing $Q(\theta,\theta^{(q)})$ can be readily obtained as

$$\pi_k^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n}\zeta_{ik}^{(q)}, \qquad \mu_k^{(q+1)} = \frac{\sum_{i=1}^{n}\zeta_{ik}^{(q)}w_i}{\sum_{i=1}^{n}\zeta_{ik}^{(q)}}, \tag{17a}$$

$$\Sigma^{(q+1)} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\zeta_{ik}^{(q)}\left(w_i-\mu_k^{(q+1)}\right)\left(w_i-\mu_k^{(q+1)}\right)^{\mathrm{T}}. \tag{17b}$$

3.3 Logistic regression

In contrast with LDA, LR starts with a model of the conditional distribution of Z given $W=w$. Specifically, assume that the conditional probabilities of each class given $W=w$ are given by

$$p_k(w;\theta) = \frac{\exp\left(\beta_k^T\widetilde{w}\right)}{1+\sum_{\ell=1}^{K-1}\exp\left(\beta_\ell^T\widetilde{w}\right)}, \quad k=1,\ldots,K-1 \tag{18a}$$

$$p_K(w;\theta) = \frac{1}{1+\sum_{\ell=1}^{K-1}\exp\left(\beta_\ell^T\widetilde{w}\right)}, \tag{18b}$$

where $p_k(w;\theta)$ is a notation for $\Pr(Z=k|W=w;\theta)$, β_k is a $p+1$-dimensional vector of coefficients, $\theta=(\beta_1^T,\ldots,\beta_{K-1}^T)^T$ is the vector of all parameters in the

model, and $\widetilde{w} = (1, w^T)^T$ is an extended input vector. Logistic regression proceeds by maximizing the conditional likelihood

$$L_c(\theta) = \prod_{i=1}^{n} \Pr(Z_i = z_i | w_i; \theta) \tag{19a}$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{K} p_k(w; \theta)^{z_{ik}}. \tag{19b}$$

After plugging the expression of the conditional class probabilities (18) into (19b), and taking the logarithm, we obtain the following expression for the complete-data conditional log-likelihood,

$$\ell_c(\theta) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K-1} z_{ik} \beta_k^T \widetilde{w}_i - \log \left(1 + \sum_{k=1}^{K-1} \beta_k^T \widetilde{w}_i \right) \right\}. \tag{20}$$

Under the same cognitive independence assumption (10) as before, the evidential likelihood is

$$L(\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} pl_{ik}\, p_k(w_i; \theta). \tag{21}$$

We can easily check that $L(\theta) = L_c(\theta)$ whenever $pl_{ik} = z_{ik}$ for all (i, k), i.e., when there is no label uncertainty. On the other hand, in case of maximal uncertainty, i.e., when $pl_{ik} = 1$ for all (i, k), we have $L(\theta) = 1$ for all θ, and the model parameters can no longer be estimated.

The E^2M algorithm for the maximization of the evidential likelihood (21) can be described as follows (Quost 2014). In the E-step, we compute the expectation of the complete-data log-likelihood (20) with respect to the combined probability mass function

$$p_Z\left(z | pl; \theta^{(q)}\right) = \prod_{i=1}^{n} p_{Z_i}\left(z_i | pl_i; \theta^{(q)}\right),$$

with

$$p_{Z_i}\left(k | pl_i; \theta^{(q)}\right) = \frac{pl_{ik}\, p_k\left(w_i; \theta^{(q)}\right)}{\sum_\ell pl_{i\ell}\, p_\ell\left(w_i; \theta^{(q)}\right)}, \quad k = 1, \ldots, K.$$

We get

$$Q(\theta, \theta^{(q)}) = \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K-1} \zeta_{ik}^{(q)} \beta_k^T \widetilde{w}_i - \log \left(1 + \sum_{k=1}^{K-1} \beta_k^T \widetilde{w}_i \right) \right\}, \tag{22}$$

with

$$\zeta_{ik}^{(q)} = \mathbb{E}\left(Z_{ik} | pl; \theta^{(q)}\right) = \frac{pl_{ik}\, p_k\left(w_i; \theta^{(q)}\right)}{\sum_\ell pl_{i\ell}\, p_\ell\left(w_i; \theta^{(q)}\right)}. \tag{23}$$

The maximization of (22) cannot be performed in one step and requires an iterative optimization procedure, such as the Newton-Raphson (NR) algorithm. It is actually not necessary to maximize function $Q(\theta, \theta^{(q)})$: we may simply make a step uphill, i.e., find some new estimate $\theta^{(q+1)}$ such that $Q(\theta^{(q+1)}, \theta^{(q)}) > Q(\theta^{(q)}, \theta^{(q)})$. Such a procedure is classically called a *Generalized EM algorithm* (McLachlan and Krishnan 1997, p. 24). An uphill step starting from the previous estimate $\theta^{(q)}$ can be made by carrying out one iteration of the NR algorithm with line search, i.e., by using the following update rule,

$$\theta^{(q+1)} = \theta^{(q)} - \eta \left[\frac{\partial^2 Q(\theta, \theta^{(q)})}{\partial \theta \partial \theta^T} \right]^{-1}_{\theta=\theta^{(q)}} \left. \frac{\partial Q(\theta, \theta^{(q)})}{\partial \theta} \right|_{\theta=\theta^{(q)}}, \tag{24}$$

where η is the step size. Typically, η is initially set to 1 and, if the objective function has not increased, it is repeatedly cut by two until an uphill step is achieved. The gradient vector and Hessian matrix can be computed using matrix operations as in standard LR (see, e.g., Hasan et al. 2016; Li 2013 for details).

3.4 Complexity

Since partially supervised learning is a more difficult problem than fully supervised learning, we can expect partially supervised LDA or LR to require more computational resources than their classical counterparts. Yet, partially supervised learning algorithm have roughly the same complexity as the corresponding supervised learning procedures.

In classical LDA the maximum likelihood estimates (MLEs) have a closed-form expression, whereas in presence of soft labels they are determined after repeatedly computing the expected labels (16) and the model parameters (17a)–(17b) (Quost and Denoeux 2016). However, each iteration of the E^2M algorithm has the same complexity as that of the MLE calculation in classical LDA: consequently, the running time of partially supervised supervised LDA is roughly equal to that of classical LDA multiplied by the number of iterations of E^2M.

Similarly, partially supervised logistic regression involves repeatedly estimating the expected class labels (23) and the model parameters (24). However, parameter estimation with classical LR already involves an iterative procedure (the NR algorithm). As stated in Sect. 3.3, we can perform only one iteration of the NR algorithm at each M-step of the E^2M algorithm. Further, each iteration of the NR algorithm has the same complexity as one iteration of E^2M. As a consequence, there is no significant difference between the running times of classical and partially supervised LR.

4 Performance evaluation and model selection

Performance evaluation and model selection are fundamental issues when designing classifiers. Typically, a loss function is defined, and the performance of a classifier is measured by its expected loss, or *risk*. If the loss is assumed to be equal to 1 for

misclassification and 0 in case for correct classification, then the risk is the error probability. This probability may be estimated from an independent test set, or using cross-validation when the amount of data at hand is too low. In the case of data with soft labels, error (or risk) estimation becomes more difficult, because the true class is uncertain. This issue is addressed in this section.

4.1 Lower and upper expected loss

Assume that, for some test pattern i with soft label m_i, we get the predicted class $\widehat{z}_i$. Let $\lambda(\widehat{z}_i, z_i)$ denote the loss incurred if the true class is z_i. The actual loss is unknown, because we only have partial knowledge of z_i, expressed by mass function m_i. However, we can compute lower and upper expected values (Dempster 1967) as follows:

$$\lambda_*(\widehat{z}_i, m_i) = \sum_{A \subseteq \mathcal{Z}} m_i(A) \min_{k \in A} \lambda(\widehat{z}_i, k) \tag{25a}$$

$$\lambda^*(\widehat{z}_i, m_i) = \sum_{A \subseteq \mathcal{Z}} m_i(A) \max_{k \in A} \lambda(\widehat{z}_i, k). \tag{25b}$$

The lower and upper expected losses can be seen as, respectively, optimistic and pessimistic assessments of the unknown loss $\lambda(\widehat{z}_i, z_i)$. A trade-off between these two assessments can be achieved by computing a convex sum

$$\lambda_\rho(\widehat{z}_i, m_i) = \rho\lambda^*(\widehat{z}_i, m_i) + (1 - \rho)\lambda_*(\widehat{z}_i, m_i), \tag{26}$$

where $\rho \in [0, 1]$ is a pessimism index (Jaffray 1989; Strat 1990). Although the particular form of Eq. (26) can be justified axiomatically (Jaffray 1989), the theory tells us nothing about the choice of ρ, which models the attitude of the decision maker. In the absence of a better argument, we propose to fix $\rho = 0.5$, based on a symmetry consideration.

In the special case where λ is the 0–1 loss function, defined by $\lambda(\widehat{z}_i, z_i) = I(\widehat{z}_i \neq z_i)$, we have

$$\min_{k \in A} \lambda(\widehat{z}_i, k) = \begin{cases} 0 & \text{if } \widehat{z}_i \in A \\ 1 & \text{otherwise.} \end{cases}$$

Consequently,

$$\lambda_*(\widehat{z}_i, m_i) = \sum_{A \subseteq \mathcal{Z} : \widehat{z}_i \notin A} m_i(A) \tag{27a}$$

$$= 1 - \sum_{A \subseteq \mathcal{Z} : \widehat{z}_i \in A} m_i(A) \tag{27b}$$

$$= 1 - pl_i(\widehat{z}_i). \tag{27c}$$

Similarly,

$$\max_{k\in A} \lambda(\widehat{z}_i, k) = \begin{cases} 0 & \text{if } A = \{\widehat{z}_i\} \\ 1 & \text{otherwise.} \end{cases}$$

Hence,

$$\lambda^*(\widehat{z}_i, m_i) = 1 - m_i(\{\widehat{z}_i\}). \tag{28}$$

The width of the lower-upper expected loss interval reflects the imprecision of the mass function m_i. It is null if m_i is Bayesian, and it is equal to one if m_i is vacuous.

We can also remark that the lower expected 0–1 loss criterion (27) has an interesting alternative interpretation in terms of degree of conflict. Let $\widehat{m}_i$ be the *certain* mass function defined by $\widehat{m}_i(\{\widehat{z}_i\}) = 1$. The degree of conflict (3) between $\widehat{m}_i$ and m_i is

$$\kappa_i = \sum_{B\cap C=\emptyset} \widehat{m}_i(B) m_i(C) \tag{29a}$$

$$= \sum_{C\not\ni \widehat{z}_i} m_i(C) \tag{29b}$$

$$= 1 - pl_i(\widehat{z}_i) = \lambda_*(\widehat{z}_i, m_i). \tag{29c}$$

The lower expected loss is thus equal to the degree of conflict between the prediction $\widehat{z}_i$ and the mass function m_i representing one's knowledge about the true class.

4.2 Application to performance assessment

Given a test set of size n_t, with soft labels $\{m_i\}_{i=1}^{n_t}$, *lower and upper test error rates* can be defined by averaging, respectively, the lower and upper expected losses,

$$\underline{\lambda} = \frac{1}{n_t}\sum_{i=1}^{n_t} \lambda_*(\widehat{z}_i, m_i) = 1 - \frac{1}{n_t}\sum_{i=1}^{n_t} pl_i(\widehat{z}_i) \tag{30a}$$

$$\overline{\lambda} = \frac{1}{n_t}\sum_{i=1}^{n_t} \lambda^*(\widehat{z}_i, m_i) = 1 - \frac{1}{n_t}\sum_{i=1}^{n_t} m_i(\{\widehat{z}_i\}). \tag{30b}$$

These two quantities can be seen as, respectively, optimistic and pessimistic assessments of the test error rate. In the fully supervised case with certain mass function m_i verifying $m_i(\{z_i\}) = 1$, we have $pl_i(\widehat{z}_i) = m_i(\{\widehat{z}_i\}) = I(z_i = \widehat{z}_i)$; both $\underline{\lambda}$ and $\overline{\lambda}$ then boil down to the usual test error rate. In the unsupervised case, we have $\underline{\lambda} = 0$ and $\overline{\lambda} = 1$, meaning that the error rate cannot be assessed.

Given a set of models (such as, e.g., LDA and LR with different subsets of input variables), their performances can be assessed using cross-validation estimates of the error rates. Typically, the learning set is partitioned into B blocks of approximately equal size, and B classifiers are trained, leaving one of the B blocks aside. The classifier

performance is then assessed from its predictions on the block that was left aside. Lower and upper cross-validation estimates $\underline{\lambda}_{CV}$ and $\overline{\lambda}^{CV}$ can be obtained in this way.

Given two classifiers C_1 and C_2, and their lower and upper cross-validation error rates $\{\underline{\lambda}_{CV}^{(i)}, \overline{\lambda}_{CV}^{(i)}\}$, $i = 1, 2$, maximality leads to considering classifier C_1 as better than C_2 if $\overline{\lambda}_{CV}^{(1)} < \underline{\lambda}_{CV}^{(2)}$. This criterion yields only a partial order over classifiers. If a total order is needed, then we can base the comparison on the weighted means $(1-\rho)\underline{\lambda}_{CV}^{(i)} + \rho\overline{\lambda}_{CV}^{(i)}$, after fixing the value of the pessimism index ρ.

5 Experiments

In this section, we report on experiments with LDA and LR trained from data with soft labels. In Sect. 5.1, we will first consider simulated and real data in which noise has been artificially introduced in class labels. We will then compare the performances of classifiers trained using the noisy labels, with those of classifiers trained using soft labels encoding labeling uncertainty. In Sect. 5.2, we will describe two real applications in which soft labels naturally arise.

5.1 Simulations

5.1.1 Experimental procedure

One synthetic dataset and seven real datasets are considered in this experiment. The synthetic dataset, synthetic with $n = 200$ instances was generated according to a Gaussian mixture model with $g = 3$ classes in proportions $\pi_1 = 0.45$, $\pi_2 = 0.35$ and $\pi_3 = 0.2$ in a $p = 2$-dimensional space, and

$$\mu_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \quad \mu_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \mu_3 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The real datasets are classical datasets of the literature. Their main characteristics are reported in Table 1. In these data sets, the true class label is known for each learning

Table 1 Real datasets used in Sect. 5.1

Dataset	Amount of instances	Number of classes	Dimension
Breastcancer	569	2	30
Ionosphere	350	2	34
Iris	150	3	4
Satimage	6435	6	36
Sonar	208	2	60
Vowel	990	11	10
Wine	178	3	13

instance. To study the influence of different levels of label uncertainty on classifier performance, we need to artificially introduce some uncertainty. For that purpose, we generated, for each instance, a probability p_i at random according to a beta distribution with fixed variance (set to 0.04) and expectation $\overline{p}$ set to $0.1, 0.2, \ldots, 0.9$ for the synthetic dataset, and $0.1, 0.2, \ldots, 0.7$ for the others. Then, with probability p_i, the label was replaced with a label picked at random. This process simulates the fact that, if the user only has partial information on class labels, but is forced to provide crisp labels, then some errors will inevitably be introduced.

We then generated sets of labels according to four different strategies. The first strategy is to ignore label uncertainty and consider the noisy labels as the true ones. Another approach, referred to hereafter as *adaptive semi-supervised learning*, is to discard the most uncertain labels; in our simulations, a label z_i was kept if $p_i \leq 0.5$, and the instance was considered as unlabeled otherwise. The other two strategies use soft labels. In the first case, we consider that the user knows each noise level p_i. We then *discount* [see Eq. (1)] the label information with a discount rate equal to p_i. If k^* is the observed label, we thus set $pl_{ik^*} = 1$ and $pl_{i\ell} = p_i$ for all $\ell \neq k^*$. The second strategy for generating soft labels is similar, except that each p_i is replaced by the average noise level $\overline{p}$. Thus, in this situation, only generic knowledge about the mislabeling process is used, instead of specific information related to each instance. This strategy uses less information, but we might argue that it is much easier to set up than the previous one, which requires an expert to assess each probability p_i of a label being erroneous.

For the synthetic dataset, we used a test set of size 1000 from the same distribution as the training data. For the real datasets, the classifiers were trained using 2/3 of the instances picked at random, the remaining ones being used for testing. Two classifiers were considered: LDA and LR. When fitting LDA to the data, the "perfect" model (i.e., obtained from the data without corrupted labels) was first determined, by considering ten different starting points (the initial proportions were set to $1/g$, the initial covariance matrices to the identity matrix, and the expectations vectors were sampled at random in the training set) and retaining the best solution (i.e., the one with the highest log-likelihood). The corresponding "actual" parameters were then used as starting values for the four LDA corresponding to the various labeling strategies mentioned above. For logistic regression, the initial parameter matrix was set to the $(p + 1) \times (g - 1)$ null matrix, and we used quadratic penalization with coefficient $\lambda = 0.01$. For each dataset, this whole procedure (training and test instance selection, introduction of label noise, model training, and model testing) was repeated 50 times, and average error rates and 95% confidence intervals were computed.

5.1.2 Results and discussion

Overall, taking into account label uncertainty almost always significantly improves classification accuracy, as compared to noisy labels or the semi-supervised strategy. This is particularly true for LDA, where the benefit of using soft labels is critical when the level of label noise is high. For a majority of datasets, soft labels based on the average mislabeling rate are almost as effective as those based on individual mislabeling rates (see, e.g., Figs. 2a, b, 3a). For some datasets, however, there is a

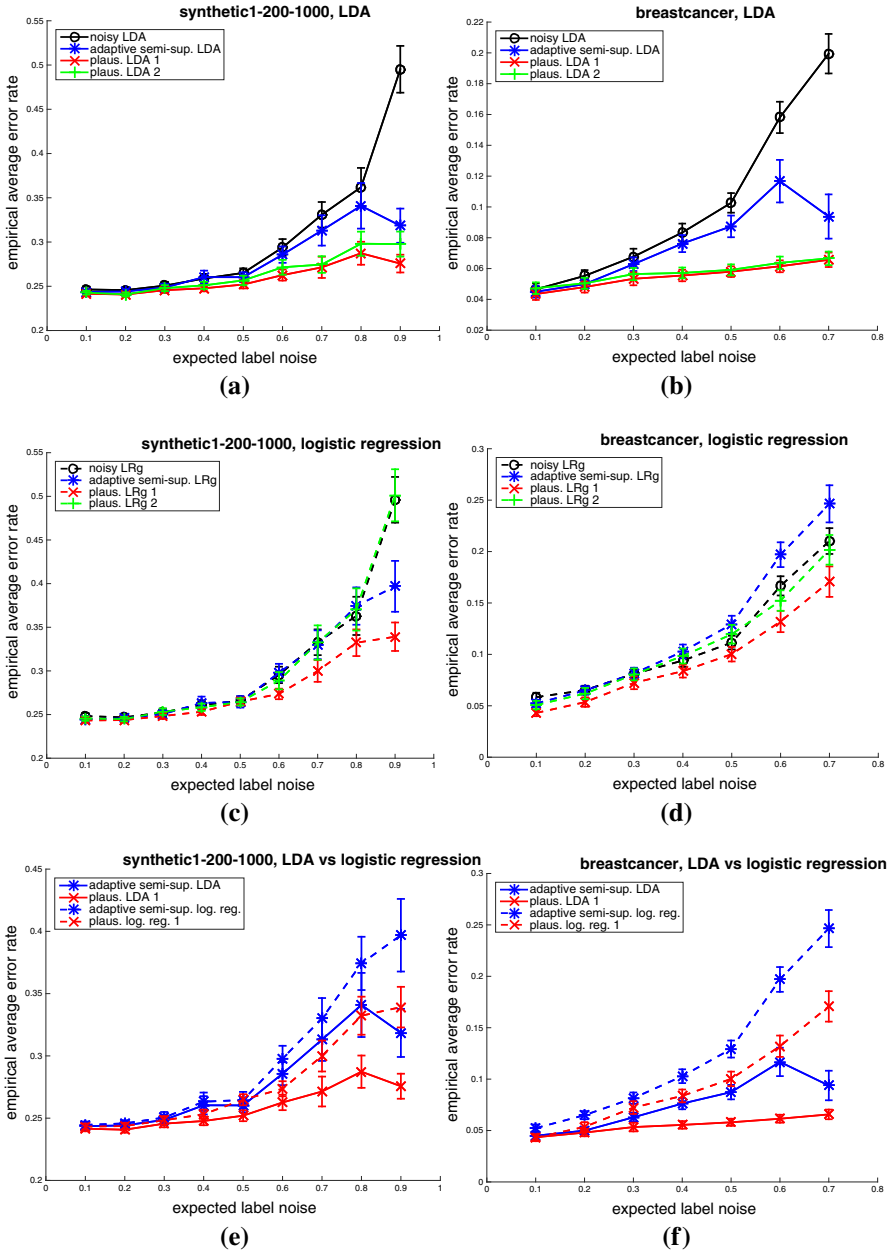

Fig. 2 synthetic (**a**, **c**, **e**) and breastcancer (**b**, **d**, **f**) data. Top: LDA, middle: LR, bottom: LDA versus LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i, green/+: soft labels based on mean noise level $\overline{p}$; plain lines: LDA, dashed lines: logistic regression

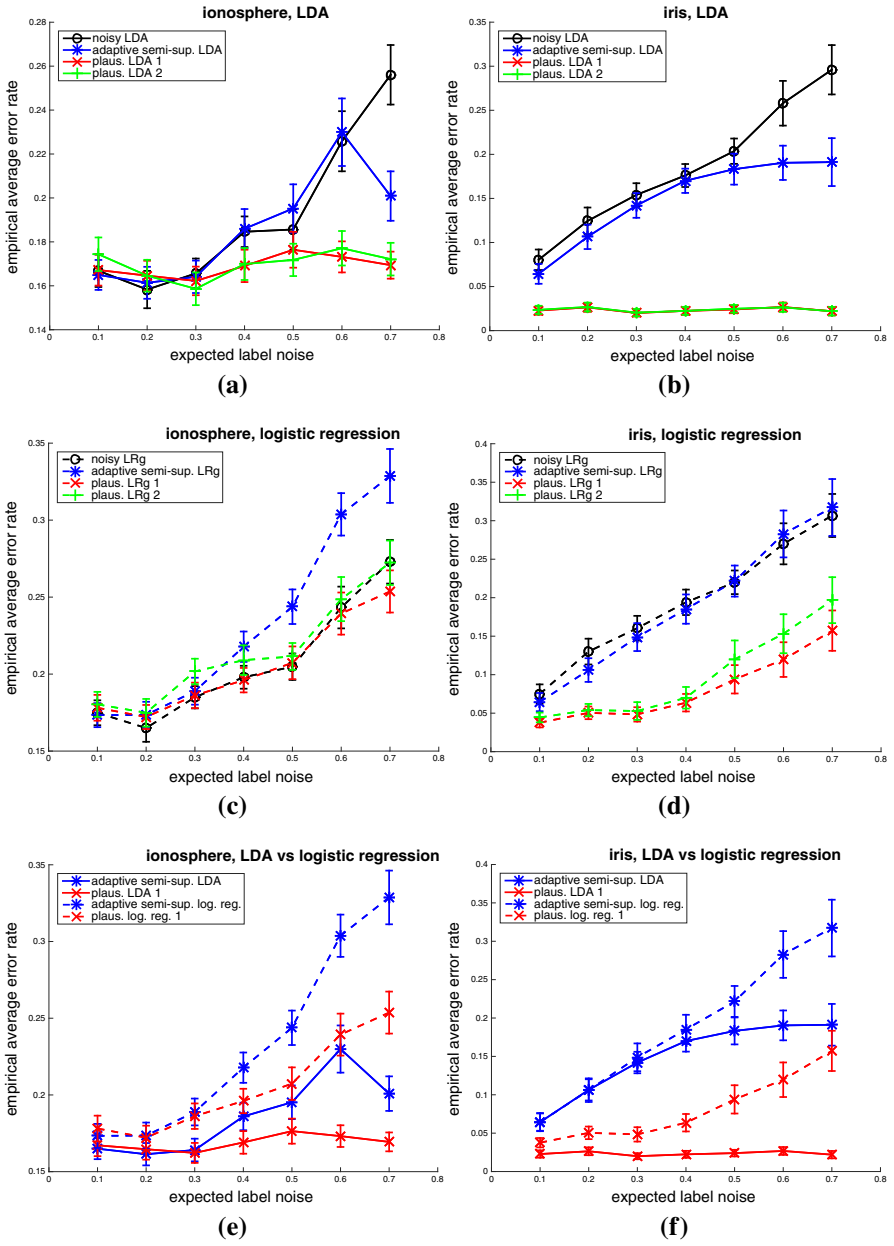

Fig. 3 ionosphere (**a**, **c**, **e**) and iris (**b**, **d**, **f**) data. Top: LDA, middle: LR, bottom: LDA versus LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i, green/+: soft labels based on mean noise level $\overline{p}$; plain lines: LDA, dashed lines: logistic regression

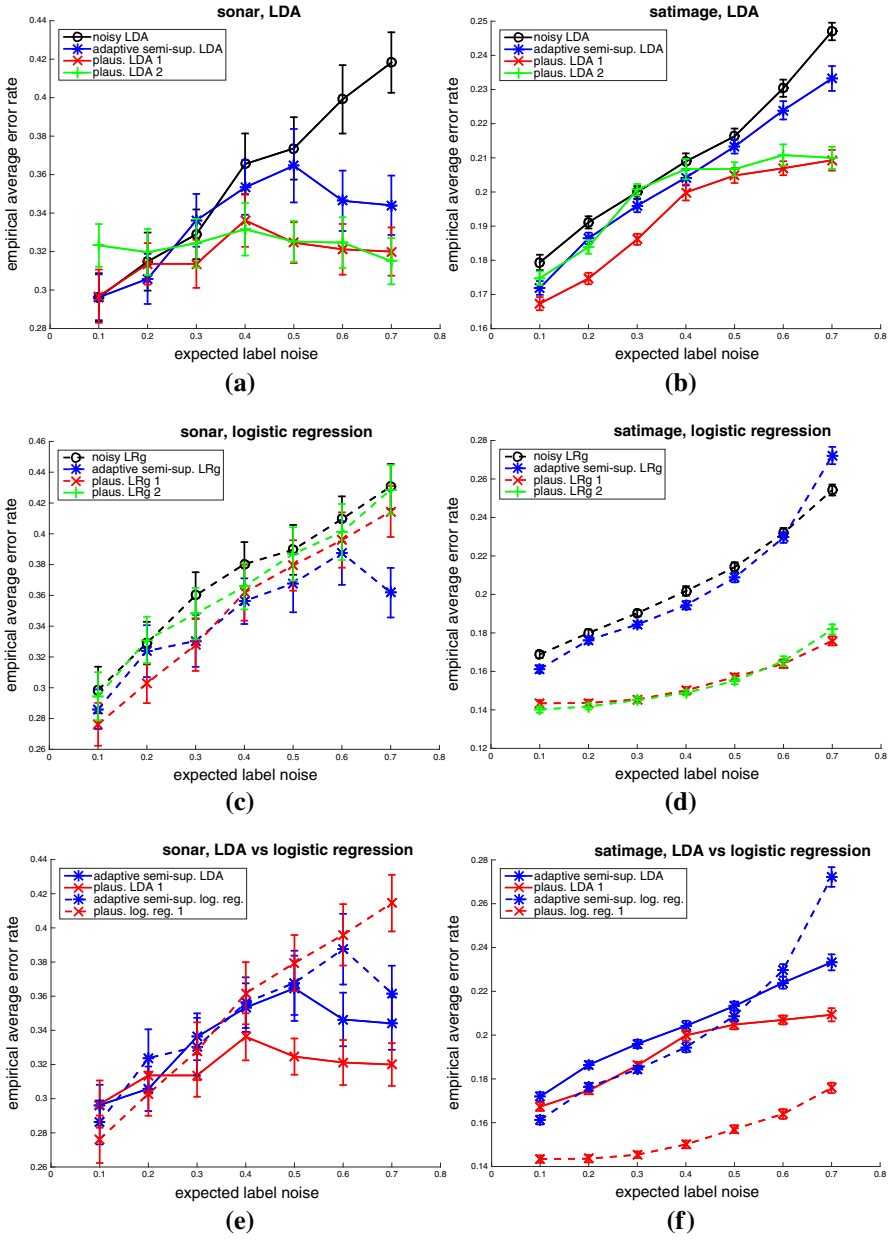

Fig. 4 sonar (**a**, **c**, **e**) and satimage (**b**, **d**, **f**) data. Top: LDA, middle: LR, bottom: LDA versus LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i, green/+: soft labels based on mean noise level $\overline{p}$; plain lines: LDA, dashed lines: logistic regression

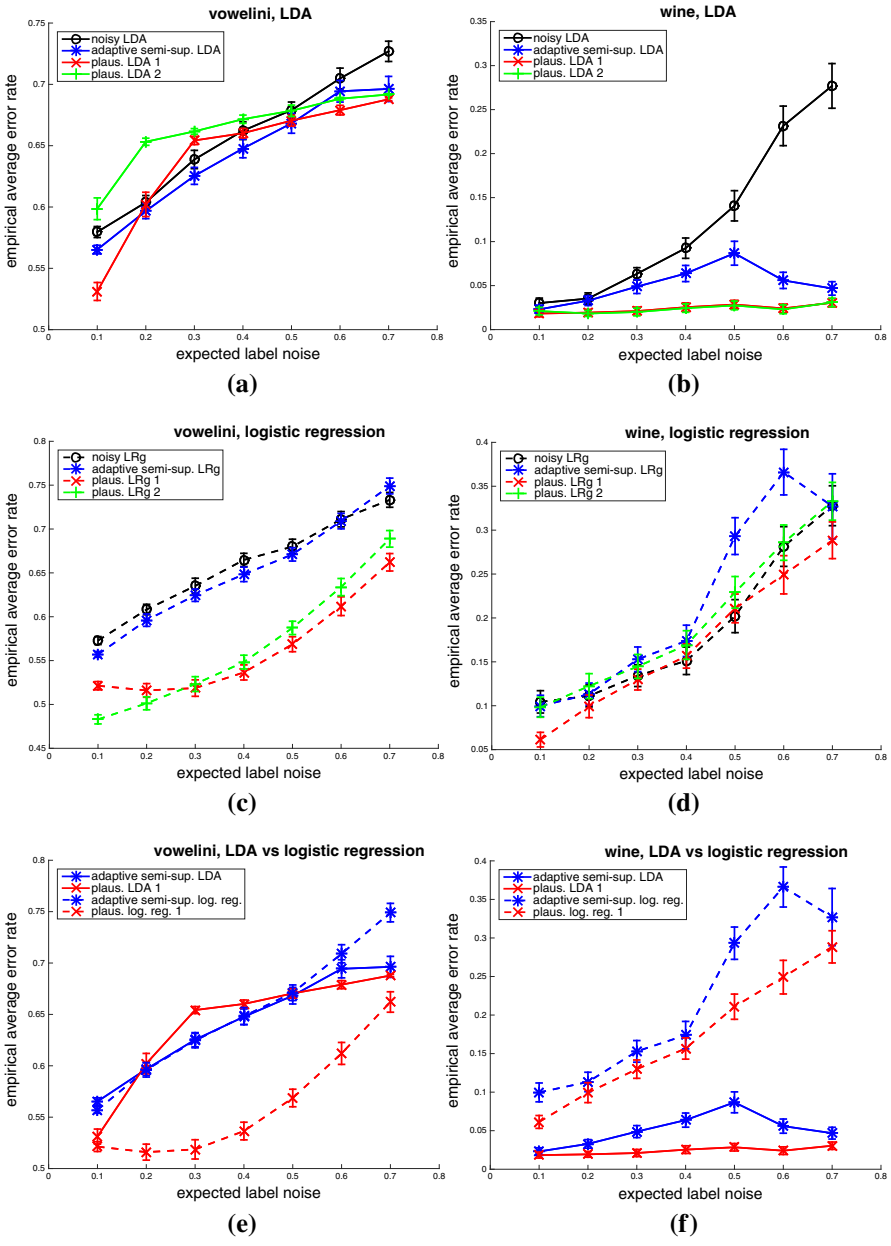

Fig. 5 vowel (**a**, **c**, **e**) and wine (**b**, **d**, **f**) data. Top: LDA, middle: LR, bottom: LDA versus LR. Black/o: noisy labels, blue/*: adaptive semi-supervised learning, red/x: soft labels based on individual noise level p_i, green/+: soft labels based on mean noise level $\overline{p}$; plain lines: LDA, dashed lines: logistic regression

Table 2 Results of Mardia's test of multivariate normality to the each of the 11 classes in the Vowel dataset

Class	Skewness	χ^2-value	p value	Kurtosis	z-value	p value
1	45.9	688	1.11e−49	116	−1.153	2.49e−1
2	42.0	631	2.70e−41	110	−3.182	1.46e−3
3	46.0	690	6.06e−50	113	−2.103	3.55e−2
4	42.5	638	2.30e−42	109	−3.267	1.09e−3
5	44.3	665	3.48e−46	114	−1.979	4.78e−2
6	47.2	707	1.53e−52	118	−0.588	5.56e−1
7	44.0	660	1.65e−45	109	−3.517	4.36e−4
8	36.0	539	7.91e−29	108	−3.608	3.09e−4
9	29.1	437	1.76e−16	109	−3.256	1.13e−3
10	44.4	666	2.52e−46	113	−2.296	2.17e−2
11	36.8	552	1.89e−30	115	−1.529	1.26e−1

The statistics are: Mardia's multivariate skewness statistic with its χ^2 and p values, and Mardia's multivariate kurtosis statistic with its z and p values

significant difference between both types of soft labels, the more informative one allowing for better performances (see Figs. 4b, 5a).

The benefit of using soft labels together with LR also exists but it is less important than it is with LDA. Generally, soft labels make it possible to improve classification accuracy, but the difference with the adaptive semi-supervised strategy—and sometimes even noisy labels—is less remarkable, especially when mislabeling probability is high. For most datasets, LR and LDA have similar performances with low label noise, but LDA becomes superior when mislabeling probability increases. This observation is consistent with our initial intuition exposed in Sect. 1, since a discriminative classifier such as LR cannot handle unsupervised learning, in contrast with LDA, which is based on a generative model. Two noticeable exceptions are the Satimage and Vowel datasets, for which LR performs critically better than LDA, with both sets of soft labels (Figs. 4f, 5e). This result may be explained by a strong departure from normality for these two datasets. To check this assumption, we applied Mardia's test of multivariate normality (Mardia 1970) to both datasets. The results, reported in Tables 2 and 3, show that the departure from normality is, indeed, highly significant in both cases. Overall, LDA tends to outperform LR for a majority of datasets, especially when soft labels are highly uncertain. However, LDA can also perform poorly in some cases. In real applications, it is thus necessary to run both classifiers and select the best model using, e.g., cross-validation and the performance measures proposed in Sect. 4.

5.2 Real applications

In this section, we present two real applications where class labels are elicited from human subjects or "experts". The first application described in Sect. 5.2.1 concerns the detection of K-complexes in EEG data. In this application, the ground truth cannot be

Table 3 Results of Mardia's test of multivariate normality to the each of the 6 classes in the Satimage dataset

Class	Skewness	χ^2-value	p value	Kurtosis	z-value	p value
1	344	87861	0	2032	248.4	0
2	227	26561	0	1644	69.9	0
3	363	82257	0	2020	229.5	0
4	321	33475	0	1769	96.0	0
5	181	21357	0	1613	62.4	0
6	210	52795	0	1774	150.5	0

The statistics are: Mardia's multivariate skewness statistic with its χ^2 and p values, and Mardia's multivariate kurtosis statistic with its z and p values

determined objectively and we need to have the data labeled by physicians. Uncertainty is reflected by disagreement between experts. The second application, described in Sect. 5.2.2, is about expression recognition in facial images. The expression on a face cannot always be determined unambiguously by human subjects. Five subjects were asked to rate the plausibility of each of the six basic expressions for each image, and the resulting credal assessments were combined by Dempster's rule. In each of the two applications, we will compare the performances of LDA and those of LR, both with soft labels and with crisp labels ignoring labeling uncertainty.

5.2.1 K-complex detection in EEG sleep data

The methods introduced in this paper were applied to the problem of detecting K-complexes in sleep EEG, using the data described in Richard (1998), Richard and Lengellé (1999). The K-complex is a transient EEG pattern which plays a major role in sleep stage assessment. It has a duration of 500–1500 ms, and is characterized by a sharp upward wave followed by a downward one. Its amplitude is three times background activity (Richard and Lengellé 1999). The discrimination of K-complexes from background activity is generally considered as a very complex pattern recognition problem.

The data used in this experiment consisted of 1178 EEG signals encoded as 64-dimensional patterns. Some of these signals were negative examples containing paroxysmal delta bursts, a phenomenon bearing some resemblance to K-complexes. The other signals consisted of patterns which, after visual inspection by five physicians, had been classified as containing a K-complex by at least one of them. Among these examples, those categorized in the K-complex class by a majority of experts were considered as positive examples, the others as negative ones. Each example (positive or negative) was then assigned a soft label consisting of a Bayesian mass function m_i such that

$$m_i(\{1\}) = k_i/5, \quad m_i(\{0\}) = 1 - k_i/5, \tag{31}$$

where 1 and 0 represent, respectively, the positive (K-complex) and negative (delta wave) class, and k_i denotes the number of experts who classified the pattern as positive.

Table 4 Sleep data: number of instances n_k classified in the positive class by k experts

k	0	1	2	3	4	5
n_k	397	340	178	116	88	59

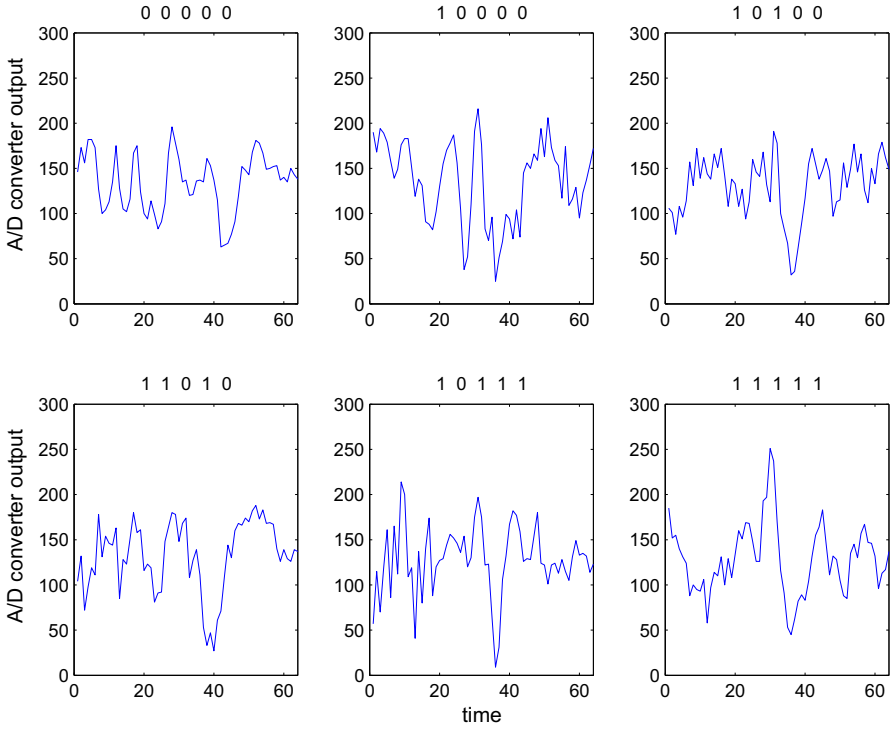

Fig. 6 Six examples of EEG signals. The figures above each plot indicates the classification given of the five experts (1 for K-complex and 0 for delta-wave)

The numbers of cases for each value of k_i are shown in Table 4. In total, $116 + 88 + 59 = 263$ instances (22.33%) were classified as positive by a majority of experts, while the remaining 915 instances (77.67%) are almost certainly negative, or were not recognized as positive by a majority of experts. Examples of signals with their classification by the five physicians are shown in Fig. 6.

Both LR and LDA were applied to these data. To reduce the input dimension, Principal Component Analysis (PCA) was first used as a preprocessing step, and the number of components was varied between 1 and 20. The LR and LDA classifiers were trained using three different sets of labels:

1. Soft labels (31), taking into account the proportion of experts in favor of each class;
2. Crisp labels, corresponding to the majority decision;

3. "Semi-supervised labels": instances classified as positive by two or three experts were considered as ambiguous and were labeled by the vacuous mass function $m_?$; the other instances were labeled unambiguously according to the majority class.

The results were evaluated using two measures: the expected loss (30), and the error rate, assuming the majority class to be the true class. We note that, as mass functions m_i are Bayesian, the lower and upper expected losses are equal in this case. Again, since the data are randomly separated into training and test instances, 10-fold cross-validation was used in order to compute average error rates. The mean cross-validation expected loss and error rate with corresponding 95% confidence intervals are represented as functions of the number of principal components for the LR and LDA with the three sets of labels in Fig. 7. For LR, training the classifier with soft labels clearly improves the performances, according to both criteria (Fig. 7a, b). For LDA, this is true for 5–8 principal components, with non significant differences for smaller or larger numbers of components (Fig. 7c, d). For both methods, the best results are obtained with 12 principal components. Figure 8 shows boxplots of expected losses and error rates for LR and LDA with 12 principal components, trained using with the three label sets. Logistic regression outperforms LDA for this dataset, and the best results are obtained when training the classifier with soft labels, whatever the performance measure.

For this dataset, LR was thus clearly able to exploit the information contained in soft labels to construct a better classifier. This phenomenon can be explained by the fact that ambiguous patterns, which have a great chance of being mislabeled, have less informative soft labels. Their influence on the final parameter estimates is thus downweighted by the E^2M algorithm.

5.2.2 Facial expression recognition

As another application of partially supervised learning with soft labels, we considered the task of recognizing facial expressions from face images. Here, soft labels may arise from the difficulty of labeling faces with a single expression unambiguously. We considered 216 images data from the CMU-Pittsburgh image database (Kanade et al. 2000), with 36 images for each of the six basic expressions (see Fig. 9). Following Dubuisson et al. (2002), we used aligned and cropped versions of these images of size 60×70, corresponding to vectors of 4200 pixels.

The database was split into a learning set and a test set of 108 images each (with 18 images per expression). Images in the learning set were viewed by five subjects, who assigned a soft label to each image by assessing plausibilities for each of the six expressions. Figure 10 shows two examples of images with the corresponding plausibility assessments, for which there was some disagreement between assessors. For instance, for the image of Fig. 10a, one subject considered that it expresses disgust for sure (white horizontal bar), while others hesitated between anger and disgust, or between disgust and sadness. The assessments from the five experts were discounted (Shafer 1976) and combined by Dempster's rule using (4). For image i, the unnormalized contour function pl_i^* was thus

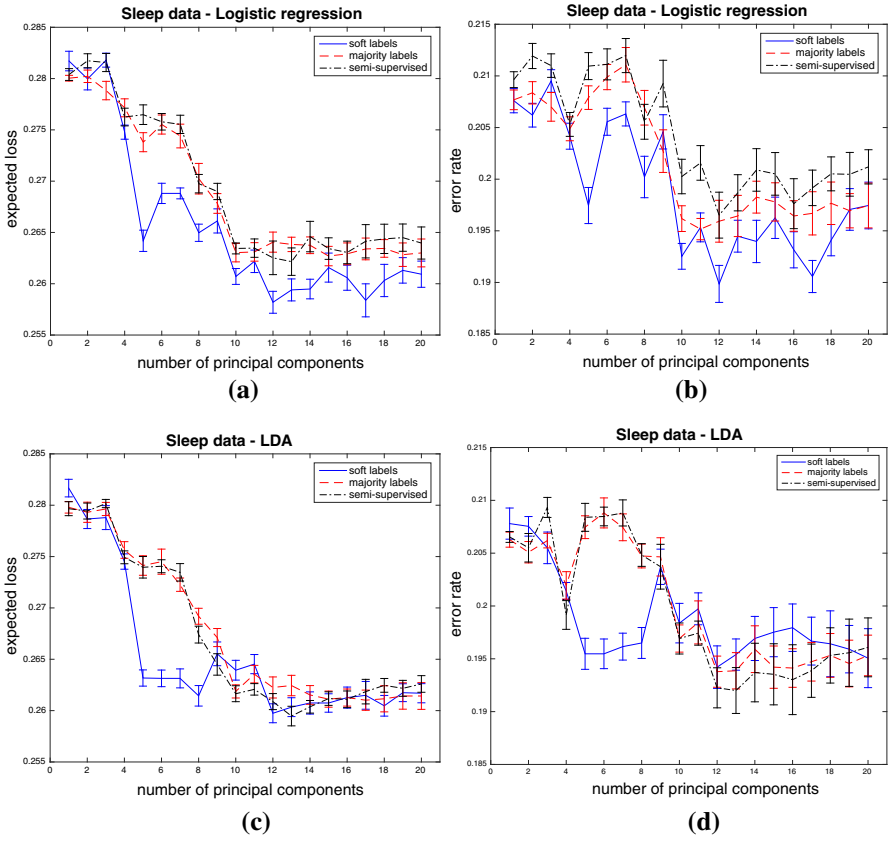

Fig. 7 Performances of LR (**a**, **b**) and LDA (**c**, **d**) for the sleep data, as a function of the number of principal components. The criteria are the expected loss (**a**, **c**) and the error rate taking the majority labels as ground truth (**b**, **d**). In each graph, the three curves corresponds to three sets of labels: soft labels, majority labels, and "semi-supervised" labels (see details in text)

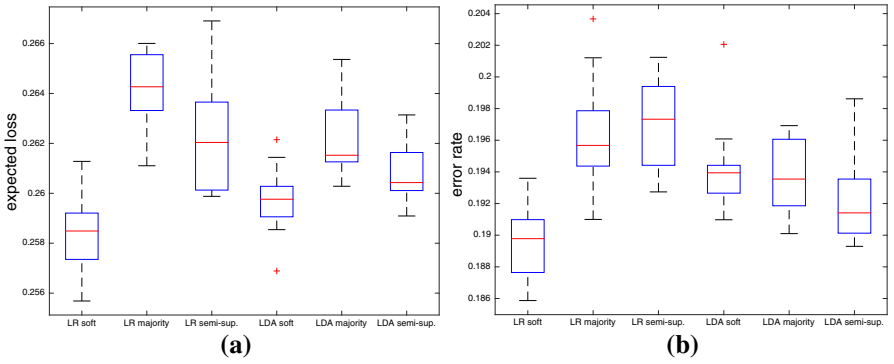

Fig. 8 Expected losses (**a**) and error rates (**b**) for classifiers trained with 12 principal components as input (sleep data). The classifiers are, from left to right: LR with soft labels, majority labels and semi-supervised labels; LDA with soft labels, majority labels and semi-supervised labels

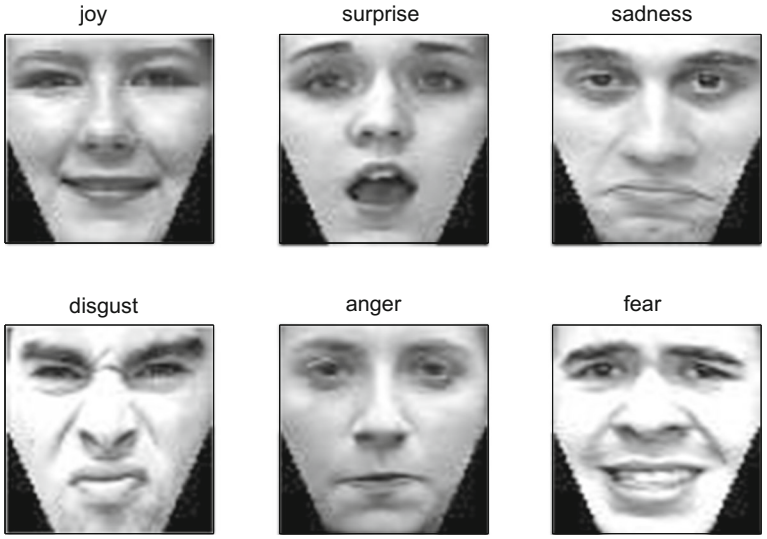

Fig. 9 Examples of each of the six basic expressions

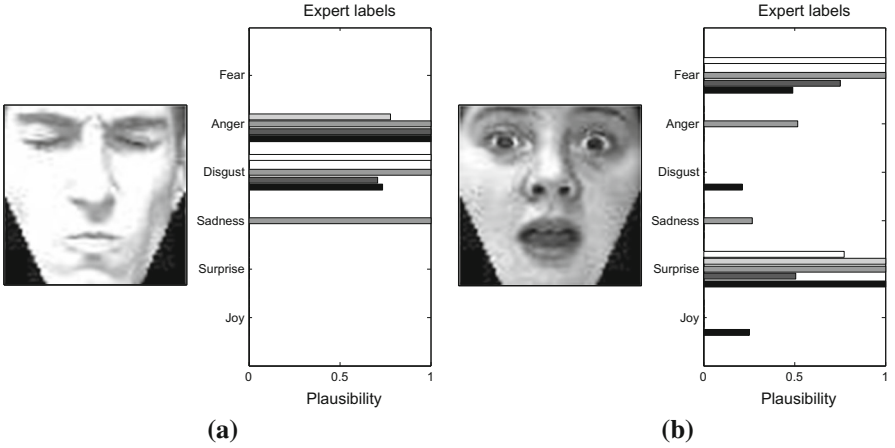

Fig. 10 Two examples of plausibility assessments provided by the five subjects. The plausibility values are shown as horizontal bars next to each of the two images, with one gray level for each subject

$$pl_i^*(k) = \prod_{j=1}^{5} \left[0.1 + 0.9\, pl_{ij}(\omega_k)\right], \tag{32}$$

where $pl_{ij}(k)$ is the plausibility of expression k for image i assessed by subject j. The plausibilities were then normalized to ensure that $\max_k pl_i(k) = 1$. Figure 11 shows two examples of individual and combined plausibility assessments. We can remark that combination rule (32) eliminates alternatives considered as impossible by most

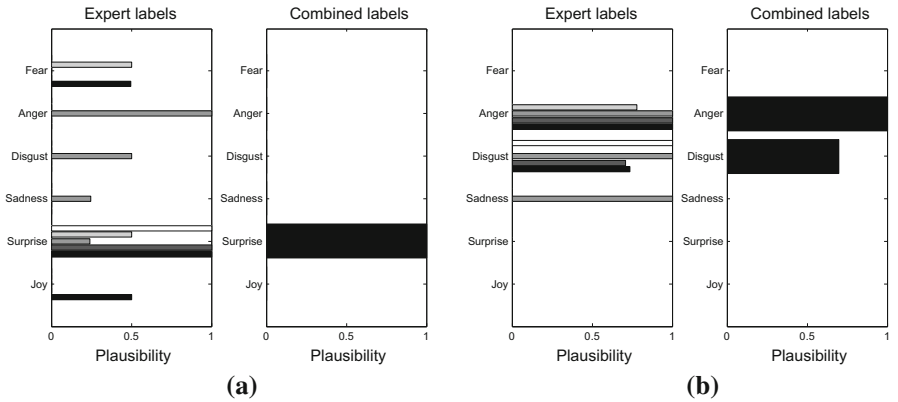

Fig. 11 Two examples of individual and combined plausibility assessments

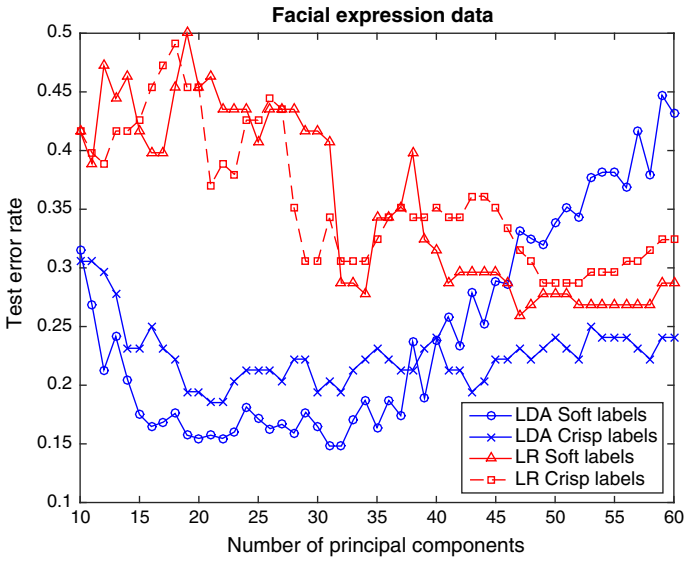

Fig. 12 Facial expression data: test error rates versus number of principal components for LDA and LR, with soft and crisp labels

subjects, even if they were considered as completely plausible by one assessor, such as anger in Fig. 11a or sadness in Fig. 11b.

As in Sect. 5.2.1, PCA was used to reduce the dimension of the data. Logistic regression and LDA were then applied to the data with different numbers of principal components, and two sets of labels: the combined soft labels, and crisp labels corresponding to the most plausible expression. For logistic regression, we used quadratic penalization with coefficient $\lambda = 10^{-4}$. The results are reported in Fig. 12, which shows the test error rate as a function of the number of components, for each of the two classifiers and the two sets of labels. For this dataset, LDA clearly outperforms LR, with the lowest error rate obtained for around 30 principal components. For LDA,

the best results are obtained with the soft labels, which can be explained by the smaller importance given to ambiguous cases. However, the LDA classifier trained with soft labels is more prone to overfitting, as shown by the sharp increase of error rate with the input dimension. In contrast, LR does not work too well for this data, with a lowest error rate almost twice as large as that obtained by LDA. For LR, there also seems to be some advantage of using soft labels for input dimensions larger than 40, but this advantage is not so clear as it is for LDA.

This second experiment using a real data set with soft labels provides additional evidence of the potential gain of using soft labels and confirms similar findings reported in Cherfi et al. (2012) or Ramasso and Denœux (2013), for instance. As with the sleep data analyzed in Sect. 5.2.1, this gain of performance can be explained by the smaller influence of potentially mislabeled data on the final classifier.

6 Conclusions

Whereas a lot of methods exist for supervised or unsupervised classification, we have seen in recent years a growing interest for more general forms of learning tasks, in which information about the class of learning instances is partial. In this paper, we have considered a very general framework for addressing such problems, which consists in representing partial class information by Dempster–Shafer mass functions constituting soft labels. Two types of linear classifiers have been considered: LDA based on a generative model, and LR based on a discriminative model. Both classifiers can be trained in a partially supervised mode by maximizing an evidential likelihood function, which can be done efficiently using the evidential EM algorithm.

As a generative model represents the joint distribution of the input vector and the response variable, it can be trained in an unsupervised setting; consequently, it can be expected to perform well with scant class information. In contrast, a discriminative model relies on weaker assumptions and can be expected to perform better than a generative model when the model is inadequate, provided sufficient class information is available. Our experimental results support these assumptions. Using artificial and real data for which class labels have been perturbed by noise, we have shown that soft labels taking into account labelling uncertainty can be successfully exploited to achieve significantly lower error rates, as compared to precise but noisy labels. Both LDA and LR classifiers are able to exploit partial information contained in soft labels. However, LDA usually performs better, especially when the mislabeling probability is high, in which case the soft labels become very imprecise and learning becomes almost unsupervised. However, LR outperforms LDA for some datasets. It is thus necessary to consider several models and select the best ones, which can also be done in a partially supervised setting using lower and upper expected losses.

We have also considered two real applications in which soft labels naturally arise from the absence of ground truth class information: detection of K-complex in EEG sleep data and recognition of facial expressions in images. In these two cases, soft labels have been shown to allow for better performances, as compared to using the most plausible label for each instance and ignoring labeling uncertainty. Whereas large amounts of unlabeled data (e.g., images) can be retrieved from the internet or

can be generated by sensors, reliably labeled data remain relatively scarce because of the cost of visually inspecting the data to determine their class, when no ground truth is available. In this context, our results suggest that partial class information can be useful, provided labeling uncertainty is suitably represented in the form of soft labels.

Although linear classifiers have been exclusively considered in this study for simplicity, there is obviously no conceptual difficulty with extending our approach to non linear parametric classifiers such as, e.g., quadratic or mixture discriminant analysis, generalized additive models, kernel logistic regression, etc. More fundamentally, other likelihood-based approaches to inference from uncertain data, such has proposed in Hüllermeier (2014) or Couso and Dubois (2017), could be considered and compared with our method. Finally, the elicitation of soft labels from single or multiple experts is also an important topic that remains to be thoroughly investigated.

References

Abassi L, Boukhris I (2016) Crowd label aggregation under a belief function framework. In: Lehner F, Fteimi N (eds) Proceedings of 9th international conference on knowledge science, engineering and management, KSEM 2016, Passau, Germany, 5–7 Oct 2016. Springer, Cham, pp 185–196

Bishop CM (2006) Pattern recognition and machine learning. Springer, Berlin

Chapelle O, Schölkopf B, Zien A (2006) Semi-supervised learning. MIT Press, Cambridge

Cherfi ZL, Oukhellou L, Côme E, Denœux T, Aknin P (2012) Partially supervised independent factor analysis using soft labels elicited from multiple experts: application to railway track circuit diagnosis. Soft Comput 16(5):741–754

Côme E, Oukhellou L, Denœux T, Aknin P (2009) Learning from partially supervised data using mixture models and belief functions. Patt Recognit 42(3):334–348

Cour T, Sapp B, Taskar B (2011) Learning from partial labels. J Mach Learn Res 12:1225–1261

Couso I, Dubois D (2017) Maximum likelihood under incomplete information: toward a comparison of criteria. In: Ferraro MB, Giordani P, Vantaggi B, Gagolewski M, Gil M Ángeles, Grzegorzewski P, Hryniewicz O (eds) Soft methods for data science. Springer, Cham, pp 141–148

Dempster AP (1967) Upper and lower probabilities induced by a multivalued mapping. Ann Math Stat 38:325–339

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Denœux T (1995) A *k*-nearest neighbor classification rule based on Dempster–Shafer theory. IEEE Trans Syst Man Cybern 25(05):804–813

Denœux T (2013) Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans Knowl Data Eng 25(1):119–130

Denœux T (2014) Likelihood-based belief function: justification and some extensions to low-quality data. Int J Approx Reason 55(7):1535–1547

Denoeux T, Kanjanatarakul O (2016) Beyond fuzzy, possibilistic and rough: an investigation of belief functions in clustering. In: Proceedings of the 8th international conference on soft methods in probability and statistics SMPS 2016, soft methods for data science, advances in intelligent and soft computing, AISC, vol 456. Springer, Rome, Italy, pp 157–164

Denœux T, Masson MH (2004) EVCLUS: evidential clustering of proximity data. IEEE Trans Syst Man Cybern B 34(1):95–109

Denœux T, Skarstein-Bjanger M (2000) Induction of decision trees for partially classified data. In: Proceedings of SMC'2000. IEEE, Nashville, TN, pp 2923–2928

Denœux T, Zouhal LM (2001) Handling possibilistic labels in pattern classification using evidential reasoning. Fuzzy Sets Syst 122(3):47–62

Denœux T, Sriboonchitta S, Kanjanatarakul O (2016) Evidential clustering of large dissimilarity data. Knowl Based Syst 106:179–195

Dubuisson S, Davoine F, Masson MH (2002) A solution for facial expression representation and recognition. Signal Process Image Commun 17(9):657–673

Elouedi Z, Mellouli K, Smets P (2001) Belief decision trees: theoretical foundations. Int J Approx Reason 28:91–124
Hasan A, Wang Z, Mahani A (2016) Fast estimation of multinomial logit models: R package mnlogit. J Stat Softw 75(1):1–24
Heitjan DF, Rubin DB (1991) Ignorability and coarse data. Ann Stat 19(4):2244–2253
Hüllermeier E (2014) Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int J Approx Reason 55(7):1519–1534
Hüllermeier E, Beringer J (2005) Learning from ambiguously labeled examples. In: Proceedings of the 6th international symposium on intelligent data analysis (IDA-05), Madrid, Spain
Jaffray JY (1989) Linear utility theory for belief functions. Oper Res Lett 8(2):107–112
Kanade T, Cohn J, Tian Y (2000) Comprehensive database for facial expression analysis. In: Proceedings of the fourth international conference of face and gesture recognition, Grenoble, France, pp 46–53
Li J (2013) Logistic regression. Course notes. http://sites.stat.psu.edu/~jiali/course/stat597e/notes2/logit.pdf
Liu ZG, Pan Q, Dezert J, Mercier G (2015) Credal c-means clustering method based on belief functions. Knowl Based Syst 74:119–132
Liu ZG, Pan Q, Dezert J, Mercier G (2017) Hybrid classification system for uncertain data. IEEE Trans Syst Man Cybern Syst (**in press**). https://doi.org/10.1109/TSMC.2016.2622247
Ma L, Destercke S, Wang Y (2016) Online active learning of decision trees with evidential data. Patt Recognit 52:33–45
Mardia KV (1970) Measures of multivariate skewness and kurtosis with applications. Biometrika 57(3):519–530
McLachlan GJ, Krishnan T (1997) The EM algorithm and extensions. Wiley, New York
McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
Nguyen N, Caruana R (2008) Classification with partial labels. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '08. ACM, New York, NY, USA, pp 551–559
Peters G, Crespo F, Lingras P, Weber R (2013) Soft clustering: fuzzy and rough approaches and their extensions and derivatives. Int J Approx Reason 54(2):307–322
Press SJ, Wilson S (1978) Choosing between logistic regression and discriminant analysis. J Am Stat Assoc 73(364):699–705
Quost B (2014) Logistic regression of soft labeled instances via the evidential EM algorithm. In: Cuzzolin F (ed) Proceedings of the third international conference on belief functions: theory and applications, BELIEF 2014. Oxford, UK, 26–28 Sept 2014. Springer, Cham, pp 77–86
Quost B, Denoeux T (2016) Clustering and classification of fuzzy data using the fuzzy EM algorithm. Fuzzy Sets Syst 286:134–156
Ramasso E, Denœux T (2013) Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. IEEE Trans Fuzzy Syst 21(6):1–11
Richard C (1998) Une méthodologie pour la détection à structure imposée. applications au plan temps-fréquence. Ph.D. thesis, Université de Technologie de Compiègne
Richard C, Lengellé R (1999) Data driven design and complexity control of time-frequency detectors. Sig Process 77:37–48
Rjab AB, Kharoune M, Miklos Z, Martin A (2016) Characterization of experts in crowdsourcing platforms. In: Vejnarová J, Kratochvíl V (eds) Proceedings of 4th international conference on belief functions: theory and applications, BELIEF 2016, Prague, Czech Republic, 21–23 Sept 2016. Springer, Cham, pp 97–104
Shafer G (1976) A mathematical theory of evidence. Princeton University Press, Princeton
Strat TM (1990) Decision analysis using belief functions. Int J Approx Reason 4(5–6):391–417
Sutton-Charani N, Destercke S, Denoeux T (2013) Learning decision trees from uncertain data with an evidential EM approach. In: 12th international conference on machine learning and applications, 2013, vol 1, pp 111–116
Sutton-Charani N, Destercke S, Denœux T (2014) Training and evaluating classifiers from evidential data: application to E2M decision tree pruning. In: Cuzzolin F (ed) Proceedings of the third international conference on belief functions: theory and applications, BELIEF 2014. Oxford, UK, 26–28 Sept 2014. Springer, Cham, pp 87–94
Trabelsi S, Elouedi Z, Mellouli K (2007) Pruning belief decision tree methods in averaging and conjunctive approaches. Int J Approx Reason 46(3):568–595

Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets Syst 1:3–28
Zhou K, Martin A, Pan Q (2014) Evidential-EM algorithm applied to progressively censored observations. In: Laurent A, Strauss O, Bouchon-Meunier B, Yager RR (eds) Proceedings of 15th international conference on information processing and management of uncertainty in knowledge-based systems, IPMU 2014, Montpellier, France, Part III, 15–19 July 2014. Springer, Cham, pp 180–189