**Soft Computing manuscript No.**
(will be inserted by the editor)

# Evidential joint calibration of binary SVM classifiers

**Pauline Minary**[1,2] · **Frédéric Pichon**[1] · **David Mercier**[1] · **Eric Lefevre**[1] · **Benjamin Droit**[2]

Received: date / Accepted: date

**Abstract** In order to improve overall performance with respect to a classification problem, a path of research consists in using several classifiers and to fuse their outputs. To perform this fusion, some approaches merge the classifier outputs using a rule of combination. This requires that the outputs be made comparable beforehand, which is usually done thanks to a probabilistic calibration of each classifier. The fusion can also be performed by concatenating the classifier outputs into a vector and applying a joint probabilistic calibration to this vector. Recently, extensions of probabilistic calibration techniques of an individual classifier have been proposed using evidence theory, in order to better represent the uncertainties inherent to the calibration process. In this paper, we adapt this latter idea to joint probabilistic calibration techniques, leading to evidential versions of joint calibration techniques. In addition, our proposal was tested on generated and real datasets and the results showed that it either outperforms or is comparable to state-of-the-art approaches.

**Keywords** Belief functions · Information fusion · Evidential calibration · Classification

## 1 Introduction

An important path of research in classification consists in using several classifiers, which are trained with different data or based on different training models, instead of relying on a single one (Kuncheva 2004). Since they do not necessarily give the same output after observing a given object, a central issue in this approach consists in figuring out how to exploit these outputs to classify this object.

There are different ways of performing the fusion of some classifier outputs (Kuncheva 2004; Tulyakov et al. 2008). These various fusion methods are usually separated into two categories: the non-trainable and trainable combiners.

In the first category, the outputs returned by the classifiers after observing a given object are combined using a predetermined rule of combination. As the used classifiers are different, their outputs are not scaled with respect to each other, and thus have to be made comparable before being combined. A step called calibration (Platt 1999) is thus usually performed to transform each output into a probability. In particular, the

This paper is an extended and revised version of (Minary et al. 2017).

Pauline Minary
pauline.minary@reseau.sncf.fr

Frédéric Pichon
frederic.pichon@univ-artois.fr

David Mercier
david.mercier@univ-artois.fr

Eric Lefevre
eric.lefevre@univ-artois.fr

Benjamin Droit
benjamin.droit@reseau.sncf.fr

[1]Univ. Artois, EA 3926,
Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
Béthune, F-62400, France.

[2]SNCF Réseau,
Département des Télécommunications,
La Plaine Saint Denis, France.

three calibration techniques the most commonly used are based on binning (Zadrozny and Elkan 2001), isotonic regression (Zadrozny and Elkan 2002) and logistic regression (Platt 1999). These calibration techniques suffer from an over-fitting problem, especially when only few training data are available. Within this scope, Xu et al. (2016) recently proposed a refinement of the main calibration procedures within a framework for reasoning under uncertainty called evidence theory (Shafer 1976; Smets and Kennes 1994). This theory allows Xu *et al.* to model more precisely the uncertainties inherent to such calibration process and thus to prevent the over-fitting issue. Xu et al. (2016) used this refinement to propose an approach of the non-trainable kind for binary classification problems. This latter approach consists in: using several SVM classifiers returning confidence scores, calibrating each of the returned scores using an evidential calibration technique, hence transforming each of the score into a belief function, and finally merging them using Dempster's rule of combination (Shafer 1976).

The second category regroups the approaches using the concatenation of the outputs of the classifiers as an input vector for another classifier. In particular, the approach defined in (Zhong and Kwok 2013) is a member of that category as a vector of scores obtained from an ensemble of classifiers is provided as an input vector to a probabilistic classifier based on multiple isotonic regression. Note that such kind of approach may be regarded as a probabilistic *joint* calibration as it learns how to convert a vector of scores into a probability, that is it calibrates jointly the classifiers. In addition, as logistic regression can also be defined with multiple inputs (Hosmer et al. 2013), one may envisage to extend this kind of approach to the logistic model.

Both categories present some disadvantages. As already mentioned, the calibration techniques used in the non-trainable combiners are prone to an over-fitting problem. In addition, non-trainable combiners rely on a fixed rule of combination; as explained by Duin (2002), a predetermined rule may be the best combination only under very strict conditions, and an improved result may be obtained using an approach of the trainable combiner category. For the trainable combiners, a training set common to all classifiers is required, and the combiner must be re-learned each time a new classifier is added to the system. Furthermore, trainable combiner approaches corresponding to a probabilistic joint calibration may also be prone to the over-fitting problem inherent to probabilistic calibration.

Within this scope, we propose in this paper to study the application of the appealing element of Xu *et al.*'s approach (Xu et al. 2016), *i.e.*, the evidential extension of calibration, to joint calibration techniques. As a result, we obtain methods that transform the vector of scores returned by the classifiers for a given object into a belief function.

This paper is organized as follows. First, necessary background on evidence theory are recalled in Section 2. In Section 3, probabilistic calibration methods of a single classifier are presented, followed by their extension using the evidence theory. Then, probabilistic joint calibrations and their extension to the evidential framework that we propose, are exposed in Section 4. In Section 5, the proposed approach is compared experimentally to other approaches, and in particular to Xu *et al.* non-trainable combiner approach relying on evidential calibration of individual classifiers and to probabilistic joint calibration. Finally, conclusion and perspectives are given in Section 6.

## 2 Evidence theory

Basic notions of the theory of evidence (Shafer 1976; Smets and Kennes 1994) are first exposed in Section 2.1. Applications of this theory to statistical inference and prediction, which are useful to derive calibration in the evidential framework, are addressed in Section 2.2 and 2.3.

### 2.1 Basic notions

Evidence theory, also referred to as belief function theory, is a general framework for modeling uncertainty. Let $\boldsymbol{\omega}$ be a variable whose possible values belong to the finite set $\Omega = \{\omega_1, \cdots, \omega_K\}$. In this theory, uncertainty with respect to the actual value $\omega_0$ taken by $\boldsymbol{\omega}$ is represented using a *Mass Function* (MF) defined as a mapping $m^\Omega : 2^\Omega \to [0,1]$ verifying $m^\Omega(\emptyset) = 0$ and

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

The quantity $m^\Omega(A)$ corresponds to the belief committed exactly to the hypothesis $\omega_0 \in A$ and nothing more specific. Any subset $A$ of $\Omega$ such that $m^\Omega(A) > 0$ is called a focal set of $m^\Omega$. When the focal sets are nested, $m^\Omega$ is said to be consonant.

Equivalent representations of a mass function exist. In particular, the belief and plausibility functions are respectively defined by

$$Bel^\Omega(A) = \sum_{B \subseteq A} m^\Omega(B), \quad \forall A \subseteq \Omega, \tag{2}$$

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

The degree of belief $Bel^{\Omega}(A)$ measures the amount of evidence strictly in favour of the hypothesis $\omega_0 \in A$, while the plausibility $Pl^{\Omega}(A)$ is the amount of evidence not contradicting it. The plausibility function restricted to singletons is called the contour function, denoted $pl^{\Omega}$ and defined by

$$pl^{\Omega}(\omega) = Pl^{\Omega}(\{\omega\}), \quad \forall \omega \in \Omega. \tag{4}$$

When a mass function is consonant, the plausibility function can be recovered from its contour function as follows:

$$Pl^{\Omega}(A) = \sup_{\omega \in A} pl^{\Omega}(\omega), \quad \forall A \subseteq \Omega. \tag{5}$$

Given two independent MFs $m_1^{\Omega}$ and $m_2^{\Omega}$, it is possible to combine them using Dempster's rule of combination. The result of this combination is a MF $m_{1\oplus 2}^{\Omega}$ defined by

$$m_{1\oplus 2}^{\Omega}(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1^{\Omega}(B) m_2^{\Omega}(C), \qquad \forall A \neq \emptyset, \tag{6}$$

where

$$\kappa = \sum_{B \cap C = \emptyset} m_1^{\Omega}(B) m_2^{\Omega}(C), \tag{7}$$

represents the degree of conflict between $m_1^{\Omega}$ and $m_2^{\Omega}$, and $m_{1\oplus 2}^{\Omega}(\emptyset) = 0$. If $\kappa = 1$, there is a total conflict between the two pieces of evidence and they cannot be combined.

Different decision strategies exist to make a decision about the actual value $\omega_0$ of $\boldsymbol{\omega}$, given a MF $m^{\Omega}$ (Denœux 1997). In particular, the value $\omega \in \Omega$ having the smallest so-called *upper* or *lower expected costs* may be selected. The upper and lower expected costs of some value $\omega \in \Omega$, respectively denoted by $R^*(\omega)$ and $R_*(\omega)$, are defined as

$$R^*(\omega) = \sum_{A \subseteq \Omega} m^{\Omega}(A) \max_{\omega' \in A} c(\omega, \omega'), \tag{8}$$

$$R_*(\omega) = \sum_{A \subseteq \Omega} m^{\Omega}(A) \min_{\omega' \in A} c(\omega, \omega'), \tag{9}$$

where $c(\omega, \omega')$ is the cost of deciding $\omega$ when the true answer is $\omega'$. When the set of focal elements is reduced to singletons and $\Omega$, and when the costs are taken equal to 0 if $\omega = \omega'$ and 1 otherwise, the upper and lower expected costs are, respectively, defined as

$$\begin{aligned} R^*(\omega) &= 1 - m^{\Omega}(\{\omega\}), \\ &= 1 - Bel^{\Omega}(\{\omega\}). \end{aligned} \tag{10}$$

$$\begin{aligned} R_*(\omega) &= 1 - m^{\Omega}(\{\omega\}) - m^{\Omega}(\Omega), \\ &= 1 - Pl^{\Omega}(\{\omega\}). \end{aligned} \tag{11}$$

Choosing the value $\omega$ minimizing the lower (resp. upper) expected costs is called the optimistic (resp. pessimistic) strategy.

To avoid making wrong decisions in the risky cases, *i.e.*, when the expected costs are high, a reject decision may be introduced. Formally, a reject cost $R_{rej} \in [0,1]$ is introduced and a decision to reject is made when $R_{rej}$ is lower than the other expected costs.

### 2.2 Statistical inference

The theory of evidence can be used for statistical inference. Consider $\theta \in \Theta$ an unknown parameter, $x \in \mathbb{X}$ some observed data and $f_\theta(x)$ the density function generating the data. Statistical inference consists in making statements about $\theta$ after observing the data $x$. Shafer (1976) proposed to represent knowledge about $\theta$ given $x$ by a consonant belief function $Bel_x^{\Theta}$ based on the likelihood function $L_x : \theta \to f_\theta(x)$ (see also justifications by Denœux (2014)), whose contour function is the normalized likelihood function:

$$pl_x^{\Theta}(\theta) = \frac{L_x(\theta)}{\sup_{\theta' \in \Theta} L_x(\theta')}, \qquad \forall \theta \in \Theta. \tag{12}$$

Let us consider an important particular case. Assume that we observe a random variable $X$, which has a binomial distribution with parameters $n \in \mathbb{N}$ and $\theta \in [0,1]$, *i.e.*, $X \sim \mathcal{B}(n, \theta)$. In that case, we have

$$f_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{13}$$

The likelihood-based belief function has the following contour function:

$$pl_x^{\Theta}(\theta) = \frac{\theta^x (1-\theta)^{n-x}}{\hat{\theta}^x (1-\hat{\theta})^{n-x}}, \tag{14}$$

for all $\theta \in \Theta = [0,1]$, where $\hat{\theta} = \frac{x}{n}$ is the Maximum Likelihood Estimate (MLE) of $\theta$. Figure 1 shows the contour function of the binomial distribution, with $n = 30$ and $x = 10$.

### 2.3 Forecasting

Let us now suppose that we have some knowledge about $\theta \in \Theta$ after observing some data $x$, given under a form of a consonant belief function $Bel_x^{\Theta}$. The aim of forecasting is to make statements about a not yet observed data $Y \in \mathbb{Y}$, whose conditional distribution $g_{x,\theta}(y)$ given

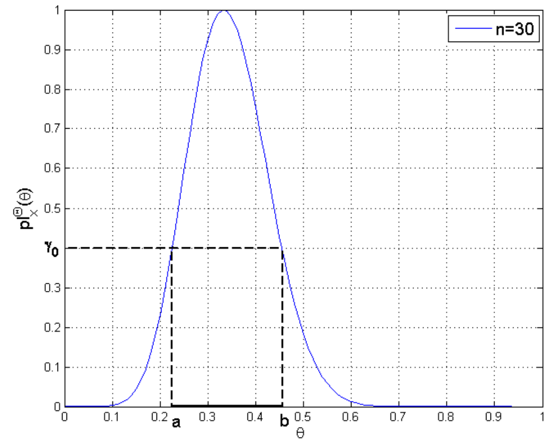


Fig. 1: Contour function of a binomial distribution, with $n = 30$ and $x = 10$.

$X = x$ depends on $\theta$. A solution to this problem, proposed by Kanjanatarakul et al. (2014, 2016), consists in using the fact that $Bel_x^\Theta$ is equivalent to a random set, and in using the sampling model of Dempster (Dempster 1966) to deduce a belief function on $\mathbb{Y}$. We detail these two points below.

Let us recall that the focal sets of $Bel_x^\Theta$ are the level sets of $pl_x^\Theta$, defined by (Nguyen 2006)

$$\Gamma_x(\gamma) = \{\theta \in \Theta | pl_x^\Theta(\theta) \geq \gamma\}, \qquad \forall \gamma \in [0,1]. \tag{15}$$

For instance in Figure 1, for $\gamma = \gamma_0 = 0.4$, the set $\Gamma_x(\gamma_0)$ is defined as the set of all values of $\theta \in \Theta$ such that $pl_x^\Theta(\theta) \geq 0.4$, *i.e.*, $\Gamma_x(\gamma_0) = [a, b] \approx [0.225, 0.454]$. Moreover, the belief function $Bel_x^\Theta$ is equivalent to the random set induced by the Lebesgue measure $\lambda$ on $[0,1]$ and the multi-valued mapping $\Gamma_x : [0,1] \rightarrow \Theta$ (Nguyen 2006). Thus, we have

$$Bel_x^\Theta(A) = \lambda(\{\gamma \in [0,1] | \Gamma_x(\gamma) \subseteq A\}), \tag{16}$$

$$Pl_x^\Theta(A) = \lambda(\{\gamma \in [0,1] | \Gamma_x(\gamma) \cap A \neq \emptyset\}), \tag{17}$$

for all $A \subseteq \Theta$.

The sampling model of Dempster proposes to express $Y$ using a function $\varphi$ depending on the parameter $\theta$ and some unobserved variable $Z \in \mathbb{Z}$, whose probability distribution $\mu$ is known and independent of $\theta$:

$$Y = \varphi(\theta, Z). \tag{18}$$

From Eqs. (15) and (18), for a given $(\gamma, z) \in [0,1] \times \mathbb{Z}$, we can assert that $Y \in \varphi(\Gamma_x(\gamma), z)$. This can be represented by a multi-valued mapping $\Gamma'_x : [0,1] \times \mathbb{Z} \rightarrow \mathbb{Y}$ defined by composing $\Gamma_x$ with $\varphi$, *i.e.*, $\Gamma'_x(\gamma, z) = \varphi(\Gamma_x(\gamma), z), \forall (\gamma, z) \in [0,1] \times \mathbb{Z}$. The product measure $\lambda \otimes \mu$ on $[0,1] \times \mathbb{Z}$ and the multi-valued mapping $\Gamma'_x$ induce the belief and plausibility functions on $\mathbb{Y}$, which are defined by

$$Bel_x^\mathbb{Y}(A) = (\lambda \otimes \mu)(\{(\gamma, z) | \varphi(\Gamma_x(\gamma), z) \subseteq A\}), \tag{19}$$

$$Pl_x^\mathbb{Y}(A) = (\lambda \otimes \mu)(\{(\gamma, z) | \varphi(\Gamma_x(\gamma), z) \cap A \neq \emptyset\}), \tag{20}$$

for all $A \subseteq \mathbb{Y}$.

Let us consider a binary case, which will be useful hereafter. Let $Y \in \mathbb{Y} = \{0,1\}$ be a random variable with a Bernoulli distribution, *i.e.*, $Y \sim \mathcal{B}(\theta)$. In that case, the function $\varphi$ can be defined as follows:

$$Y = \varphi(\theta, Z) = \begin{cases} 1, \text{ if } Z \leq \theta, \\ 0, \text{ otherwise}, \end{cases} \tag{21}$$

with $Z$ having a uniform distribution on $[0,1]$. Assume that the consonant belief function $Bel_x^\Theta$ has a unimodal and continuous contour function $pl_x^\Theta$. In that case, each level set of $Bel_x^\Theta$ is a closed interval, *i.e.*, $\Gamma_x(\gamma) = [U(\gamma), V(\gamma)]$ (Dempster 1968), and the multi-valued mapping $\Gamma'_x$ defined by composing $\Gamma_x$ with $\varphi$, is given by

$$\Gamma'_x(\gamma, z) = \varphi([U(\gamma), V(\gamma)], z) = \begin{cases} \{1\}, & \text{if } z \leq U(\gamma), \\ \{0\}, & \text{if } z > V(\gamma), \\ \{0,1\}, & \text{otherwise}. \end{cases} \tag{22}$$

By applying Eq. (19), we get

$$Bel_x^\mathbb{Y}(\{1\}) = (\lambda \otimes \mu)(\{(\gamma, z) | z \leq U(\gamma)\}), \tag{23}$$

$$Bel_x^\mathbb{Y}(\{0\}) = (\lambda \otimes \mu)(\{(\gamma, z) | z > V(\gamma)\}). \tag{24}$$

Xu et al. (2016) showed that in this situation, the belief function $Bel_x^\mathbb{Y}$ and plausibility function $Pl_x^\mathbb{Y}$ are defined by

$$Bel_x^\mathbb{Y}(\{1\}) = \hat{\theta} - \int_0^{\hat{\theta}} pl_x^\Theta(u) du, \tag{25}$$

$$Pl_x^\mathbb{Y}(\{1\}) = \hat{\theta} + \int_{\hat{\theta}}^1 pl_x^\Theta(v) dv, \tag{26}$$

where $\hat{\theta}$ maximizes $pl_x^\Theta$.

Let us consider again the particular case of Section 2.2, where $X \sim \mathcal{B}(n, \theta)$. In that case, the contour function on $\Theta$ defined in Eq. (14) is unimodal and continuous, as illustrated in Figure 1. Thus, to represent knowledge about an unobserved data $Y \in \mathbb{Y}$, with $Y \sim \mathcal{B}(\theta)$, we can apply Eqs. (25) and (26) and Xu *et al.* showed that the obtained belief and plausibility functions boil down in that case to (Xu et al. 2016):

$$Bel_x^\mathbb{Y}(\{1\}) = \begin{cases} 0, & \text{if } \hat{\theta} = 0, \\ \hat{\theta} - \frac{B(\hat{\theta}; x+1, n-x+1)}{\hat{\theta}^x (1-\hat{\theta})^{n-x}}, & \text{if } 0 < \hat{\theta} < 1, \\ \frac{n}{n+1}, & \text{if } \hat{\theta} = 1, \end{cases} \tag{27}$$

$$Pl_x^{\mathbb{Y}}(\{1\}) = \begin{cases} \frac{1}{n+1}, & \text{if } \hat{\theta} = 0, \\ \hat{\theta} + \frac{\overline{B}(\hat{\theta};x+1,n-x+1)}{\hat{\theta}^x(1-\hat{\theta})^{n-x}}, & \text{if } 0 < \hat{\theta} < 1, \\ 1, & \text{if } \hat{\theta} = 1, \end{cases} \tag{28}$$

where $\underline{B}$ and $\overline{B}$ are respectively the lower and upper incomplete beta functions, defined when $a$ and $b$ are integers and $0 < z < 1$ by

$$\underline{B}(z;a,b) = \sum_{j=a}^{a+b-1} \frac{(a-1)!(b-1)!}{j!(a+b-1-j)!} z^j (1-z)^{a+b-1-j}, \tag{29}$$

and

$$\overline{B}(z;a,b) = \underline{B}(1-z;b,a). \tag{30}$$

## 3 Calibration of a single binary SVM classifier

Let us consider an object, whose true label $y$ is such that $y \in \mathbb{Y} = \{0,1\}$, and a confidence score $s \in \mathbb{R}$ returned by a classifier after observing this object. To learn how to interpret what this score represents with respect to $y$, a step called calibration may be used. This step relies on a training set $\mathcal{X}$, which contains $n$ other objects for which the label is known, and for which we observed the score that the classifier returned, *i.e.*, $\mathcal{X} = \{(s_1,y_1),...,(s_n,y_n)\}$ where $s_i$ represents the score given by the classifier for the $i^{th}$ object whose true label is $y_i$. The calibration procedures commonly used are the binning (Zadrozny and Elkan 2001), isotonic regression (Zadrozny and Elkan 2002) and logistic regression (Platt 1999). This paper focuses on binning and logistic regression as the isotonic regression can be seen as an intermediary approach between these two (Zadrozny and Elkan 2002). The probabilistic version of these two calibrations is described in Section 3.1, followed by their extension to the evidential framework in Section 3.2.

### 3.1 Probabilistic calibration of a single classifier

Given a score $s \in \mathbb{R}$ returned by a classifier after observing a given object, the aim of the calibration in the probabilistic framework consists in estimating the probability distribution $P^{\mathbb{Y}}(\cdot|s)$.

### *3.1.1 Binning*

The binning approach consists in dividing the score spaces into $B_U$ different bins, for example $(-3;-2]$, $(-2;-1]$, etc. For each bin $j$, the number $k_j$ of couples $(s_i,y_i) \in \mathcal{X}$ such as $y_i = 1$ and $s_i$ is in bin $j$, and the number $n_j$ of couples $(s_i,y_i) \in \mathcal{X}$ such as $s_i$ is in bin $j$ can be obtained. Then, for a score $s$ such that $s$ belongs to bin $j$, we have

$$P^{\mathbb{Y}}(y = 1|s) = \frac{k_j}{n_j}. \tag{31}$$

### *3.1.2 Logistic regression*

The calibration based on logistic regression proposed by Platt (1999) is a more elaborate method, which is based on fitting a sigmoid function $h$ defined by

$$P^{\mathbb{Y}}(y = 1|s) \approx h_s(\sigma) = \frac{1}{1 + e^{(\sigma_0 + \sigma_1 s)}}, \tag{32}$$

where the parameter $\sigma = (\sigma_0, \sigma_1) \in \mathbb{R}^2$ is chosen as the one maximizing the following likelihood function:

$$L_{\mathcal{X}}(\sigma) = \prod_{i=1}^{n} p_i^{t_i}(1-p_i)^{1-t_i}, \tag{33}$$

with

$$p_i = \frac{1}{1 + e^{(\sigma_0 + \sigma_1 s_i)}}, \tag{34}$$

and

$$t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if } y_i = 1, \\ \frac{1}{N_- + 2} & \text{if } y_i = 0, \end{cases} \tag{35}$$

where $N_+$ and $N_-$ are respectively the number of positive and negative samples in the training set $\mathcal{X}$.

Yet, it is usually easier to maximize the log-likelihood instead, which is defined by

$$\ell_{\mathcal{X}}(\sigma) = \mathrm{log} L_{\mathcal{X}}(\sigma) \tag{36}$$

$$= \sum_{i}^{n} \big(t_i \mathrm{log}(p_i) + (1-t_i)\mathrm{log}(1-p_i)\big). \tag{37}$$

Since the logarithm function is a strictly increasing function, maximizing the logarithm of the likelihood is the same as maximizing the likelihood. The parameter $\sigma$ maximizing this log-likelihood function can be approximated using iterative methods such as gradient descent. As the log-likelihood function of the logistic regression is concave (Minka 2003), all local maxima are global maxima and thus, an unique solution is found for $\sigma$.

We may notice that the less training samples are available, the more the estimated probabilities are uncertain. Within this scope, Xu et al. (2016) proposed to refine the above calibrations using the theory of evidence, in order to better handle the uncertainties. The following section recalls the evidential versions of the binning and logistic regression calibration procedures.

### 3.2 Evidential calibration of a single classifier

Xu *et al.* have recently extended the probabilistic calibration methods to the evidential framework (Xu et al. 2016). In their approach, the calibration of a given score $s$ is seen as a prediction problem of a Bernoulli variable $Y \in \mathbb{Y} = \{0,1\}$ with parameter $\theta$, where uncertainty on $\theta$ depends on $s$. They studied different models to estimate the uncertainty on $\theta$, and highlighted in particular the benefits of the so-called likelihood-based model. Thus, this paper focuses on the evidential extension of binning and logistic regression calibrations based on this likelihood model. These evidential calibration procedures yields a MF $m^{\mathbb{Y}}(\cdot|s)$ (rather than a probability distribution), equivalently represented by the belief and plausibility functions $Bel^{\mathbb{Y}}(\cdot|s)$ and $Pl^{\mathbb{Y}}(\cdot|s)$.

#### *3.2.1 Binning*

For a given bin $j$, binning can be seen as a binomial experiment, where the number of examples $n_j$ corresponds to the number of trials and the number of positive examples $k_j$ represents the number of successes. Thus, it corresponds to the particular case of estimation considered in Section 2.2, and used for forecasting in Section 2.3. Considering that the given score $s$ is in bin $j$, the likelihood-based contour function defined in Eq. (14) becomes

$$pl^{\Theta}_{\mathcal{X}}(\theta|s) = \frac{\theta^{k_j}(1-\theta)^{n_j-k_j}}{\hat{\theta}^{k_j}(1-\hat{\theta})^{n_j-k_j}}, \tag{38}$$

where $\hat{\theta} = \frac{k_j}{n_j}$ is the Maximum Likelihood Estimate (MLE) of $\theta$. The belief and plausibility functions $Bel^{\mathbb{Y}}(\cdot|s)$ and $Pl^{\mathbb{Y}}(\cdot|s)$ are then simply obtained using Eq. (27) and (28) with $x = k_j$ and $n = n_j$.

#### *3.2.2 Logistic regression*

Logistic-based calibration can also be extended in the evidential framework through the likelihood model. Xu et al. (2016) express uncertainty on the parameter $\sigma = (\sigma_0, \sigma_1)$ of the sigmoid function, by a consonant belief function $Bel^{\Sigma}$, whose contour function is defined by

$$pl^{\Sigma}_{\mathcal{X}}(\sigma) = \frac{L_{\mathcal{X}}(\sigma)}{L_{\mathcal{X}}(\hat{\sigma})}, \quad \forall \sigma \in \Sigma, \tag{39}$$

where $\hat{\sigma} = (\hat{\sigma}_0, \hat{\sigma}_1)$ is the MLE of $\sigma$ and $L_{\mathcal{X}}$ is the likelihood function defined in Eq. (33). The corresponding plausibility function is defined as

$$Pl^{\Sigma}_{\mathcal{X}}(A) = \sup_{\sigma \in A} pl^{\Sigma}_{\mathcal{X}}(\sigma), \quad \forall A \subseteq \Sigma. \tag{40}$$

As seen in Section 2.3, the belief and plausibility functions on $\mathbb{Y}$ can be deduced from the contour function $pl^{\Theta}_{\mathcal{X}}$ defined on $\Theta$. Xu *et al.* showed in (Xu et al. 2016) that this function $pl^{\Theta}_{\mathcal{X}}$ can be computed from $Pl^{\Sigma}_{\mathcal{X}}$. Indeed, as $\theta$ is defined by $\theta = h_s(\sigma)$, we get

$$pl^{\Theta}_{\mathcal{X}}(\theta|s) = \begin{cases} 0 & \text{if } \theta \in \{0,1\}, \\ Pl^{\Sigma}_{\mathcal{X}}(h_s^{-1}(\theta)) & \text{otherwise,} \end{cases} \tag{41}$$

with

$$h_s^{-1}(\theta) = \{(\sigma_0,\sigma_1) \in \Sigma | h_s(\sigma) = \theta\}, \tag{42}$$

$$= \left\{(\sigma_0,\sigma_1) \in \Sigma | \frac{1}{1+\exp(\sigma_0+\sigma_1 s)} = \theta\right\}, \tag{43}$$

$$= \left\{(\sigma_0,\sigma_1) \in \Sigma | \sigma_0 = \ln(\theta^{-1}-1) - \sigma_1 s\right\}. \tag{44}$$

Finally, Eqs. (41) and (44) yield the following function

$$pl^{\Theta}_{\mathcal{X}}(\theta|s) = \sup_{\sigma_1 \in \mathbb{R}} pl^{\Sigma}_{\mathcal{X}}(\ln(\theta^{-1}-1) - \sigma_1 s, \sigma_1), \quad \forall \theta \in [0,1]. \tag{45}$$

The value $pl^{\Theta}_{\mathcal{X}}(\theta|s)$ can be obtained by an iterative maximization algorithm, for all $\theta \in [0,1]$. The belief and plausibility functions $Bel^{\mathbb{Y}}(\cdot|s)$ and $Pl^{\mathbb{Y}}(\cdot|s)$ can then be calculated using Eqs. (25) and (26).

## 4 An evidential joint calibration approach

In a context of multiple classifiers, one may independently calibrate the score given by each classifier after observing an object, and merge them using a predetermined rule of combination. In particular, this kind of process is followed by Xu et al. (2016), where scores provided by binary SVM classifiers are transformed into belief functions using evidential calibration and combined using Dempster's rule of combination. We refer hereafter to this latter approach as the disjoint method.

We propose in this paper to use the multivariable versions of the techniques underlying the calibrations, and to apply it to the outputs of multiple classifiers, *i.e.*, to perform a joint calibration of the scores provided by the binary SVM classifiers. More specifically, in order to better handle the uncertainties of the calibration process, we propose to perform the joint calibration in the evidential framework.

For a given object, we take as input the score vector $\mathbf{s} = (s_1, s_2, ..., s_J)$, with $s_j$ the score returned by the $j^{th}$ classifier after observing the object. The required training set is defined by $\mathcal{X}' = \{(s_{11}, s_{21}, ..., s_{J1}, y_1), ..., (s_{1n}, s_{2n}, ..., s_{Jn}, y_n)\}$, where $s_{ji}$ corresponds to the score given by the $j^{th}$ classifier for the $i^{th}$ test sample, and $y_i$ the true label of this sample.

We first expose in Section 4.1 the multivariable version of binning calibration, followed by the multivariable version of the calibration based on logistic regression in Section 4.2.

## 4.1 Joint binning

The idea consists in dividing the score space into multidimensional bins (cells), or more precisely into $J$- dimensional bins with $J$ the number of classifiers. Let us illustrate the building of these cells with a $2D$ scenario, *i.e.*, when only two classifiers are considered. If the first classifier has score values between -3 and 3 and the second classifier has score values between -2 and 1, the score space is $[-3,3] \times [-2,1]$. This score space can be divided in different ways. In particular, a number of bins per classifier can be chosen and the score space can be divided uniformly based on this number. An illustration of this naive scheme is given in Figure 2, where the number of bins by classifier, denoted $B_M$, is chosen equal to 5.

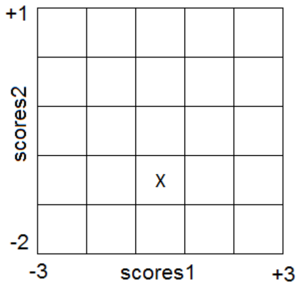


Fig. 2: Example of score space for joint binning, with $J = 2$ and $B_M = 5$.

Given a cell $c$, the number $k_c$ of tuples $(s_{1i}, ..., s_{Ji}, y_i) \in \mathcal{X}'$ such that $y_i = 1$ and $(s_{1i}, s_{2i}, ..., s_{Ji})$ belongs to cell $c$, and the number $n_c$ of tuples such that $(s_{1i}, ..., s_{Ji})$ belongs to cell $c$, can be obtained. For a given input vector $\mathbf{s} = (s_1, s_2, ..., s_J)$ such that $\mathbf{s}$ belongs to the cell $c$, we have

$$P^{\mathbb{Y}}(y = 1|\mathbf{s}) = \frac{k_c}{n_c}. \tag{46}$$

For instance, let us consider that we have $\mathbf{s} = (0.5, -1)$, *i.e.*, after observing a given example the first classifier returns the score 0.5 and the second $-1$. The probability associated to this object can thus be found by looking into the corresponding cell $c$, which is the one marked by a cross in Figure 2.

This probabilistic joint approach of binning can be extended to the evidential framework. Similarly to the single classifier case, the label $y$ of a given score vector $\mathbf{s}$ can be seen as a realization of a random variable with a Bernoulli distribution, and binning can be seen as a binomial experiment for each cell. If the score vector $\mathbf{s}$ is in cell $c$, the belief and plausibility functions associated to this score vector can be calculated using the following equations:

$$Bel^{\mathbb{Y}}(\{1\}|\mathbf{s}) = \begin{cases} 0, & \text{if } \hat{\theta} = 0, \\ \hat{\theta} - \frac{B(\hat{\theta}; k_c+1, n_c-k_c+1)}{\hat{\theta}^{k_c}(1-\hat{\theta})^{n_c-k_c}}, & \text{if } 0 < \hat{\theta} < 1, \\ \frac{n_c}{n_c+1}, & \text{if } \hat{\theta} = 1, \end{cases} \tag{47}$$

$$Pl^{\mathbb{Y}}(\{1\}|\mathbf{s}) = \begin{cases} \frac{1}{n_c+1}, & \text{if } \hat{\theta} = 0, \\ \hat{\theta} + \frac{B(\hat{\theta}; k_c+1, n_c-k_c+1)}{\hat{\theta}^{k_c}(1-\hat{\theta})^{n_c-k_c}}, & \text{if } 0 < \hat{\theta} < 1, \\ 1, & \text{if } \hat{\theta} = 1, \end{cases} \tag{48}$$

with $\hat{\theta} = \frac{k_c}{n_c}$.

## 4.2 Joint logistic regression

The logistic regression, exposed in Section 3, is used to calibrate a score given by a single classifier. Yet, the logistic model works as well when more than one input is available: it is then called a multivariable (or multiple) logistic regression (Hosmer et al. 2013). It has been widely used in many applications, such as for instance in medicine field (Bagley et al. 2001). We propose to use this multiple version of logistic regression and apply it to the vector of scores returned by different classifiers for a given object, in order to calibrate this vector.

Given a vector of scores $\mathbf{s} = (s_1, s_2, ..., s_J)$, the probabilistic joint calibration based on multiple logistic regression is defined by

$$P^{\mathbb{Y}}(y = 1|\mathbf{s}) = \frac{1}{1 + \exp(\sigma_0 + \sigma_1 s_1 + \sigma_2 s_2 + ... + \sigma_J s_J)}, \tag{49}$$

where the parameter $\sigma = (\sigma_0, ..., \sigma_J) \in \mathbb{R}^{J+1}$ is obtained by maximizing the likelihood function $L_{\mathcal{X}'}$ defined by

$$L_{\mathcal{X}'}(\sigma) = \prod_{i=1}^{n} p_i^{t_i}(1 - p_i)^{1-t_i}, \tag{50}$$

with

$$p_i = \frac{1}{1 + \exp(\sigma_0 + \sigma_1 s_{1i} + ... + \sigma_J s_{Ji})}, \tag{51}$$

and

$$t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} \text{ if } y_i = 1, \\ \frac{1}{N_- + 2} \text{ if } y_i = 0, \end{cases} \tag{52}$$

where $N_+$ and $N_-$ are respectively the number of positive and negative samples in the training set $\mathcal{X}'$. The log-likelihood can be used instead of the likelihood, and an unique solution is found for $\sigma$.

We propose to extend this joint logistic-based calibration to the evidential framework by following the same likelihood-based reasoning as for the single classifier case. The knowledge about $\sigma = (\sigma_0, ..., \sigma_J)$ can be represented by a consonant belief function whose contour function is defined by

$$pl^{\varSigma}_{\mathcal{X}'} = \frac{L_{\mathcal{X}'}(\sigma)}{L_{\mathcal{X}'}(\hat{\sigma})}, \quad \forall \sigma \in \varSigma. \tag{53}$$

Furthermore, $pl^{\Theta}_{\mathcal{X}'}$ can be computed from $Pl^{\varSigma}_{\mathcal{X}'}$:

$$pl^{\Theta}_{\mathcal{X}'}(\theta|\mathbf{s}) = \begin{cases} 0 & \text{if } \theta \in \{0, 1\}, \\ Pl^{\varSigma}_{\mathcal{X}'}(h^{-1}_{\mathbf{s}}(\theta)) & \text{otherwise,} \end{cases} \tag{54}$$

with

$$h^{-1}_{\mathbf{s}}(\theta) = \{(\sigma_0, \sigma_1, ..., \sigma_J) \in \varSigma | h_s(\sigma) = \theta\}, \tag{55}$$

$$= \left\{(\sigma_0, ..., \sigma_J) \in \varSigma | \frac{1}{1 + \exp(\sigma_0 + \sigma_1 s_1 + ... + \sigma_J s_J)} = \theta\right\}, \tag{56}$$

$$= \left\{(\sigma_0, ..., \sigma_J) \in \varSigma | \sigma_0 = \ln(\theta^{-1} - 1) - \sigma_1 s_1 - ... - \sigma_J s_J\right\}. \tag{57}$$

Thus, the contour function $pl^{\Theta}_{\mathcal{X}'}(\theta|\mathbf{s})$ is defined by

$$pl^{\Theta}_{\mathcal{X}'}(\theta|\mathbf{s}) =$$

$$\sup_{\sigma_1, ..., \sigma_J \in \mathbb{R}} pl^{\varSigma}_{\mathcal{X}'}(\ln(\theta^{-1} - 1) - \sigma_1 s_1 - ... - \sigma_J s_J, \sigma_1, ..., \sigma_J), \tag{58}$$

for all $\theta \in [0, 1]$. The vector of parameters $(\sigma_1, \sigma_2, ..., \sigma_J)$ which maximizes $pl^{\varSigma}_{\mathcal{X}'}$ can be approximated using an iterative maximization algorithm (the computational complexity of such algorithm is $O(nJ)$ per iteration). Then, the belief and plausibility functions $Bel^{\mathbb{Y}}(\cdot|\mathbf{s})$ and $Pl^{\mathbb{Y}}(\cdot|\mathbf{s})$ can be obtained trough Eq. (25) and (26).

## 5 Experiments

In this section, the performance of our proposed evidential joint calibration approach is compared to those of other approaches using different datasets, which are presented in Section 5.1. In Section 5.2, our approach is compared to the disjoint approach of Xu *et al.*. Both binning and logistic regression calibrations are studied. Then, in Section 5.3, these latter two calibrations are compared to a conceptually similar approach, that is a trainable combiner based on an evidential classifier, *i.e.*, a classifier returning a mass function after observing an object. Finally, we focus on the calibration based on multiple logistic regression and we compare the probabilistic and evidential versions of this joint calibration in Section 5.4.

### 5.1 Datasets

The experiments are conducted on five binary classification problems provided by UCI repository (Bache and Lichman 2013). They are all of different sizes, and their sample vectors have various number of features. This is presented in Table 1.

| Dataset | # instance vectors | # features |
|---|---|---|
| Australian | 690 | 14 |
| Diabetes | 768 | 8 |
| Heart | 270 | 13 |
| Ionosphere | 351 | 34 |
| Sonar | 208 | 60 |

Table 1: Number of instance vectors and number of features by vector for different datasets from UCI.

We also simulated a dataset composed of 360 randomly generated instance vectors from a multivariate normal distribution, with means $\mu_0 = (-1, 0)$ in class 0 and $\mu_1 = (1, 1)$ in class 1, and with a covariance matrix equals to $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ for both classes. Each instance vector has two features. An illustration of these data on the feature space are represented in Figure 3, where $x$ and $y$ represent respectively the first and second feature of each instance vector.

### 5.2 Comparison with Xu *et al.*'s approach (Xu et al. 2016)

The following experiment follows the same protocol as the first experiment detailed in (Xu et al. 2016). For

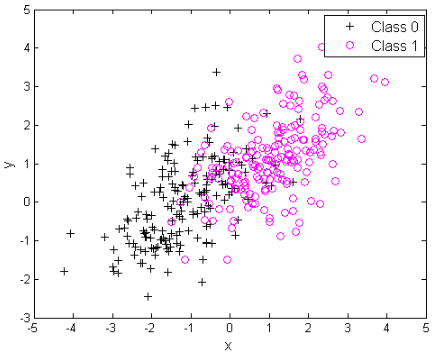


Fig. 3: Illustration of 360 instance vectors of the simulated dataset.

each dataset, three SVM classifiers are trained on non-overlapping subsets, using the LIBSVM library (Chang and Lin 2011). The numbers of examples used for training and testing for each dataset are described in Table 2. For the first two classifiers, the number of training examples is fixed while different training set sizes are considered for the third one. The training set of each classifier is partitioned into two equal sized-subsets. One of these subsets is for training the classifier, and in Xu *et al.*'s approach the second subset is for training the calibration of the classifier. In the proposed approach, the joint calibration is trained using the set composed of the concatenation of each second subset of each classifier.

| Dataset | #Train 1 | #Train 2 | #Train 3 | #Test |
|---|---|---|---|---|
| Australian | 30 | 70 | 10-60-190 | 400 |
| Diabetes | 30 | 70 | 10-50-200 | 468 |
| Heart | 20 | 40 | 10-50-140 | 70 |
| Ionosphere | 20 | 40 | 10-80-190 | 101 |
| Sonar | 20 | 40 | 10-40-90 | 58 |
| Simulated | 20 | 40 | 10-50-200 | 100 |

Table 2: Number of examples used for training and testing.

For each sample belonging to the test set, the three classifiers return a score. In Xu *et al.*'s approach, each of these scores is calibrated using the trained calibration of its corresponding classifier, and the three obtained mass functions are merged into a final mass function using Dempster's rule. In our proposed approach, the scores are grouped into a score vector and this vector is calibrated using a joint calibration, which directly returns a final mass function. In both cases, the decision corresponds to the singleton with the highest belief, since we use $\{0, 1\}$ costs without the possibility to reject, in which case upper and lower expected costs lead to the same decision. The error rate is calculated on the test set and corresponds to the number of samples misclassified over the number of tested samples. The whole process is repeated for 100 rounds of random partitioning, thus the final error rate corresponds to the average of 100 calculated error rates.

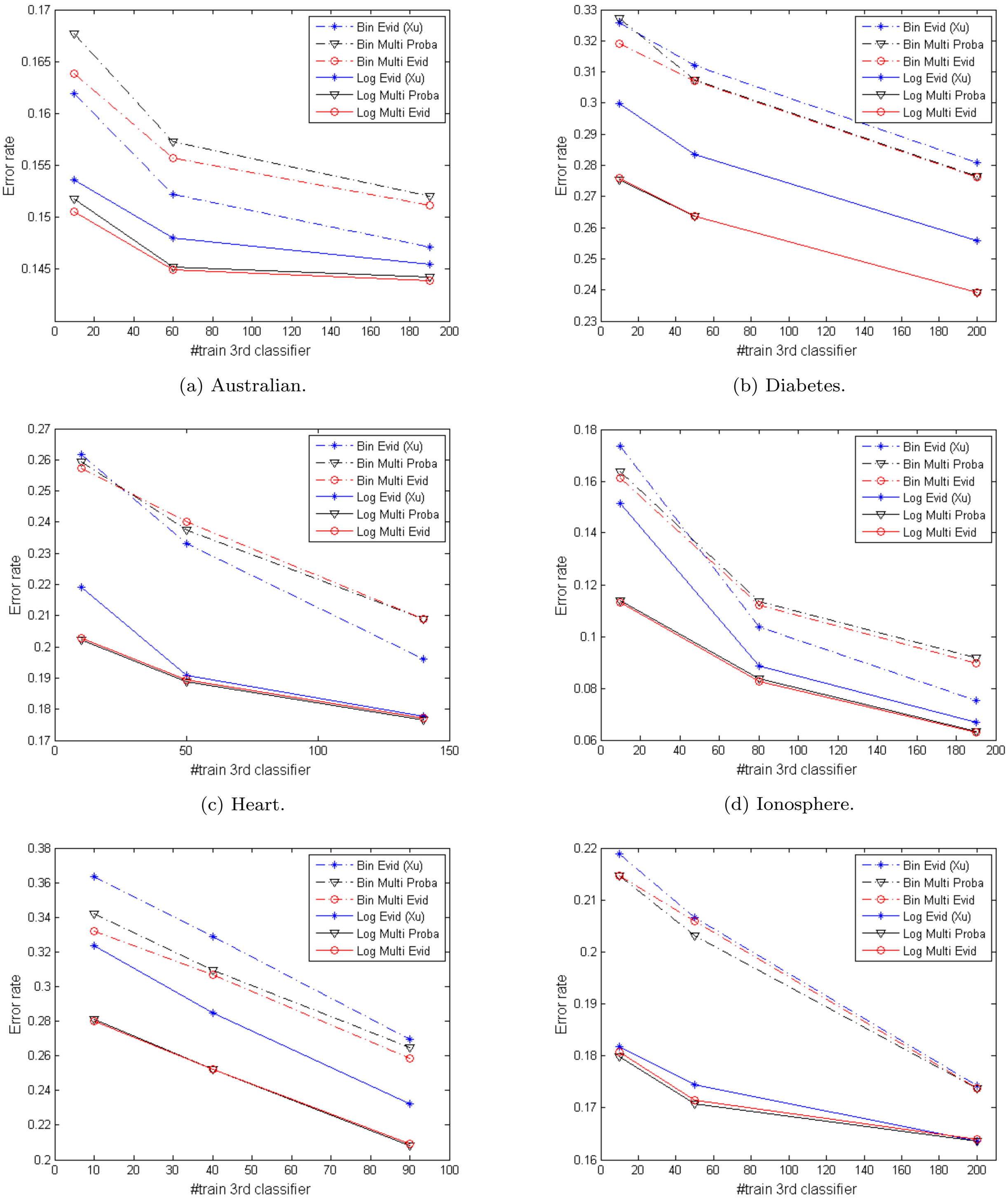
For the binning calibration, we may remark that there are in total a number of $B_U \times J$ bins in the disjoint case against $(B_M)^J$ bins for the joint binning. In order to fairly compare our approach to the disjoint one, the number of bins for each classifier is chosen such that each method has the same total number of bins. In particular, as $J = 3$, we chose respectively $B_U = 9$ and $B_M = 3$ for disjoint and joint approaches.

Figure 4 shows the results of the experiments for binning and logistic-based approaches, in the evidential framework, and for disjoint and joint cases. Results of the probabilistic version of joint calibrations are also given. As it can be noticed, the approaches based on the logistic regression are always better than those based on binning, as their obtained error rates are lower. For binning approaches, the joint case is not always better than the disjoint case, but it might come from the chosen value for $B_M$; with a higher value, the results might be better. For logistic regression, the evidential joint approach always presents better results than the evidential disjoint approach. It can also be noticed that the probabilistic and evidential joint versions nearly give the same results in this experiment. Comparison between probabilistic and evidential versions of calibration based on multiple logistic regression will be performed in Section 5.4.

### 5.3 Comparison with evidential trainable combiner approach

In the previous experiment, we compared our approach to its probabilistic version and to the so-called disjoint method, which belongs to the non trainable combiner category. In this section, we perform the same experiment but with the aim of comparing our results to those of approaches of the same category, *i.e.*, to evidential trainable combiners. Indeed, there exist other approaches similar to ours to be compared to, and in particular some methods which can take a score vector as input and return a belief function on the class of a given observed object.

The first evidential trainable combiner that we consider in this experiment relies on the evidential classifier described in (Denœux and Smets 2006) and based on the Generalized Bayesian Theorem (GBT) (Smets 1993).



(a) Australian.

(b) Diabetes.

(c) Heart.

(d) Ionosphere.

(e) Sonar.

(f) Simulated data.

Fig. 4: Average error rates using binning and logistic regression, with joint (referred to as “multi” in the figures) and disjoint approaches and with both probabilistic and evidential frameworks. The X-axis corresponds to the number of training examples used to train the third classifier.

Let us consider a classification problem with $\Omega = \{w_k\}_{k=1}^{K}$ the finite set of classes. After observing the feature vector $\mathbf{x}$ of an object, the aim is to obtain a belief function about the class label of this object, based on a training set $\mathcal{L} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i$ represents the feature vector of the $i^{th}$ object, whose true label is $y_i$. The application of the GBT gives the following MF on $\Omega$ about the class of $\mathbf{x}$ (Denœux and Smets 2006):

$$m^{\Omega}(A|\mathbf{x}) = \prod_{w_k \in A} Pl[w_k](\mathbf{x}) \prod_{w_k \in \overline{A}} (1 - Pl[w_k](\mathbf{x})), \tag{59}$$

$\forall A \subseteq \Omega$, where $\overline{A}$ denotes the complement of $A$, and $Pl[w_k](\mathbf{x})$ represents the plausibility of observing $\mathbf{x}$ under the hypothesis that the true class is $\omega_k$. In particular, Denœux and Smets (2006) have considered a special case, where

$$Pl[w_k](\mathbf{x}) = \frac{N(\mathbf{x}, k)}{N(k)}, \tag{60}$$

with $N(\mathbf{x}, k)$ the number of samples in $\mathcal{L}$ from class $w_k$ contained in a ball $S_r$ of radius $r$ and centered on $\mathbf{x}$, and $N(k)$ the total number of samples from class $w_k$ in $\mathcal{L}$.

We note that it may happen that $m^{\Omega}(\emptyset|\mathbf{x}) > 0$, and in that case the MF $m^{\Omega}(\cdot|\mathbf{x})$ can be transformed into a normalized MF $M^{\Omega}(\cdot|\mathbf{x})$ using the operation defined by

$$M^{\Omega}(A|\mathbf{x}) = \frac{m^{\Omega}(A|\mathbf{x})}{1 - m^{\Omega}(\emptyset|\mathbf{x})}, \qquad \forall A \subseteq \Omega, A \neq \emptyset, \tag{61}$$

and $M^{\Omega}(\emptyset|\mathbf{x}) = 0$.

We now apply this classifier to our binary problem, by taking the same inputs as for our approach. In particular, after observing a given object, the feature vector is now the vector of scores $\mathbf{s} = (s_1, ..., s_J)$ obtained by $J$ classifiers, and the training set $\mathcal{L}$ is now $\mathcal{X}'$. Using the definition of the MF given in Eq. (59) and the considered particular case of Eq. (60), we obtain the MF $m^{\mathbb{Y}}(\cdot|\mathbf{s})$ defined by

$$m^{\mathbb{Y}}(\{0\}|\mathbf{s}) = \frac{N(\mathbf{s}, 0)}{N(0)} \times (1 - \frac{N(\mathbf{s}, 1)}{N(1)}), \tag{62}$$

$$m^{\mathbb{Y}}(\{1\}|\mathbf{s}) = \frac{N(\mathbf{s}, 1)}{N(1)} \times (1 - \frac{N(\mathbf{s}, 0)}{N(0)}), \tag{63}$$

$$m^{\mathbb{Y}}(\{0, 1\}|\mathbf{s}) = \frac{N(\mathbf{s}, 1)}{N(1)} \times \frac{N(\mathbf{s}, 0)}{N(0)}, \tag{64}$$

and

$$m^{\mathbb{Y}}(\emptyset|\mathbf{s}) = (1 - \frac{N(\mathbf{s}, 0)}{N(0)}) \times (1 - \frac{N(\mathbf{s}, 1)}{N(1)}), \tag{65}$$

with $N(\mathbf{s}, k)$ the number of samples in $\mathcal{X}'$ from class $k$ (equal to 0 or 1), contained in a ball $S_r$ of radius $r$ and centered on $\mathbf{s}$. This MF is then normalized similarly as $m^{\Omega}(\cdot|\mathbf{x})$ is normalized using Eq. (61).

We may notice that using a ball $S_r$ to build the MFs has some similarities with our multivariable version of binning. Let us illustrate this statement with a simple example, using the dataset *Diabetes* and with $J = 2$. Figure 5 shows the scores returned by two trained classifiers for each sample of a given calibration training set. The X-axis corresponds to the scores given by the first classifier and Y-axis by the second one. A test sample is illustrated by a blue asterisk, and corresponds to $\mathbf{s} = (s_1, s_2)$ the values of the scores returned by the two classifiers. The continuous green lines correspond to the bounds of the joint binning, with $B_M = 3$, and the red circle represents the ball $S_r$ of the GBT-based classifier, with $r = 1$ and centered on $\mathbf{s}$. To build the MF $m^{\mathbb{Y}}(\cdot|\mathbf{s})$, the joint binning uses the training samples belonging to the bin containing $\mathbf{s}$, while the GBT-based classifier uses the ones contained by the ball $S_r$.

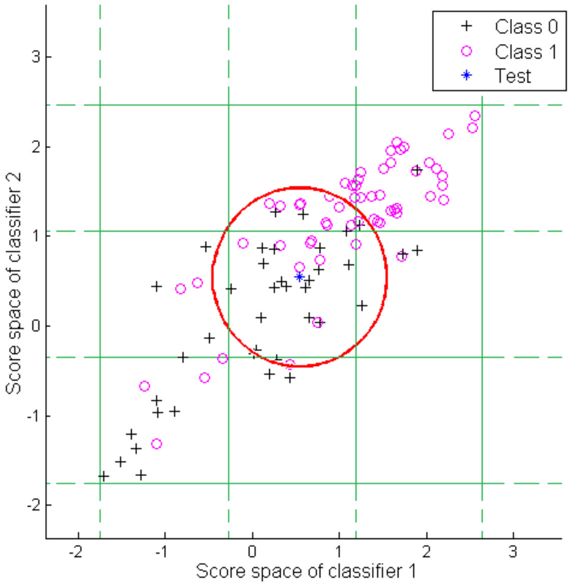


Fig. 5: Illustration of the multidimensional bins and the ball $S_r$, using *Diabetes* data.

The second evidential trainable combiner that we consider is the evidential $\kappa$ Nearest Neighbor ($\kappa$NN) classification rule (Denœux 1995), whose parameters are optimized using the procedure described in (Zouhal and Denœux 1998). Let us consider the same classification problem as for GBT-based approach, *i.e.*, obtaining a belief function about the class label of an observed object with feature vector $\mathbf{x}$, based on a training set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i$ represents the feature vector of the $i^{th}$ object, whose true label is $y_i$. If $\mathbf{x}$ is close to $\mathbf{x}_i$ according to some distance measure $d$, then it is reasonable to believe that both vectors belong to the same class. More generally, the closer $\mathbf{x}_i$ is to $\mathbf{x}$, the more reasons to believe that the class of $\mathbf{x}$ is the same as the one of $\mathbf{x}_i$. This piece of information brought by the $i^{th}$ object about the class of the observed object may be represented by a MF $m_i^{\Omega}$ defined by (Denœux 1995):

$$m_i^{\Omega}(\{\omega_k\}) = \alpha\phi_k(d_i), \tag{66}$$

$$m_i^{\Omega}(\Omega) = 1 - \alpha\phi_k(d_i), \tag{67}$$

$$m_i^{\Omega}(A) = 0, \quad \forall A \in 2^{\Omega}\backslash\{\Omega, \{\omega_k\}\}, \tag{68}$$

where $\omega_k$ is the class $y_i$ of the $i^{th}$ object, $d_i = d(\mathbf{x}, \mathbf{x_i})$ is the distance between the feature vector of the observed object and the feature vector of the $i^{th}$ object, $\alpha$ is a parameter such that $0 < \alpha < 1$ and $\phi_k$ is a decreasing function verifying $\phi_k(0) = 1$ and $\lim_{d\to\infty}\phi_k(d) = 0$. A common choice for $\phi_k$ is given by (Denœux 1995):

$$\phi_k(d) = \exp(-\alpha_k d^2), \tag{69}$$

where $\alpha_k > 0$ is a parameter associated to class $\omega_k$.

Thus, a MF may then be obtained for each sample of the training set $\mathcal{L}$. Denœux (Denœux 1995) proposed to pool by Dempster's rule the evidence of the $\kappa$ nearest neighbors, $1 \leq \kappa \leq n$, of the observed object in order to obtain a MF $m^{\Omega}(\cdot|\mathbf{x})$ about its class. Let $\kappa_{\mathbf{x}}$ denote the set of the $\kappa$ nearest objects of $\mathbf{x}$ in $\mathcal{L}$. The MF $m^{\Omega}(\cdot|\mathbf{x})$ is then defined as

$$m^{\Omega}(A|\mathbf{x}) = (\oplus_{\mathbf{x_i}\in\kappa_{\mathbf{x}}} m_i^{\Omega})(A), \quad \forall A \subseteq \Omega. \tag{70}$$

When applying this classifier to our binary problem, the feature vector $\mathbf{x}$ is now the vector of scores $\mathbf{s} = (s_1, ..., s_J)$ obtained by $J$ classifiers, and the training set $\mathcal{L}$ is now $\mathcal{X}'$.

We performed the experiment with $r = 1$ for the GBT-based approach and $\kappa = 15$ for the evidential $\kappa$NN approach[1] as some preliminary tests showed that the best results were obtained with these values.

Figure 6 shows the error rates for the GBT and $\kappa$NN-based approaches, compared to those obtained with our evidential multivariable versions of binning and logistic regression. As it can be noticed, the results obtained with the GBT and the $\kappa$NN-based classifiers are better than those obtained with the binning approach. It can be explained by the fact that in the binning approach the bounds of the multi-dimensional bins are fixed, and any test sample belonging to the same multi-dimensional bin has the same associated MF, no matter where the sample is positioned in the bin. By contrast, for the GBT classifier, the ball is centered on the considered test sample, so the neighborhood of the test sample is taken into account in a better way. A similar explanation can be provided for the $\kappa$NN classifier. Furthermore, with other values of $r$ and of $\kappa$, or with other size and number of our multi-dimensional bins, the obtained results may vary significantly, as these approaches highly rely on these parameters. Finally, we can see that the evidential joint calibration using logistic regression is always better than the GBT and $\kappa$NN-based approaches in our experiments.

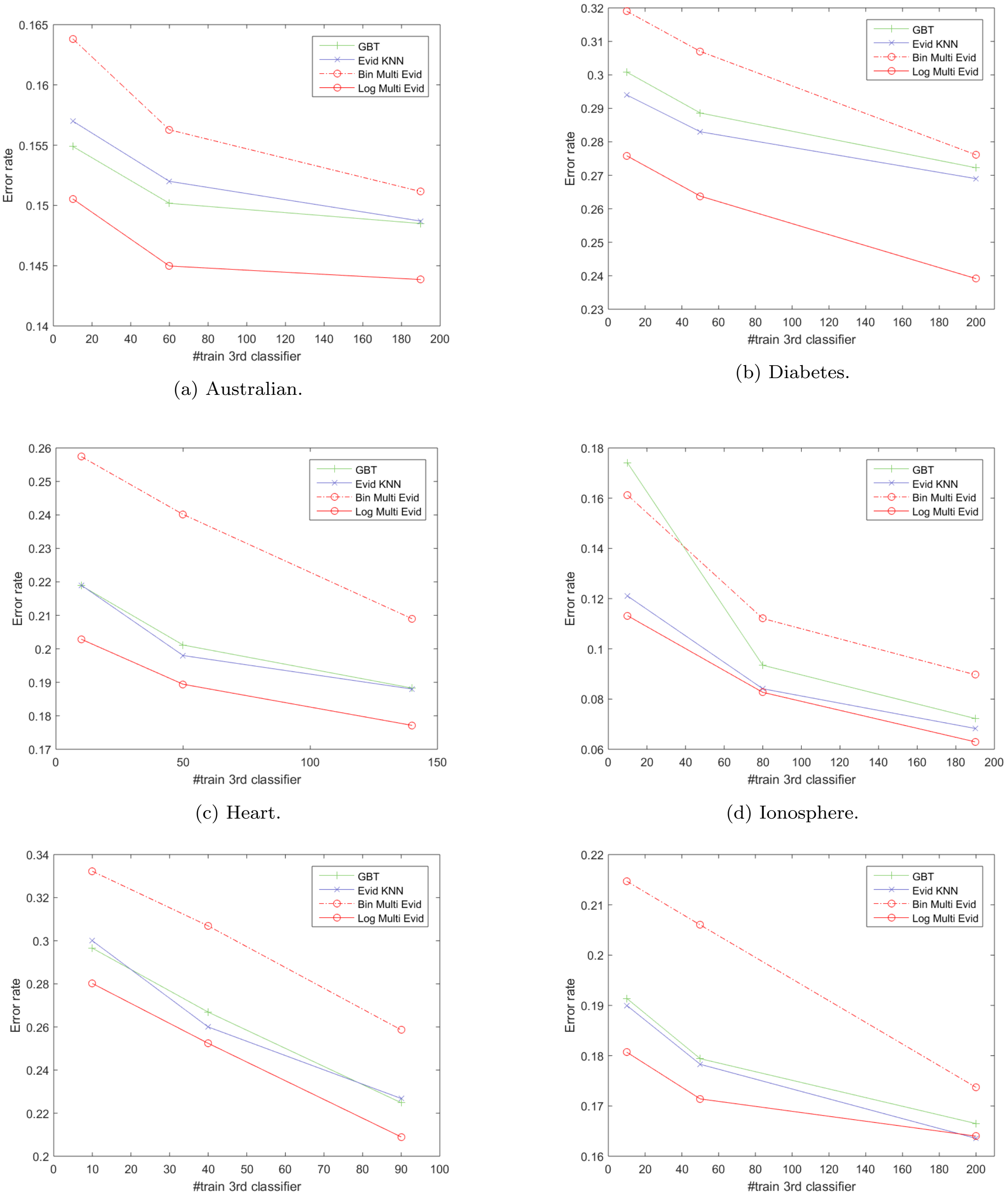
### 5.4 Comparison between evidential and probabilistic joint calibrations based on logistic regression

As seen in Sections 5.2 and 5.3, the evidential joint logistic-based calibration always presents the best results. Yet, we also noted (in Section 5.2) that the performance of the probabilistic version of this calibration were nearly the same. Thus, in this section, probabilistic and evidential versions of the calibration based on the multiple logistic regression are further compared. To do that, we introduce the possibility of a third decision for the system given a test sample, by allowing a reject option. Hence, for a given test sample, three possible decisions can be rendered: 0, 1, or $R$. This option $R$ expresses doubt and is used for some examples that are hard to classify. In addition, as recalled in Section 2.1, there are different decision-making criteria in the evidential framework and thus the evidential approach has two possible strategies of decision, either pessimistic or optimistic.

Using the simulated dataset previously defined, 290 training examples were generated: three SVM classifiers were trained with three non-overlapping subsets of 30 training examples of this set, and the joint calibration using logistic regression was trained with the remaining 200 examples of this set. Then, the same experiment was performed but the joint logistic-based calibration was trained with 15 examples instead of 200. The decision frontiers for both the pessimistic and optimistic strategies and for both cases are illustrated in Figure 7 for $R_{rej} = 0.15$.

As it can be seen, the evidential joint calibration based on the optimistic strategy tends to reject less the test samples than the two others. It is the exact opposite for the evidential joint calibration based on the pessimistic strategy, which decide to reject in more

[1] We used the software for the evidential $\kappa$NN classifier with parameter optimization available at: https://www.hds.utc.fr/~tdenoeux/dokuwiki/en/software/k-nn



(a) Australian.

(b) Diabetes.

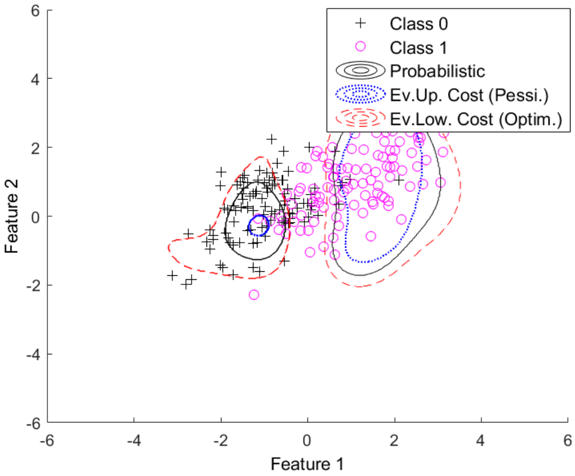
(c) Heart.

(d) Ionosphere.

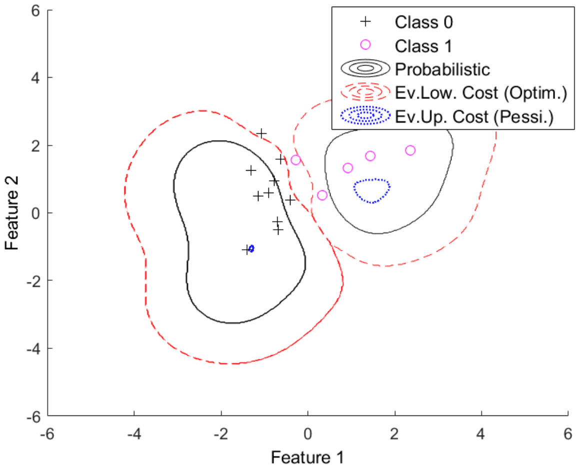
(e) Sonar.

(f) Simulated data.

Fig. 6: Average error rates using the GBT-based and $\kappa$NN-based approaches, and using the binning and logistic regression with evidential joint approaches. The X-axis corresponds to the number of training examples used to train the third classifier.



(a) Joint logistic-based calibration trained with 200 training samples

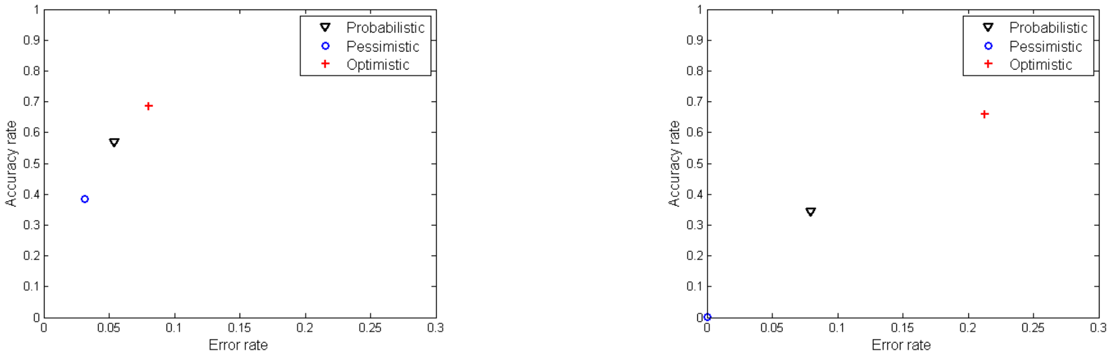


(b) Joint logistic-based calibration trained with 15 training samples

Fig. 7: Decision frontiers in feature space of the probabilistic and evidential joint calibrations based on logistic regression trained with 200 (7a) and 15 training examples (7b), and with $R_{rej} = 0.15$.

cases. The probabilistic approach is between these two. Furthermore, the frontiers associated to the pessimistic and optimistic strategies are a lot more distant from each other in Figure (7b) than in Figure (7a), *i.e.*, when there are less examples to train the joint calibration and thus more uncertainties. Probabilistic approach is only represented by one frontier, so the impact of the uncertainties is not visible. Thus, the evidential approach better reflects the uncertainties than the probabilistic one.

Let us illustrate this point further. The three SVM classifiers were still trained with three non-overlapping subsets of 30 training samples, and the calibration with 200 then 15 samples. We calculated the error rate and accuracy rate for 100 test samples and with $R_{rej} = 0.15$. Accuracy rate represents the number of correctly classified objects over the number of classified objects, *i.e.*, not over the total number of test examples as some of them are rejected. The whole process was repeated for 100 rounds of random partitioning. As it can be seen in Figure 8, if there are a lot of examples to train the joint calibration, the obtained error rates are almost equal. Yet, when less training examples are available, the two points obtained for the evidential approach are more distant from each other. This interval reflects the uncertainties, as when it is larger the uncertainties are more important. This information cannot be obtained with the probabilistic calibration, as it is represented by only one point. Thus, the joint calibration based on evidence theory better reflects the uncertainties.



(a) 200 training examples

(b) 15 training examples

Fig. 8: Obtained error rates for $R_{rej} = 0.15$ and with 200 (8a) and 15 (8b) training examples.

Finally, we performed a similar experiment with $R_{rej}$ varying from 0 to 1, on five datasets (*Australian*, *Diabetes*, *Heart*, *Ionosphere*, *Sonar*) of UCI repository (Bache and Lichman 2013) and on the simulated dataset. The only difference with the previous experiment is that the multivariable logistic regression was trained with 45 then 15 samples. Due to the size of *Sonar*, it was tested on 50 sample tests instead of 100 for the other datasets. The whole process was carried out for 100 rounds of random partitioning and Figures 9 and 10 show the obtained results.

As it can be noticed, for a given error rate, the results obtained with the pessimistic strategy has a higher (or equal) accuracy rate than the probabilistic calibration when few training examples are available (right column). Let us underline that for a fixed error rate, the accuracy rates obtained with the probabilistic calibration and the pessimistic strategy are obtained for different values of $R_{rej}$ (as seen in the previous experiment, the results of which are given in Figure 8, a given value of $R_{rej}$ leads in general to different error rates). Furthermore, when the number of training examples is more important (left column of Figures 9 and 10), the obtained results become similar for the probabilistic and evidential approaches, as should be.

## 6 Conclusion

In this paper, an evidential joint calibration approach was proposed in order to handle the scores returned by multiple SVM classifiers. This approach belongs to the category of trainable combiners as it takes a score vector as input and does not need a predetermined rule of combination. We used evidence theory to prevent the over-fitting problem and to handle better the uncertainties associated with calibration techniques. Our approach was compared to Xu *et al.*'s disjoint approach, which independently calibrates the scores of SVM classifiers using the evidence theory and combines the obtained mass functions using Dempster's rule of combination. We compared also our proposed method to two approaches belonging to the trainable combiner category and based on an evidential classifier. In both cases, the obtained results for our evidential joint calibration based on logistic regression either are better or are comparable to that of the other approaches. Furthermore, by introducing the possibility to reject a test sample, we showed the advantages of the evidential multivariable logistic-based calibration over the probabilistic version: it models more precisely the uncertainties and it exhibits better performances.

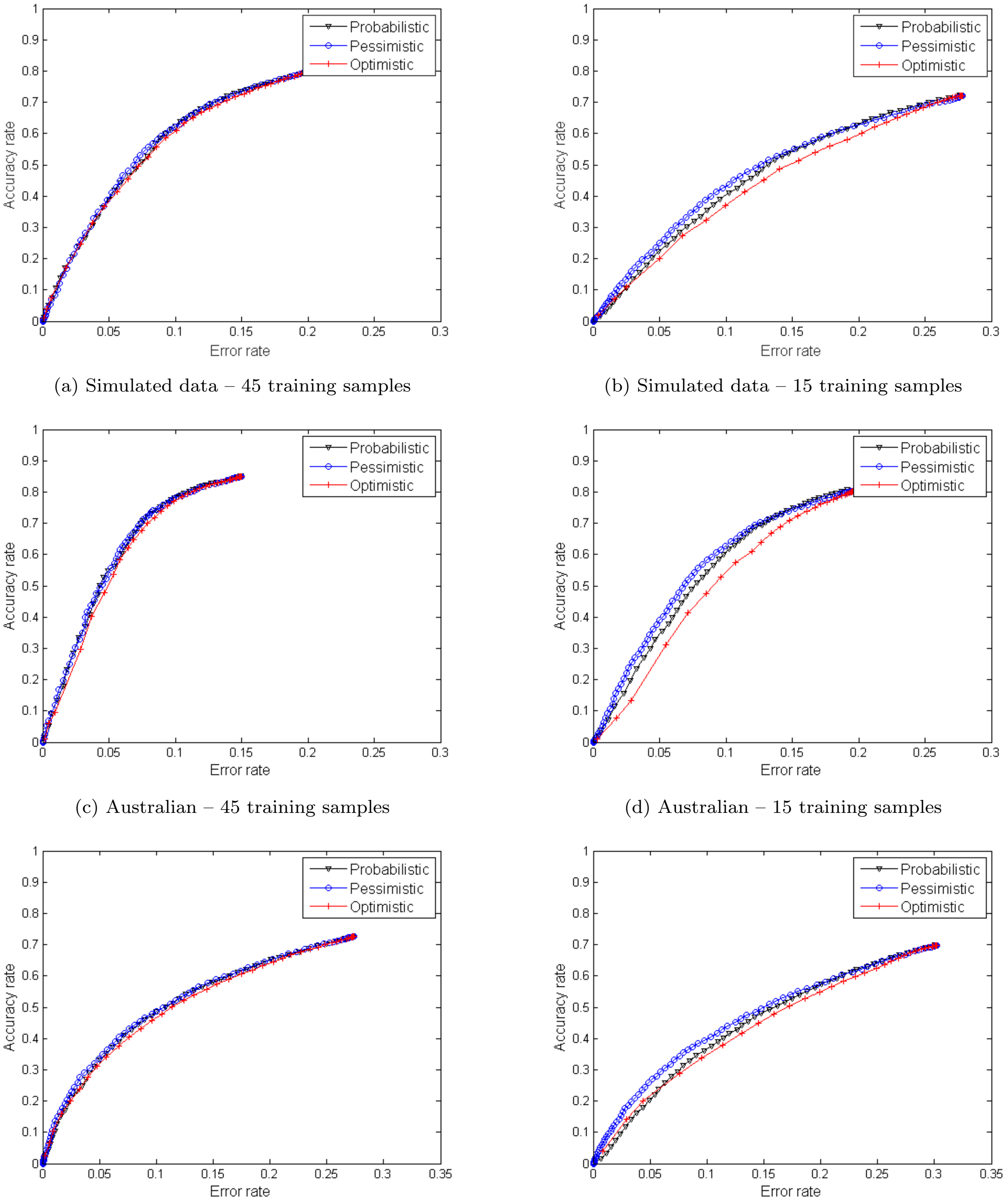
The approach presented in this paper was applied to the calibration of binary SVM classifiers, but they may also be applied to any other binary classifiers returning scores. As a matter of fact, future works include applying the evidential multivariable calibration to the face blurring application described in (Minary et al. 2016), which involves four different binary classifiers, and which was solved in (Minary et al. 2016) using the disjoint approach. The extension of the proposed evidential joint calibration to the multi-class problem may also be tackled in future works, following Xu et al. (2015) which addressed this extension in the single classifier case.

## Compliance with Ethical Standards

### Conflict of interest

Author Pauline Minary declares that she has no conflict of interest.

Author Frédéric Pichon declares that he has no conflict of interest.



(a) Simulated data – 45 training samples

(b) Simulated data – 15 training samples

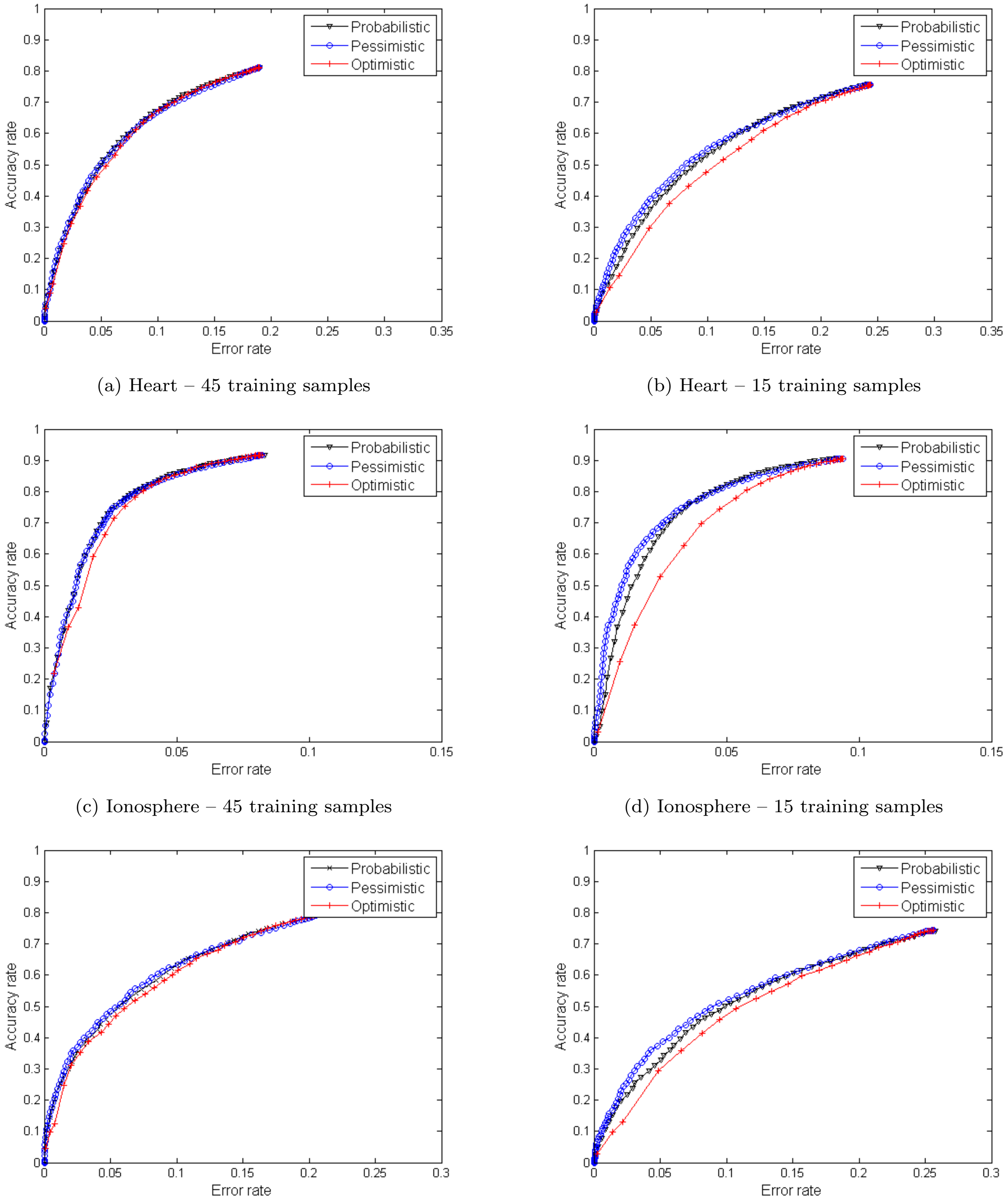
(c) Australian – 45 training samples

(d) Australian – 15 training samples

(e) Diabetes – 45 training samples

(f) Diabetes – 15 training samples

Fig. 9: Obtained error rates with 45 training samples (left) and 15 training samples (right) for the simulated dataset, *Australian* and *Diabetes*.



(a) Heart – 45 training samples

(b) Heart – 15 training samples

(c) Ionosphere – 45 training samples

(d) Ionosphere – 15 training samples

(e) Sonar – 45 training samples

(f) Sonar – 15 training samples

Fig. 10: Obtained error rates with 45 training samples (left) and 15 training samples (right) for *Heart*, *Ionosphere* and *Sonar*.

Author David Mercier declares that he has no conflict of interest.
Author Éric Lefèvre declares that he has no conflict of interest.
Author Benjamin Droit declares that he has no conflict of interest.

**Ethical approval**
This article does not contain any studies with human participants or animals performed by any of the authors.

## References

Bache K, Lichman M (2013) UCI machine learning repository. University of California, School of Information and Computer Sciences, http://archive.ics.uci.edu/ml

Bagley SC, White H, Golomb BA (2001) Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of clinical epidemiology 54(10):979–985

Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:27:1–27:27, software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Dempster A (1966) New methods for reasoning towards posterior distributions based on sample data. Annals of Mathematical Statistics 37(2):355–374

Dempster AP (1968) Upper and lower probabilities generated by a random closed interval. Annals of Mathematical Statistics pp 957–966

Denœux T (1995) A k-nearest neighbor classification rule based on dempster-shafer theory. IEEE transactions on Systems, Man, and Cybernetics 25(5):804–813

Denœux T (1997) Analysis of evidence-theoretic decision rules for pattern classification. Pattern Recognition 30(7):1095–1107

Denœux T (2014) Likelihood-based belief function: justification and some extensions to low-quality data. International Journal of Approximate Reasoning 55(7):1535–1547

Denœux T, Smets P (2006) Classification using belief functions: relationship between case-based and model-based approaches. IEEE Transactions on Systems, Man and Cybernetics B 36(6):1395–1406

Duin RPW (2002) The combining classifier: to train or not to train? In: Proceedings of the 16th International Conference on Pattern Recognition, Quebec City, Quebec, Canada, August, 2002, IEEE, vol 2, pp 765–770

Hosmer DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression, vol 398. Wiley & Sons

Kanjanatarakul O, Sriboonchitta S, Denœux T (2014) Forecasting using belief functions: an application to marketing econometrics. International Journal of Approximate Reasoning 55(5):1113–1128

Kanjanatarakul O, Denœux T, Sriboonchitta S (2016) Prediction of future observations using belief functions: A likelihood-based approach. International Journal of Approximate Reasoning 72:71–94

Kuncheva LI (2004) Combining pattern classifiers: methods and algorithms. John Wiley & Sons

Minary P, Pichon F, Mercier D, Lefevre E, Droit B (2016) An evidential pixel-based face blurring approach. In: Vejnarov J, Kratochvil V (eds) Proceedings of the Fourth International Conference on Belief Functions, Prague, Czech Republic, September 21-23, Springer, Lecture Notes in Computer Science, vol 9861, pp 222–230

Minary P, Pichon F, Mercier D, Lefevre E, Droit B (2017) Evidential joint calibration of binary svm classifiers using logistic regression. In: Proceedings of the 11th International Conference on Scalable Uncertainty Management, Granada, Spain, October 4-6, 2017, Lecture Notes in Artificial Intelligence, Springer, 7 pages

Minka TP (2003) Algorithms for maximum-likelihood logistic regression. Tech. Rep. 758, Carnegie Mellon University

Nguyen HT (2006) An Introduction to Random Sets. Chapman and Hall/CRC press

Platt JC (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10(3):61–74

Shafer G (1976) A mathematical theory of evidence, vol 1. Princeton University Press

Smets P (1993) Belief functions: the disjunctive rule of combination and the generalized Bayesian theorem. International Journal of Approximate Reasoning 9(1):1–35

Smets P, Kennes R (1994) The Transferable Belief Model. Artificial Intelligence 66:191–243

Tulyakov S, Jaeger S, Govindaraju V, Doermann D (2008) Review of classifier combination methods. In: Marinai S, Fujisawa H (eds) Machine Learning in Document Analysis and Recognition, Berlin, Germany, Springer, pp 361–386

Xu P, Davoine F, Denœux T (2015) Evidential multinomial logistic regression for multiclass classifier calibration. In: Proceedings of the 18th International Conference on Information Fusion, Washington, DC, USA, July 6-9, 2015, IEEE, pp 1106–1112

Xu P, Davoine F, Zha H, Denœux T (2016) Evidential calibration of binary SVM classifiers. International Journal of Approximate Reasoning 72:55–70

Zadrozny B, Elkan C (2001) Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann, pp 609–616

Zadrozny B, Elkan C (2002) Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2002, ACM, pp 694–699

Zhong W, Kwok JT (2013) Accurate probability calibration for multiple classifiers. In: Proceedings of the Twenty-Third international Joint Conference on Artificial Intelligence, Beijing, China, August, 2013, pp 1939–1945

Zouhal LM, Denœux T (1998) An evidence-theoretic k-nn rule with parameter optimization. IEEE Transactions on Systems, Man and Cybernetics C 28(2):263–271