Evidential calibration of binary SVM classifiers

Philippe Xu, Franck Davoine, Hongbin Zha, Thierry Denoeux

Evidential calibration of binary SVM classifiers

Philippe Xu[a], Franck Davoine[a], Hongbin Zha[b], Thierry Denœux[a]

[a]*Sorbonne universités, Université de technologie de Compiègne, CNRS, Heudiasyc, France*
[b]*Peking University, Key Laboratory of Machine Perception (MOE), Beijing, China*

Abstract

In machine learning problems, the availability of several classifiers trained on different data or features makes the combination of pattern classifiers of great interest. To combine distinct sources of information, it is necessary to represent the outputs of classifiers in a common space via a transformation called calibration. The most classical way is to use class membership probabilities. However, using a single probability measure may be insufficient to model the uncertainty induced by the calibration step, especially in the case of few training data. In this paper, we extend classical probabilistic calibration methods to the evidential framework. Experimental results from the calibration of SVM classifiers show the interest of using belief functions in classification problems.

Keywords: Classifier calibration, theory of belief functions, Dempster-Shafer theory, evidence theory, support vector machine

1. Introduction

The combination of pattern classifiers is an important issue in machine learning [25]. In many applications, the availability of several sources of information, coming from different sensors, training data, models or human experts, makes the use of combination techniques very attractive. Indeed, an ensemble of diverse classifiers often gives better results than do any of single one.

The combination strategies can generally be separated into two kinds: trainable and non-trainable combiners [18, 34]. In the first case, the outputs of each classifier are used as inputs in a new round of training. A simple way is to concatenate the initial outputs into a feature vector and learn a new classifier using classical machine learning techniques. Using trainable approaches is appealing as they may be asymptotically optimal [18]. In some cases, as in the bagging [4] and boosting [20] algorithms, the base classifiers are especially designed to be used in such learning schemes. However, one drawback of such methods is the additional training step. Not only do they need to have a training set common to all base classifiers, but new classifiers can hardly be added afterward. In

Email address: philippe.xu@hds.utc.fr (Philippe Xu)
This paper is a revised and extended version of [36].

many practical situations, new sensors can be included in the system and new sources of information or training data can become available after the classifiers have been trained. In such cases, we may not wish to train the whole combiner again every time.

The non-trainable approaches consist in combining directly the outputs of the base classifiers, using a pre-defined combination rule. The simplest way is the majority vote. It is one of the few cases where no particular attention has to be paid to the outputs of the base classifiers. Otherwise, the various outputs have to be made comparable beforehand. For this purpose, they are often transformed into class membership posterior probabilities [28, 37, 38]. This step is called calibration.

In this paper, we address the latter kind of combinations. In this context, the performance of each base classifier is not of primary importance, instead the calibration step becomes the major issue [7, 3]. One of the main difficulties in calibration is to avoid over-fitting, and this is especially critical when dealing with few training data. Classical probability theory can manage over-fitting only to a certain degree. However, more general theories [24] have been developed over the last decades to better handle cases with little information available. The theory of belief functions, also known as the Dempster-Shafer theory [30] or evidence theory, is a popular one in the information fusion community.

Many probabilistic learning algorithms such as naive Bayes, k-nearest neighbors, neural networks or decision trees have their evidential counterparts [10, 11, 33, 15, 19]. For kernel based methods, getting probabilistic outputs is more challenging. In the particular case of support vector machines (SVM), Bartlett and Tewari [2] showed that sparsity is lost if the posterior probabilities are to be estimated for all SVM scores. An additional training step, called calibration, as proposed by Platt [28], can give probabilistic outputs while retaining the sparseness of the SVM.

In this paper, we investigate the calibration of binary classifiers using belief functions. We extend three existing probabilistic calibration methods to the evidential framework and apply them to calibrate SVM classifiers. The rest of this paper is organized as follows. In Section 2, we review existing probabilistic calibration methods and discuss their limitations. An introduction to the theory of belief functions is given in Section 3. Evidential extensions of probabilistic calibration methods are then addressed in Section 4. Finally, some experimental results about the calibration of binary SVM classifiers are presented in Section 5.

2. Probabilistic calibration

Let $\mathcal{X} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be some training data in a binary classification problem, where $x_i \in \mathbb{R}$ is the score returned by a pre-trained classifier for the i-th training sample with label $y_i \in \{0, 1\}$. Given a test sample of score $s \in \mathbb{R}$ and unknown label $y \in \{0, 1\}$, the aim of calibration is to estimate the posterior class probability $P(y = 1|s)$. Several calibration methods can be found in the literature. Binning [37], isotonic regression [38] and logistic regression [28] are the most commonly used.

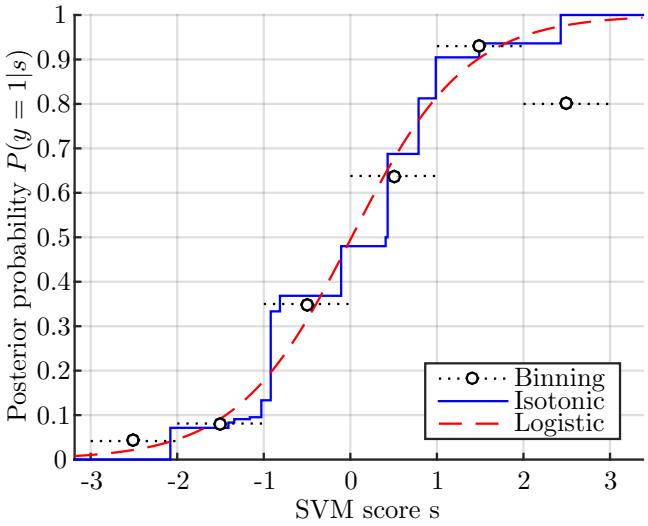


Figure 1: SVM scores calibration on the *Australian* dataset.

Binning [37] is a rather simple way to perform calibration by partitioning the score space into bins. For the j-th bin, which is an interval $[\underline{s}_j, \overline{s}_j]$, we count the number of positive examples k_j over all the n_j training examples whose score falls into this particular bin. Given a test sample of score $s \in [\underline{s}_j, \overline{s}_j]$ and unknown label $y \in \{0, 1\}$, the posterior probability $P\left(y = 1 | s \in [\underline{s}_j, \overline{s}_j]\right)$ is simply approximated by the empirical proportion $\hat{\tau}_j = k_j / n_j$.

Figure 1 illustrates the calibration results of an SVM classifier on the UCI[2] *Australian* dataset. This dataset concerns consumer credit applications. A set of 690 cases are described by 15 attributes. In our setting, the data are divided into a training set of size 390 and a test set of size 400. The white dots represent the outputs of a binning calibration with the following bins: $(-3, -2], (-2, -1], \ldots, (+2, +3]$. Calibration methods using the binning approach do not use any prior knowledge about the shape of the posterior probability with respect to the score. However, in many practical situations, the scores are seen as confidence measures. This implies that the transformation from a score to a probability measure should be done using a non-decreasing function. This assumption is strong, but it is often reasonable in practice.

An alternative to binning that incorporates such prior constraint is isotonic regression [38]. It consists in fitting a stepwise-constant, non-decreasing, i.e., isotonic, function $g : \mathbb{R} \to [0, 1]$ to the training data by minimizing the mean-

[2] http://archive.ics.uci.edu/ml.

squared error

$$MSE(g, \mathcal{X}) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \left[g\left(x_i\right) - y_i \right]^2 . \tag{1}$$

The optimal function $\hat{g}$ can be computed efficiently using the pair-adjacent violators (PAV) algorithm [1]. The solid line in Figure 1 shows the result of isotonic calibration.

Platt [28] further constrains the calibration problem using logistic regression. Niculescu-Mizil and Caruana [27] showed that logistic regression is well-adapted for calibrating maximum margin methods like SVM. Moreover, it is less prone to over-fitting as compared to binning and isotonic regression, especially when relatively few training data are available. Logistic regression calibration consists in fitting a sigmoid function

$$P\left(y = 1|s\right) \approx h_s(\theta) = \frac{1}{1 + \exp\left(\theta_0 + \theta_1 s\right)}. \tag{2}$$

The parameter $\theta = (\theta_0, \theta_1) \in \mathbb{R}^2$ of the sigmoid function is determined by maximizing the likelihood function on the training data,

$$L_{\mathcal{X}}(\theta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i} \quad \text{with} \quad p_i = \frac{1}{1 + \exp(\theta_0 + \theta_1 x_i)}. \tag{3}$$

To reduce over-fitting and prevent θ_0 from becoming infinite when the training examples are perfectly separable, Platt proposed to use an out-of-sample data model by replacing y_i and $1 - y_i$ by t_+ and t_- defined as

$$t_+ = \frac{n_+ + 1}{n_+ + 2} \quad \text{and} \quad t_- = \frac{1}{n_- + 2}, \tag{4}$$

where n_+ and n_- are respectively the number of positive and negative training samples. This ensures $L_{\mathcal{X}}$ to have a unique supremum $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1)$. The dashed curve in Figure 1 shows the result of logistic calibration.

One limitation of probabilistic calibration methods is that the uncertainty due to the number of training data is not taken into account. Figure 2 shows the empirical distribution of the SVM scores in a binning calibration framework. For a bin that contains many data, the estimate of the associated proportion is more certain. This can be illustrated by confidence intervals. We can see that for the $(+2, +3]$ bin, the confidence interval is larger, meaning that the proportion estimate is more uncertain. The same phenomenon occurs for isotonic and logistic regressions: the less training sample, the more uncertain the estimated parameters. In order to better handle such uncertainties, a more powerful representation needs to be used instead of probability.

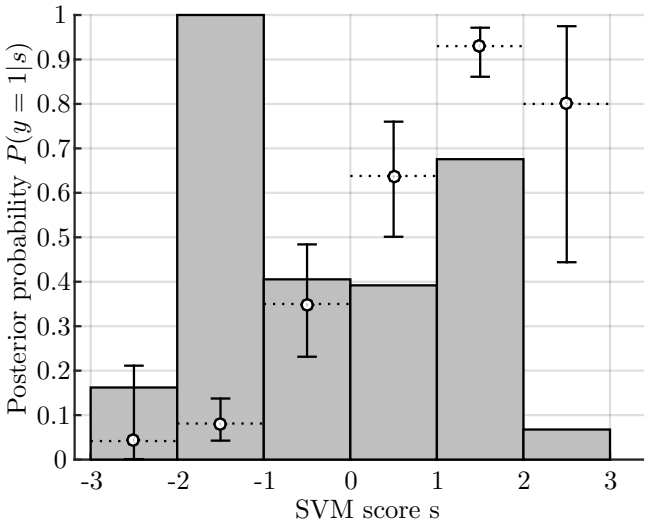


Figure 2: SVM scores distribution and confidence interval over the binomial proportion of each bin. The intervals are computed at the 95% confidence level.

3. Theory of belief functions

The theory of belief functions, also known as Dempster-Shafer theory [30], is a generalization of probability theory. It can be used for both prediction and statistical inference. In this section, we first introduce some basic notions of the theory of belief functions in a finite space. Then, we present the formulation of likelihood-based belief functions. Our presentation follows the work of Denœux [13] for statistical inference and the work of Kanjanatarakul et al. [23] for its applications to forecasting.

3.1. Predictive belief functions

Let $\Omega = \{\omega_1, \ldots, \omega_C\}$ be a finite set of classes. In the case of binary classification problems, we would have $\Omega = \{0, 1\}$. A *mass function* defined over the frame of discernment Ω is a function $m^\Omega : 2^\Omega \to [0, 1]$ verifying:

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1, \qquad m^\Omega(\emptyset) = 0. \tag{5}$$

Given an object whose class ω is in Ω, our belief about ω can be represented by a mass function over Ω. The quantity $m^\Omega(A)$, for a given subset $A \subseteq \Omega$, represents the belief committed to the hypothesis $\omega \in A$. The particular mass $m^\Omega(\Omega)$ represents the amount of ignorance. All subsets $A \subseteq \Omega$ such that $m^\Omega(A) > 0$ are called *focal elements*. A mass function whose focal elements are all singletons actually defines a probability distribution and will be referred to as Bayesian.

Given two mass functions m_1^Ω and m_2^Ω, generated from independent pieces of evidence, one can combine them into a new mass function $m_{1,2}^\Omega$, using Dempster's rule of combination:

$$m_{1,2}^\Omega(\emptyset) = 0, \qquad m_{1,2}^\Omega(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1^\Omega(B) m_2^\Omega(C), \tag{6}$$

where

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

represents the amount of conflict between the two mass functions. In this paper, the closed-world assumption is used [31], thus only normalized mass functions with no mass on the empty set are used. When the two mass functions are Bayesian, i.e., probability distributions, Dempster's rule becomes equivalent to Bayes' rule of combination with a uniform class prior distribution.

A mass function can also be represented by a *belief* or a *plausibility* function, defined, respectively, as:

$$Bel^\Omega(A) = \sum_{B \subseteq A} m^\Omega(B) \quad \text{and} \quad Pl^\Omega(A) = \sum_{B \cap A \neq \emptyset} m^\Omega(B), \quad \forall A \subseteq \Omega. \tag{8}$$

The degree of belief $Bel^\Omega(A)$ represents the amount of evidence strictly supporting the hypothesis $\omega \in A$, while the plausibility $Pl^\Omega(A) = 1 - Bel^\Omega\left(\overline{A}\right)$ is the amount of evidence not contradicting it. A function used to represent and assess the class of an object will be referred to as *predictive*.

There exist several ways to predict a label from a mass function. A popular way is to transform the mass function into a probability distribution and choose the singleton with maximum probability. This can be done using the pignistic probability $BetP^\Omega$ [32] defined as:

$$BetP^\Omega(\omega) = \sum_{A \subseteq \Omega | \omega \in A} \frac{m^\Omega(A)}{|A|}, \qquad \forall \omega \in \Omega. \tag{9}$$

The pignistic transformation consists in distributing the mass of any subset to its singletons uniformly. It is also possible to transform a probability distribution into a belief function by using the inverse pignistic transformation [17]. Another, simpler and computationally more efficient way to predict a label from a mass function is to choose the singleton with maximum plausibility. The plausibility over the singletons is defined by the contour function

$$pl^\Omega(\omega) \quad = \quad Pl^\Omega(\{\omega\}), \quad \forall \omega \in \Omega. \tag{10}$$

In the case of a binary classification problem, choosing the singleton with maximum mass, belief, plausibility or pignistic probability actually leads to the same decision.

3.2. Likelihood-based belief function

The theory of belief functions can also be used for statistical inference. Shafer [30] originally proposed to use a "likelihood-based" belief function for statistical inference. This approach was further justified by Denœux [13, 14]. Knowledge about some parameters can then be used for prediction as in [23].

Let $X \in \mathbb{X}$ be some observable data and $\theta \in \Theta$ the unknown parameter of the density function $f_\theta(x)$ generating the data. Information about θ can be inferred given the outcome x of a random experiment. Shafer [30] proposed to build a belief function Bel_x^Θ from the likelihood function. After observing $X = x$, the likelihood function $L_x : \theta \mapsto f_\theta(x)$ is normalized to yield the following contour function:

$$pl_x^\Theta(\theta) \quad = \quad \frac{L_x(\theta)}{\sup_{\theta' \in \Theta} L_x(\theta')}, \quad \forall \theta \in \Theta, \tag{11}$$

where sup denotes the supremum operator. The consonant plausibility function associated to this contour function is

$$Pl_x^\Theta(A) \quad = \quad \sup_{\theta \in A} pl_x^\Theta(\theta), \quad \forall A \subseteq \Omega. \tag{12}$$

The focal sets of Bel_x^Θ are defined as

$$\Gamma_x(\gamma) \quad = \quad \{\theta \in \Theta \mid pl_x^\Theta(\theta) \geq \gamma\}, \quad \forall \gamma \in [0,1]. \tag{13}$$

To manipulate belief functions defined over a continuous space, the random sets formalism [26] is often used. Given the Lebesgue measure λ on $[0,1]$ and the multi-valued mapping $\Gamma_x : [0,1] \to 2^\Theta$, we have

$$\begin{array}{rcl} Bel_x^\Theta(A) & = & \lambda\left(\{\gamma \in [0,1] \mid \Gamma_x(\gamma) \subseteq A\}\right) \\ Pl_x^\Theta(A) & = & \lambda\left(\{\gamma \in [0,1] \mid \Gamma_x(\gamma) \cap A \neq \emptyset\}\right) \end{array}, \qquad \forall A \subseteq \Theta. \tag{14}$$

A complete description of the theory of random sets and its relation to the theory of belief functions can be found in [26].

3.3. Forecasting

Suppose that we now have some knowledge about the parameter $\theta \in \Theta$ after observing some training data x. The *forecasting* problem consists in making some predictions about some random quantity $Y \in \mathbb{Y}$ whose conditional distribution $g_{x,\theta}(y)$ given $X = x$ depends on θ. A belief function on $\mathbb{Y}$ can be derived from the sampling model proposed by Dempster [8]. For some unobserved auxiliary variable $Z \in \mathbb{Z}$ with known probability distribution μ independent of θ, we define a function $\varphi : \Theta \times \mathbb{Z} \to \mathbb{Y}$ so that

$$Y \quad = \quad \varphi(\theta, Z). \tag{15}$$

A multi-valued mapping $\Gamma'_x : [0,1] \times \mathbb{Z} \to 2^{\mathbb{Y}}$ is defined by composing Γ_x with φ

$$\begin{array}{rccc} \Gamma'_x : & [0,1] \times \mathbb{Z} & \to & 2^{\mathbb{Y}} \\ & (\gamma, z) & \mapsto & \varphi(\Gamma_x(\gamma), z). \end{array} \tag{16}$$

A belief function on $\mathbb{Y}$ can then be derived from the product measure $\lambda \otimes \mu$ on $[0,1] \times \mathbb{Z}$ and the multi-valued mapping Γ'_x

$$Bel_x^{\mathbb{Y}}(A) = (\lambda \otimes \mu)\left(\{(\gamma, z) \mid \varphi\left(\Gamma_x\left(\gamma\right), z\right) \subseteq A\}\right), \tag{17a}$$

$$Pl_x^{\mathbb{Y}}(A) = (\lambda \otimes \mu)\left(\{(\gamma, z) \mid \varphi\left(\Gamma_x\left(\gamma\right), z\right) \cap A \neq \emptyset\}\right), \tag{17b}$$

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

3.4. Binary case example

Let $Y \in \mathbb{Y}$ be a binary random variable, i.e., $\mathbb{Y} = \{0, 1\}$. Let $\tau \in T$, where $T = [0, 1]$, be the proportion parameter of its associated Bernoulli distribution $\mathcal{B}(\tau)$. The random variable Y can be generated by a function φ defined as

$$Y = \varphi(\tau, Z) = \begin{cases} 1 & \text{if } Z \leq \tau, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where Z has a uniform distribution on $[0, 1]$. Suppose we have some information about the parameter τ given under the form of a belief function Bel_x^T that is induced by a random closed interval $\Gamma_x(\gamma) = [U(\gamma), V(\gamma)]$. In particular, it is the case if Bel_x^T is the consonant belief function associated to a unimodal contour function pl_x^T. We get

$$\Gamma'_x(\gamma, z) = \varphi\left([U(\gamma), V(\gamma)], z\right) \quad = \quad \begin{cases} \{1\} & \text{if } Z \leq U(\gamma), \\ \{0\} & \text{if } Z > V(\gamma), \\ \{0, 1\} & \text{otherwise.} \end{cases} \tag{19}$$

The *predictive* belief function $Bel_x^{\mathbb{Y}}$ can then be computed as

$$Bel_x^{\mathbb{Y}}(\{1\}) = (\lambda \otimes \mu)(\{(\gamma, z) \mid Z \leq U(\gamma)\}) \tag{20a}$$

$$= \int_0^1 \mu(\{z \mid z \leq U(\gamma)\}) f(\gamma) d\gamma \tag{20b}$$

$$= \int_0^1 U(\gamma) f(\gamma) d\gamma = \mathbb{E}(U) \tag{20c}$$

and

$$Bel_x^{\mathbb{Y}}(\{0\}) = (\lambda \otimes \mu)(\{(\gamma, z) \mid Z > V(\gamma)\}) \tag{21a}$$

$$= 1 - (\lambda \otimes \mu)(\{(\gamma, z) \mid Z \leq V(\gamma)\}) \tag{21b}$$

$$= 1 - \mathbb{E}(V). \tag{21c}$$

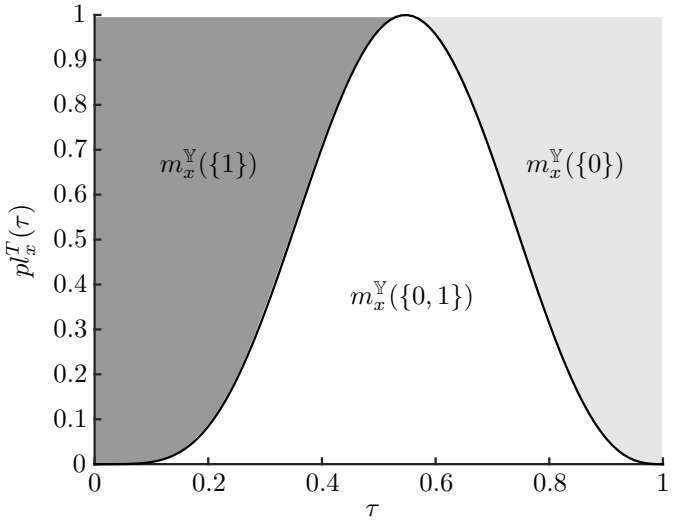


Figure 3: Predictive mass function $m_x^{\mathbb{Y}}$ based on the contour function pl_x^T.

As U and V take only non-negative values, these quantities have the following expressions:

$$Bel_x^{\mathbb{Y}}(\{1\}) = \int_0^{+\infty} (1 - F_U(u))du \tag{22a}$$

$$= \int_0^{\hat{\tau}} (1 - pl_x^T(u))du \tag{22b}$$

$$= \hat{\tau} - \int_0^{\hat{\tau}} pl_x^T(u)du \tag{22c}$$

and

$$Pl_x^{\mathbb{Y}}(\{1\}) = 1 - Bel_x^{\mathbb{Y}}(\{0\}) \tag{23a}$$

$$= \int_0^{+\infty} (1 - F_V(v))dv \tag{23b}$$

$$= \hat{\tau} + \int_{\hat{\tau}}^{1} pl_x^T(v)dv, \tag{23c}$$

where $\hat{\tau}$ is the value maximizing pl_x^T. In many practical situations, the belief function $Bel_x^{\mathbb{Y}}$ cannot be expressed analytically. However, they can be approximated either by Monte Carlo simulation using Eqs. (20) and (21) or by numerically estimating the integrals of Eqs. (22) and (23). The predictive mass function $m_x^{\mathbb{Y}}$ can be represented by the areas of regions delimited by the contour function pl_x^T, as shown in Figure 3.

4. Evidential calibration

In this section, we extend the three probabilistic calibration methods presented in Section 2 to the evidential framework. Instead of estimating a posterior probability distribution $P(\cdot|s)$, we now aim at estimating a mass function $m_s^{\mathbb{Y}}$. As opposed to the Bayesian approach where there is single probability of success $P(y=1|s)$, the evidential method returns two values: the belief $Bel_s^{\mathbb{Y}}(\{1\})$ and the plausibility $Pl_s^{\mathbb{Y}}(\{1\})$.

4.1. Binning

Binning can be formulated as a binomial proportion estimation problem. For a bin $[\underline{s}_j, \overline{s}_j]$, we are given n_j trials with k_j successes, and the goal is to estimate the associated unknown binomial proportion $\tau_j \in [0,1]$. One simple way to get a predictive belief function in such a configuration is to use Dempster's model [9], which leads to the following mass function:

$$m_{\text{D}}^{\mathbb{Y}}(\{1\}) = \frac{k_j}{n_j+1}, \quad m_{\text{D}}^{\mathbb{Y}}(\{0\}) = \frac{n_j-k_j}{n_j+1}, \quad m_{\text{D}}^{\mathbb{Y}}(\{0,1\}) = \frac{1}{n_j+1}. \tag{24}$$

It can be interpreted, similarly to the Laplace estimator, as having observed one sample prior to the trial but with unknown label.

From a statistical inference point of view, confidence intervals are often used to better model the uncertainty due to a small sample size. A confidence interval $[\underline{\tau}_j, \overline{\tau}_j]$ at confidence level $1-\alpha \in [0,1]$, i.e., $P\left(\underline{\tau}_j \leq \tau_j \leq \overline{\tau}_j\right) = 1-\alpha$, can be represented by the following contour function defined over T:

$$pl_{\text{CI}}^T(\tau_j) \quad = \quad \begin{cases} 1 & \text{if } \underline{\tau}_j \leq \tau_j \leq \overline{\tau}_j, \\ \alpha & \text{otherwise.} \end{cases} \tag{25}$$

This contour function can be used within the Eqs. (22) and (23) to derive the associated predictive mass function, which is defined as

$$\begin{aligned} m_{\text{CI}}^{\mathbb{Y}}(\{1\}) &= Bel_{\text{CI}}^{\mathbb{Y}}(\{1\}) && \text{(26a)} \\ &= \underline{\tau}_j - \int_0^{\underline{\tau}_j} pl_{\text{CI}}^T(\tau) d\tau && \text{(26b)} \\ &= (1-\alpha)\underline{\tau}_j, && \text{(26c)} \end{aligned}$$

and

$$\begin{aligned} m_{\text{CI}}^{\mathbb{Y}}(\{0\}) &= 1 - Pl_{\text{CI}}^{\mathbb{Y}}(\{1\}) && \text{(27a)} \\ &= 1 - \overline{\tau}_j - \int_{\overline{\tau}_j}^1 pl_{\text{CI}}^T(\tau) d\tau && \text{(27b)} \\ &= (1-\alpha)(1-\overline{\tau}_j),. && \text{(27c)} \end{aligned}$$

For the Clopper-Pearson interval [6], the bounds are defined as

$$\underline{\tau}_j = B\left(\frac{\alpha}{2}; k_j, n_j - k_j + 1\right), \quad \overline{\tau}_j = B\left(1 - \frac{\alpha}{2}; k_j + 1, n_j - k_j\right), \tag{28}$$

where $B(q; \beta, \gamma)$ is the q-th quantile of a beta distribution with shape parameters β and γ. The choice of the confidence level is often arbitrary, a confidence level of 95% is a common one. The mass function defined in Eqs. (26) and (27) is similar to the one proposed in [12] but discounted by a factor α.

An alternative to confidence intervals is the use of the likelihood function as proposed by Denœux [13]. If the relative likelihood function is used as contour function for τ_j, we get

$$pl_{\text{L}}^{T}(\tau_j) \quad = \quad \frac{\tau_j^{k_j}(1-\tau_j)^{n_j-k_j}}{\hat{\tau}_j^{k_j}(1-\hat{\tau}_j)^{n_j-k_j}}, \tag{29}$$

which gives the following predictive mass function:

$$m_{\text{L}}^{\mathbb{Y}}(\{1\}) = \begin{cases} 0 & \text{if } \hat{\tau}_j = 0, \\ \hat{\tau}_j - \dfrac{\underline{B}(\hat{\tau}_j; k_j + 1, n_j - k_j + 1)}{\hat{\tau}_j^{k_j}(1-\hat{\tau}_j)^{n_j-k_j}} & \text{if } 0 < \hat{\tau}_j < 1, \\ \dfrac{n_j}{n_j+1} & \text{if } \hat{\tau}_j = 1, \end{cases} \tag{30a}$$

$$m_{\text{L}}^{\mathbb{Y}}(\{0\}) = \begin{cases} \dfrac{n_j}{n_j+1} & \text{if } \hat{\tau}_j = 0, \\ 1 - \hat{\tau}_j - \dfrac{\overline{B}(\hat{\tau}_j; k_j + 1, n_j - k_j + 1)}{\hat{\tau}_j^{k_j}(1-\hat{\tau}_j)^{n_j-k_j}} & \text{if } 0 < \hat{\tau}_j < 1, \\ 0 & \text{if } \hat{\tau}_j = 1, \end{cases} \tag{30b}$$

where $\underline{B}$ and $\overline{B}$ are, respectively, the lower and upper incomplete beta functions defined as

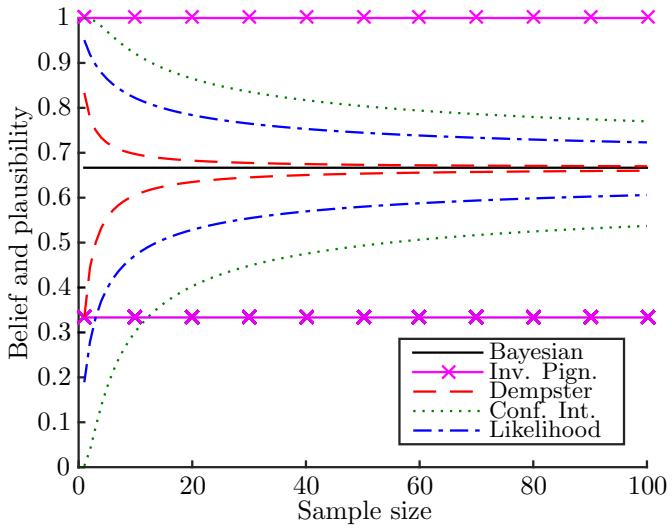
$$\underline{B}(z; a, b) = \int_0^z t^{a-1}(1-t)^{b-1}dt, \tag{31a}$$

$$\overline{B}(z; a, b) = \int_z^1 t^{a-1}(1-t)^{b-1}dt = \underline{B}(1-z; b, a). \tag{31b}$$

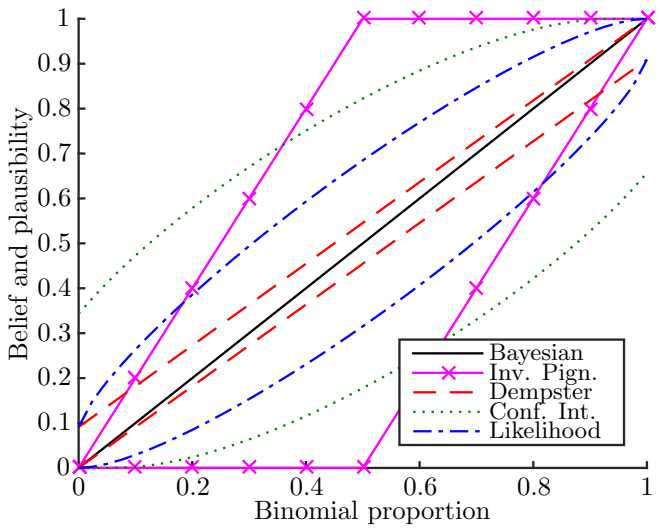
They can be computed exactly for integer values of a and b as

$$\underline{B}(z; a, b) = \sum_{j=a}^{a+b-1} \frac{(a-1)!(b-1)!}{j!(a+b-1-j)!} z^j (1-z)^{a+b-1-j}. \tag{32}$$

Figure 4 illustrates the belief and plausibility of success obtained with different sample sizes and binomial proportions. In Figure 4a, we can see that when the sample size grows, Dempster's model converges very rapidly to the Bayesian estimate. In contrast, the Clopper-Pearson interval-based model is more con-



(a)



(b)

Figure 4: (a) Belief and plausibility of success given a proportion of 2/3 w.r.t. the sample size. (b) Belief and plausibility of success given a sample of size 10 w.r.t. the binomial proportion. The Clopper-Pearson confidence interval was computed with a confidence level of 95%.

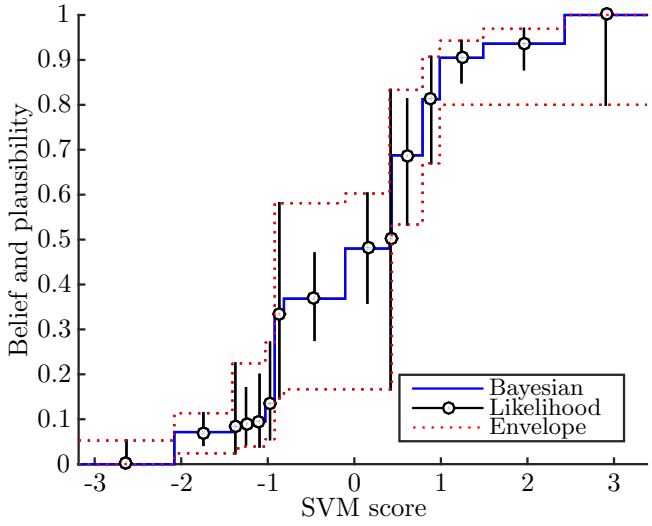


Figure 5: Isotonic regression

servative. The likelihood-based model gives intermediate results. The Bayesian estimator and its associated inverse pignistic transform do not take into account the sample sizes. It can be noted that using the Laplace estimator instead of the Bayesian estimator would yield different estimates for different sample size. However, any given binomial proportion estimate can still be generated from an infinite number of configurations. In Figure 4b, it is interesting to note that, when the empirical proportion $\hat{\tau}$ is equal to 0 or 1, the likelihood-based model yields the same result as Dempster's model.

4.2. Isotonic regression

The calibration result from isotonic regression can be seen as a particular case of binning. All the previous methods can thus be used. In Figure 5, the vertical segments depict the likelihood-based interval for each bin defined by the isotonic regression. We can see, however, that the lower and upper envelopes defined by these intervals are not isotonic. A simple way to get an isotonic envelope is to first scan the bins in increasing order and keep the highest upper bound computed so far to define the upper envelope. Then we scan the bins in decreasing order and keep the lowest lower bound to define the lower envelope. The doted curves in Figure 5 illustrate the obtained belief and plausibility functions.

4.3. Logistic regression

For logistic regression, the parameter $\theta \in \mathbb{R}^2$ of the sigmoid function h, defined in Eq. (2), needs to be estimated. After observing the score s of a test sample, its label $y \in \{0, 1\}$ can be seen as the realization of a random variable

Y with a Bernoulli distribution $\mathcal{B}(\tau)$, where $\tau = h_s(\theta) \in [0,1]$. By formulating the logistic regression as a generalized linear model [21], normal approximation intervals can be used to compute a confidence interval over $h_s(\hat{\theta})$ for any score s. The predictive mass function (26) can then be used.

From a likelihood point of view, the training data $\mathcal{X}$ generates the likelihood function $L_{\mathcal{X}}$, which can be used to define a plausibility function $Pl^{\Theta}_{\mathcal{X}}$ over the parameter $\theta \in \Theta$ as follows:

$$Pl^{\Theta}_{\mathcal{X}}(A) = \sup_{\theta \in A} pl^{\Theta}_{\mathcal{X}}(\theta), \quad \forall A \subseteq \Theta, \tag{33}$$

where

$$pl^{\Theta}_{\mathcal{X}}(\theta) = \frac{L_{\mathcal{X}}(\theta)}{L_{\mathcal{X}}(\hat{\theta})}, \quad \forall \theta \in \Theta. \tag{34}$$

As described in Section 3.4, a predictive belief function $Bel^{\mathbb{Y}}_{\mathcal{X},s}$ can be derived from the contour function $pl^{T}_{\mathcal{X},s}$. The function $pl^{T}_{\mathcal{X},s}$ can be computed from $Pl^{\Theta}_{\mathcal{X}}$ as

$$pl^{T}_{\mathcal{X},s}(\tau) = \begin{cases} 0 & \text{if } \tau \in \{0,1\} \\ Pl^{\Theta}_{\mathcal{X}}\left(\{\theta \in \Theta \mid \tau = h_s(\theta)\}\right) & \text{otherwise,} \end{cases} \tag{35}$$

$$= \begin{cases} 0 & \text{if } \tau \in \{0,1\} \\ Pl^{\Theta}_{\mathcal{X}}\left(h_s^{-1}(\tau)\right) & \text{otherwise,} \end{cases} \tag{36}$$

where

$$h_s^{-1}(\tau) = \left\{(\theta_0,\theta_1) \in \Theta \;\middle|\; \frac{1}{1+\exp(\theta_0+\theta_1 s)} = \tau \right\} \tag{37}$$

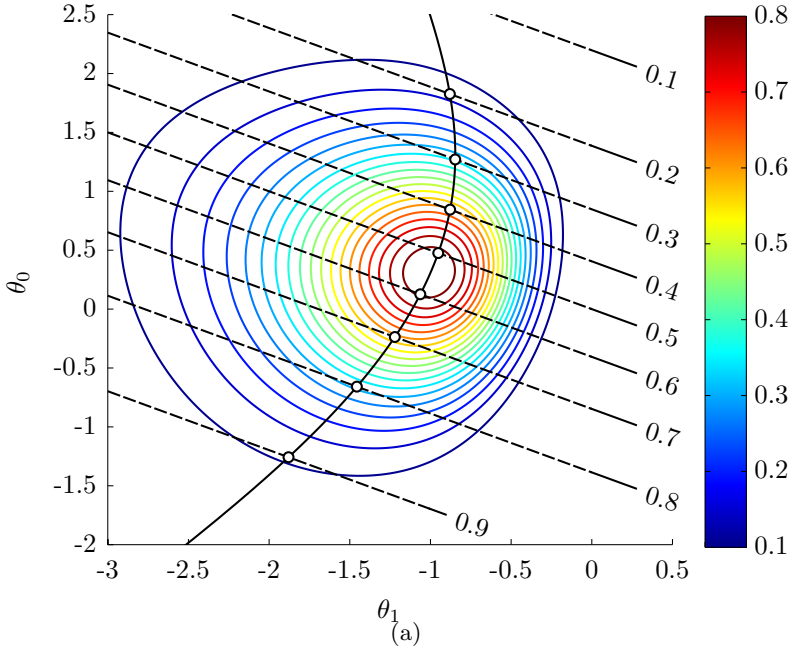
$$= \left\{(\theta_0,\theta_1) \in \Theta \;\middle|\; \exp(\theta_0+\theta_1 s) = \tau^{-1} - 1 \right\} \tag{38}$$

$$= \left\{(\theta_0,\theta_1) \in \Theta \;\middle|\; \theta_0 = \ln\left(\tau^{-1}-1\right) - \theta_1 s \right\}, \tag{39}$$

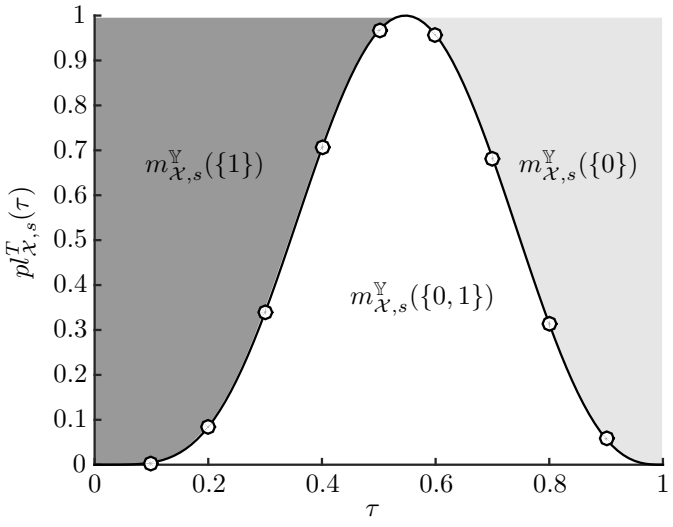
which finally yields

$$pl^{T}_{\mathcal{X},s}(\tau) = \sup_{\theta_1 \in \mathbb{R}} pl^{\Theta}_{\mathcal{X}}\left(\ln\left(\tau^{-1}-1\right) - \theta_1 s, \theta_1\right), \quad \forall \tau \in (0,1). \tag{40}$$

Figure 6 illustrates the computation of the predictive belief function $Bel^{\mathbb{Y}}_{x,s}$. Figure 6a shows level sets of the contour function $pl^{\Theta}_{\mathcal{X}}$ computed from the scores of an SVM classifier trained on the *Australian* dataset. The value of $pl^{T}_{\mathcal{X},s}(\tau)$ is defined as the maximum value of $pl^{\Theta}_{\mathcal{X}}$ along the dashed line $\theta_0 = \ln\left(\tau^{-1}-1\right) - \theta_1 s$. It can be approximated by an iterative maximization algorithm. Figure 6b shows the contour function $pl^{T}_{\mathcal{X},s}$ from which the predictive mass function $m^{\mathbb{Y}}_{\mathcal{X},s}$ can be computed using Eqs. (22) and (23). The calibration results using a training dataset of size 20 and 200 are shown in Figure 7.

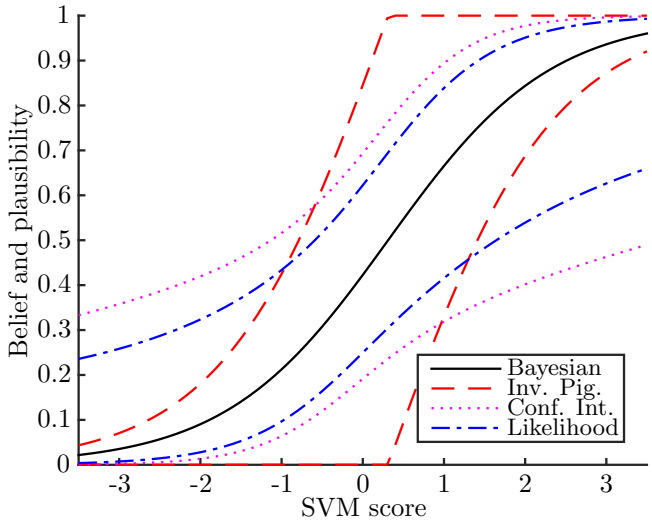


(a)

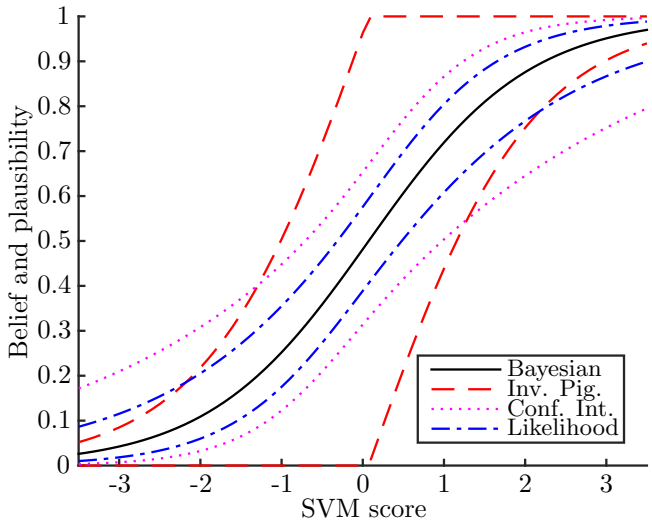


(b)

Figure 6: (a) Level sets of the contour function $pl^{\Theta}_{\mathcal{X}}$. (b) Contour function $pl^{T}_{\mathcal{X},s}$ with $s = 0.5$ and its associated predictive mass function $m^{\mathbb{Y}}_{\mathcal{X},s}$.



(a) Number of training data: 20



(b) Number of training data: 200

Figure 7: Evidential logistic regression calibration results. The confidence interval-based model was computed with a confidence level of 95%.

Table 1: Number of training data used for training and testing on different datasets from UCI.

Dataset	Train #1	Train #2	Train #3	Test
Australian	30	70	10–190	400
diabetes	30	70	10–200	468
heart	20	40	10–140	70
ionosphere	20	40	10–190	101
liver-disorders	20	40	10–190	95
sonar	20	40	10–90	58

5. Experimental evaluations

Experimental evaluations were conducted on several binary classification problems from the UCI repository. We first compared the classification results of the combination of three SVM classifiers using different calibration strategies with probabilistic and evidential models. Then, we focused on the logistic regression-based calibration method with the combination of ten SVM classifiers. The source code of all the calibration methods is available on the author's website[3]

5.1. Combination of three classifiers

The data used to compare the combination results are described in Table 1. For each dataset, three independent classifiers were trained on non-overlapping subsets of different sizes. For two of them, we fixed the number of examples and we considered different training sample sizes for the third one. For each experiment, the training data were partitioned into two subsets of equal size. The first subset was used to train the base SVM classifier and the other one for calibration. The whole process was repeated for 100 rounds of random partitioning. The LibSVM[4] library [5] was used to learn the base SVM classifiers.

Figure 8 shows the combination results using the binning strategy. We compared the Bayesian model, along with its inverse pignistic transform, to evidential ones using Dempster's model, confidence interval and likelihood-based approaches. We can see that, when the third classifier was trained with much more data than the two others, the differences in performance were larger. In contrast, when the three classifiers were trained with about the same amount of training data, the five models all led to relatively close results. The Bayesian approach and its inverse pignistic transform always performed worse when the third classifier was trained with much more data. Dempster's model and the likelihood-based one yielded very similar results. The performance of the confidence interval-based approach varied for the different datasets. Compared to the two other evidential approaches, it performed worse on the *heart* and *sonar*

[3] https://www.hds.utc.fr/~xuphilip/dokuwiki/en/data.
[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm.

datasets, similarly on the *Australian* and *ionosphere* datasets, and better on the *diabetes* and *liver-disorders* datasets.

Figures 9 and 10 show the results obtained using isotonic and logistic regressions, respectively. For these two calibration strategies, the performances of the different models were very similar. For the isotonic regression method, we can still see that the Bayesian and inverse pignistic models performed slightly worse than the evidential ones, except on the *diabetes* dataset, for which the confidence interval-based method performed worse when the third classifier was trained with relatively few training data. For the logistic regression approach, the four models led to almost similar results. In contrast with binning and isotonic regression, the Bayesian and inverse pignistic models performed as well as the evidential ones. Overall, the likelihood-based model returned the most stable and satisfactory results.

We then compared the three calibration strategies using the likelihood-based model. The results are shown in Figure 11. The best performances were obtained either by the logistic regression method or by the isotonic regression one. For the *diabetes* and *liver-disorders* datasets, the isotonic regression approach was significantly better than the two others. For the *Australian* and *heart* datasets, the logistic regression method performed the best. Finally, for the *ionosphere* and *sonar* datasets, the isotonic and logistic regression approaches had similar results. Moreover, they were both significantly better than the binning method on the *sonar* dataset.

5.2. Combination of ten classifiers

In order to emphasize the contribution of evidential calibration methods, we conducted an additional experiment in which ten classifiers were combined. The training dataset was now partitioned into ten subsets. Three scenarios were considered, as illustrated in Figure 12. In the first scenario (a), all ten classifiers were trained using the same amount of training data. In the second one (b), one half of the classifiers were trained with five times more data than the other half. Finally, in (c), one classifier was trained with $2/3^{\text{rd}}$ of the data, a second one used $1/5^{\text{th}}$ and the eight other ones shared the rest uniformly.

From the datasets used in the previous section, we only kept *Australian* and *diabetes* as they were the only ones that were large enough to train ten classifiers. And we furthermore added the *adult* dataset which was used in [28, 38].

To compare the performances of the different models, the significance of the results was evaluated from a McNemar test [16] at the 5% level. The results are shown in Table 2. The best results were always obtained by the likelihood-based model except for *adult* (a). In particular, except for the inverse pignistic transformation on the *Australian* dataset, the results were always significantly better for scenario (c). For the *adult* dataset, the likelihood-based model always gave significantly better results than the Bayesian model. We can see that the likelihood-based approach is more robust when the training sets have highly unbalanced sizes.

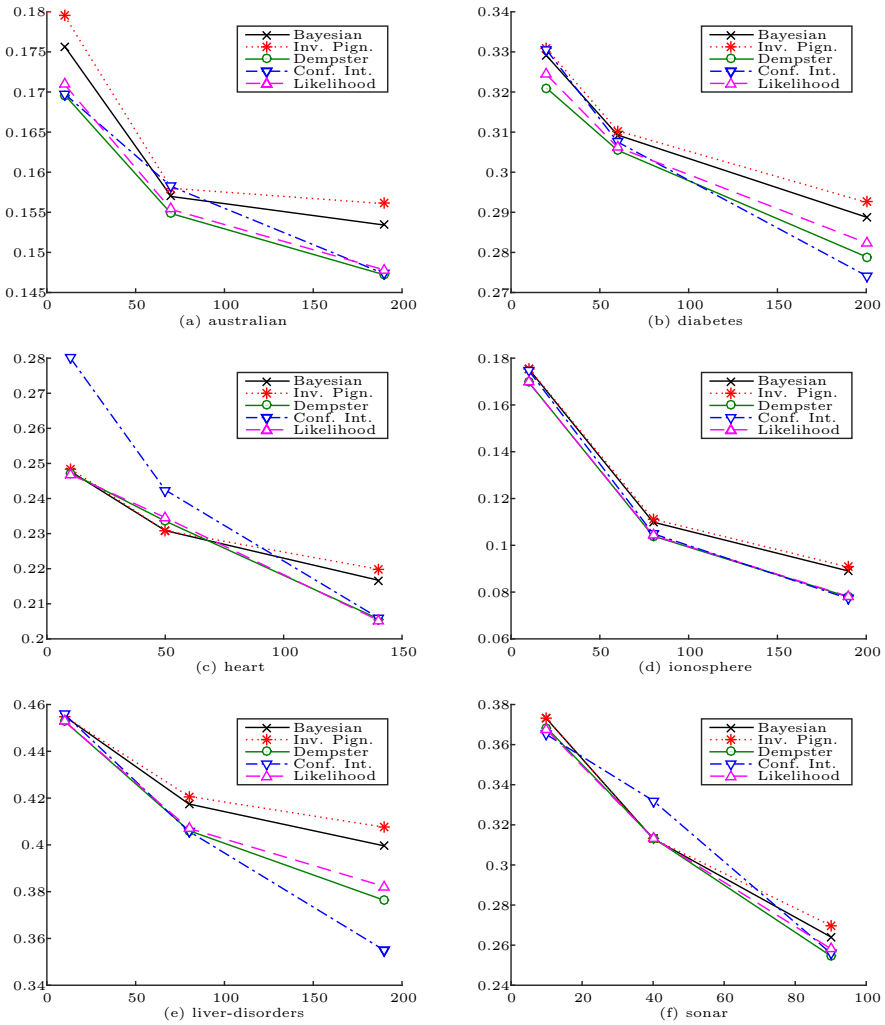


Figure 8: Classification results using binning calibration. The x-axis corresponds to the number of training data used to learn the third classifier. The y-axis corresponds to the average error rate.

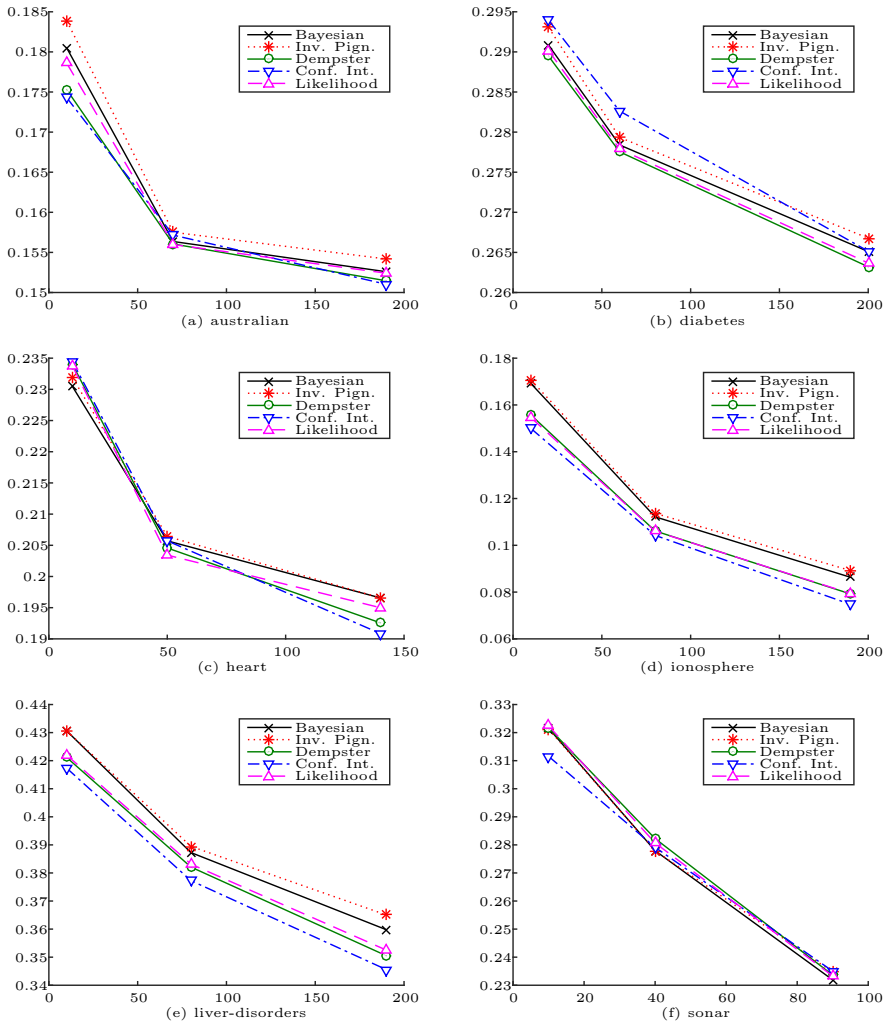


Figure 9: Classification results using isotonic regression calibration. The x-axis corresponds to the number of training data used to learn the third classifier. The y-axis corresponds to the average error rate.

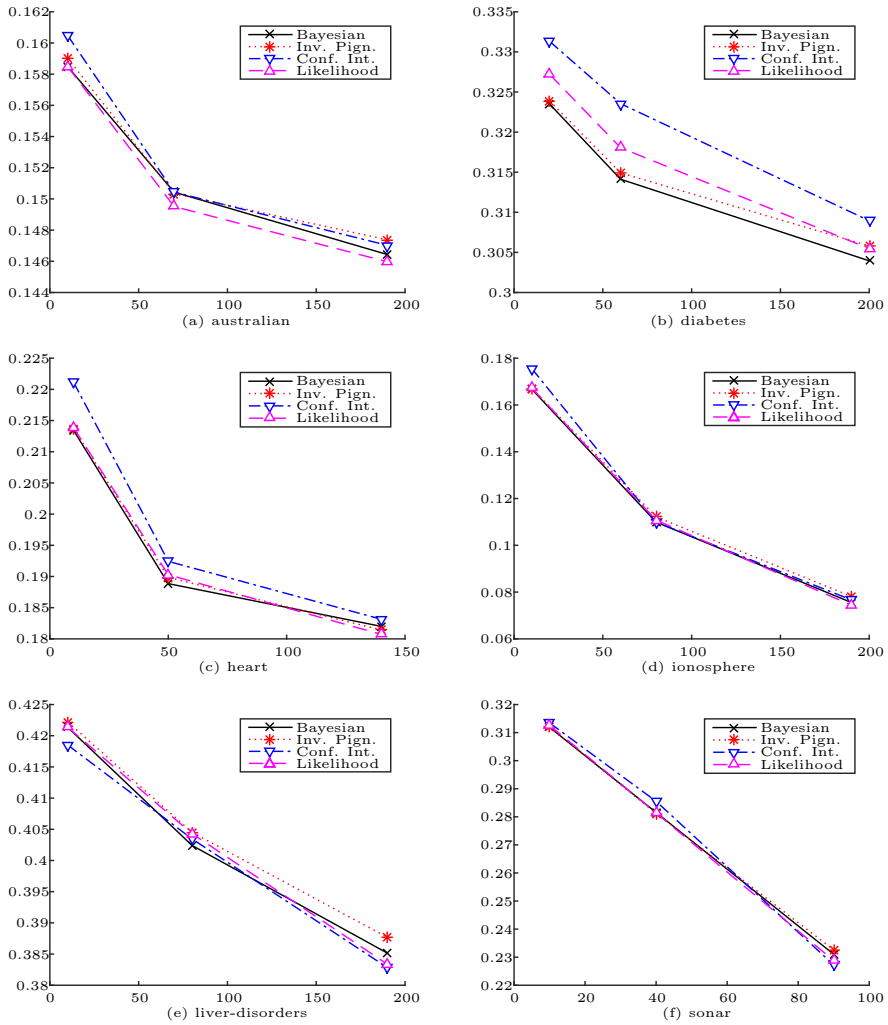


Figure 10: Classification results using logistic regression calibration. The x-axis corresponds to the number of training data used to learn the third classifier. The y-axis corresponds to the average error rate.

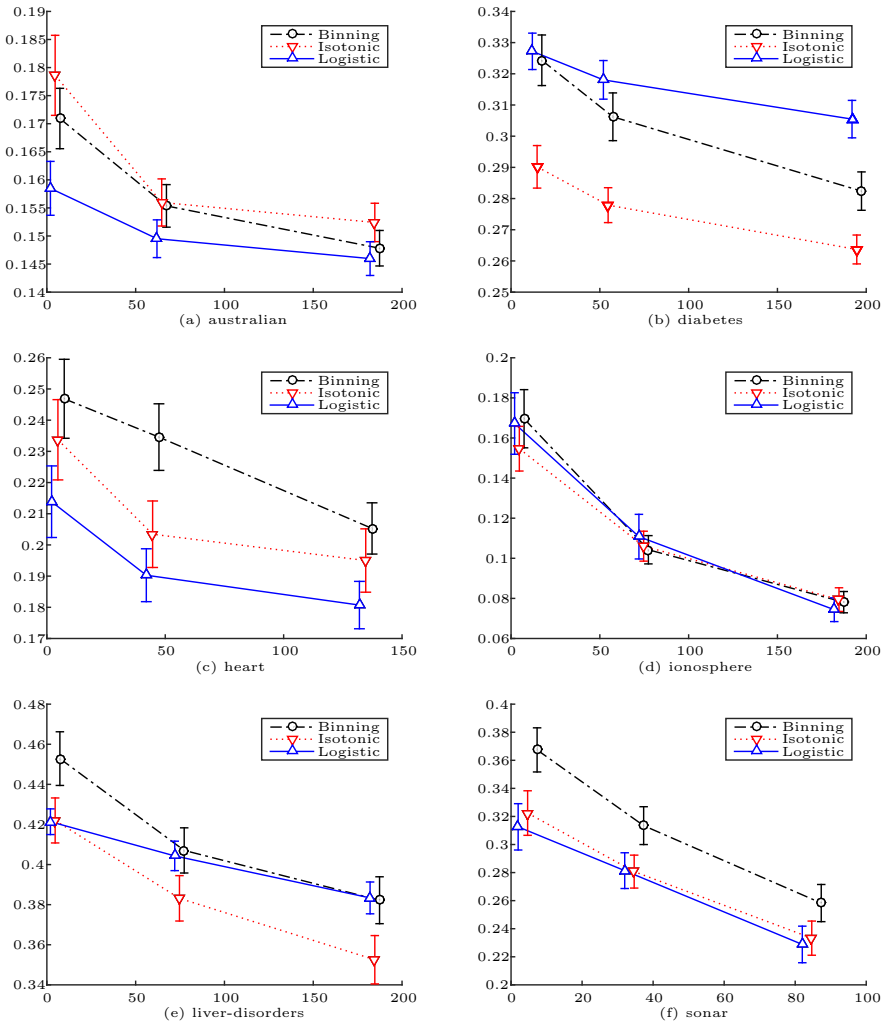


Figure 11: Classification results using the likelihood-based model. The vertical segments correspond to the confidence interval of the average error rate at a 95% confidence level.

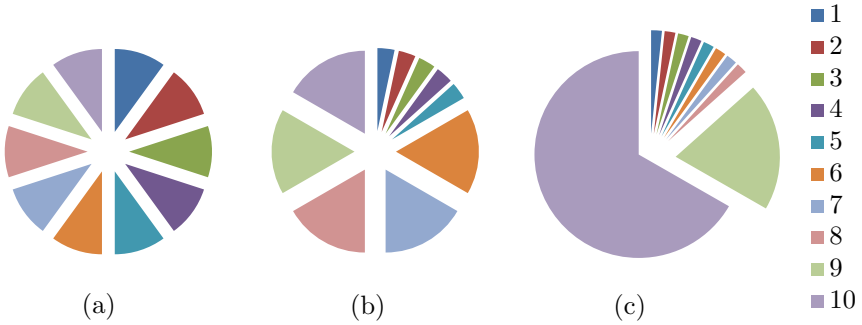


Figure 12: Proportions of data used to train each of the ten classifiers, three scenarios: (a) All classifiers use 10% of the training data. (b) One half the classifiers use $1/6^{\text{th}}$ of the data and the other half the rest. (c) One classifier uses $2/3^{\text{rd}}$ of the data, a second one uses $1/5^{\text{th}}$ and the eight other classifiers use the rest.

Table 2: Classification error rates for different datasets and scenarios. The best results are underlined and those that are not significantly different are in bold.

	Adult #train=600, #test=16,281			Australian #train=300, #test=390		
Scenario	(a)	(b)	(c)	(a)	(b)	(c)
Bayesian	16.76%	17.30%	19.10%	**14.87**%	**14.10**%	14.10%
Inv. Pign.	**16.68**%	17.21%	18.98%	**14.87**%	**14.10**%	**13.59**%
Likelihood	**16.71**%	**16.97**%	**18.35**%	**14.87**%	**13.33**%	**11.54**%

	Diabetes #train=300, #test=468		
Scenario	(a)	(b)	(c)
Bayesian	**21.58**%	**22.86**%	46.58%
Inv. Pign.	**21.37**%	**22.86**%	45.30%
Likelihood	**20.94**%	**22.65**%	**31.84**%

6. Conclusion

In this paper, we have shown how to extend classical probabilistic calibration methods using belief functions. Belief functions can better represent the uncertainty of the calibration procedure especially when few data are used. The likelihood-based model gave good overall results and can be used with binning, isotonic regression and logistic regression. The joint use of logistic regression and likelihood-based model led to the most promising results.

The methods presented in this paper were applied to the calibration of SVM classifiers but they may also be used for other classification algorithms. Their extension to multi-class problem is also possible through the use of binary decomposition such as one-vs-one [22, 35] or one-vs-all [38]. Comparison of probabilistic approaches and evidential ones [29] will be considered in future work.

Acknowledgments

This research was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the Agence Nationale de la Recherche (Reference ANR-11-IDEX-0004-02). It was supported by the ANR-NSFC Sino-French PRETIV project (References ANR-11-IS03-0001 and NSFC-61161130528).

References

[1] M. Ayer, H. Brunk, G. Ewing, W. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 5:641–647, 1955.

[2] P. L. Bartlett and A. Tewari. Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8:775–790, 2007.

[3] A. Bella, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. On the effect of calibration in classifier combination. *Applied Intelligence*, 38(4):566–585, 2013.

[4] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27.1–27.27, 2011.

[6] C. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.

[7] M. H. DeGroot and S. E. Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 32(1):12–22, 1982.

[8] A. Dempster. The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*, 48(2):365–377, 2008.

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

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

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

[12] T. Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3):228–252, 2006.

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

[14] T. Denœux. Rejoinder on "Likelihood-based belief function: justification and some extensions to low-quality data". *International Journal of Approximate Reasoning*, 55(7):1614–1617, 2014.

[15] T. Denœux and M. Skarstein-Bjanger. Induction of decision trees from partially classified data. In *Proceedings of SMC'2000*, pages 2923–2928, Nashville, TN, October 2000.

[16] T. G. Dietterich. Approximate statistical tests for comparing supervised classification algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[17] D. Dubois, H. Prade, and Ph. Smets. A definition of subjective possibility. *International Journal of Approximate Reasoning*, 48(2):352–364, 2008.

[18] R. P. W. Duin. The combining classifier: to train or not to train? In *Proceedings of the International Conference on Pattern Recognition*, volume 2, pages 765–770, Quebec, Canada, 2002.

[19] Z. Elouedi, K. Melloudi, and Ph. Smets. Belief decision trees: theoretical foundations. *International Journal of Approximate Reasoning*, 28(2–3):91–124, November 2001.

[20] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[21] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–318, 1986.

[22] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26:451–471, 1998.

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

[24] B. Khaleghi, A. Khamis, F. Karray, and S. Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44, 2013.

[25] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[26] H. T. Nguyen. *An Introduction to Random Sets*. Chapman and Hall/CRC Press, Boca Raton, Florida, 2006.

[27] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the International Conference on Machine Learning*, pages 625–632, Bonn, Germany, 2005.

[28] J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large-Margin Classifiers*, pages 61–74. MIT Press, 1999.

[29] B. Quost, T. Denœux, and M.-H. Masson. Pairwise classifier combination using belief functions. *Pattern Recognition Letters*, 28(5):644–653, 2007.

[30] G. Shafer. *A mathematical theory of evidence*. Princeton University Press, Princeton, New Jersey, 1976.

[31] Ph. Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, May 1990.

[32] Ph. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2):191–234, 1994.

[33] N. Sutton-Charani, S. Destercke, and T. Denœux. Classification trees based on belief functions. In T. Denoeux and M.-H. Masson, editors, *Belief Functions: Theory and Applications*, volume 164 of *Advances in Intelligent and Soft Computing*, pages 77–84. Springer Berlin Heidelberg, 2012.

[34] S. Tulyakov, S. Jaeger, V. Govindaraju, and D. Doermann. Review of classifier combination methods. In S. Marinai and H. Fujisawa, editors, *Machine Learning in Document Analysis and Recognition*, volume 90 of *Studies in Computational Intelligence*, pages 361–386. Springer Berlin Heidelberg, 2008.

[35] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

[36] Ph. Xu, F. Davoine, and T. Denœux. Evidential logistic regression for binary SVM classifier calibration. In F. Cuzzolin, editor, *Belief Functions: Theory and Applications*, volume 8764 of *Lecture Notes in Computer Science*, pages 49–57. Springer International Publishing, 2014.

[37] B. Zadrozny and C. Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *International Conference on Machine Learning*, pages 609–616, Williamstown, Maryland, 2001.

[38] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, New York, USA, 2002. ACM.