Evidential joint calibration of binary SVM classifiers using logistic regression

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

[1]Univ. Artois, EA 3926,
Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A),
Béthune, F-62400, France.
{frederic.pichon,david.mercier,eric.lefevre}@univ-artois.fr
[2]SNCF Réseau, Département des Télécommunications,
La Plaine Saint Denis, France.
{pauline.minary,benjamin.droit}@reseau.sncf.fr

Abstract. In a context of multiple classifiers, a calibration step based on logistic regression is usually used to independently transform each classifier output into a probability distribution, to be then able to combine them. This calibration has been recently refined, using the evidence theory, to better handle uncertainties. In this paper, we propose to use this logistic-based calibration in a multivariable scenario, *i.e.*, to consider jointly all the outputs returned by the classifiers, and to extend this approach to the evidential framework. Our evidential approach was tested on generated and real datasets and presents several advantages over the probabilistic version.

Keywords: Belief functions, Information fusion, Evidential calibration.

1 Introduction

Using several classifiers to obtain different information on a given object and combining their outputs is a means to obtain better classification performance. These classifiers may be trained with different data or may not rely on the same training models. Thus, their outputs may not be of the same type or not scaled with each other. To be able to combine them, they first have to be made comparable: a technique called calibration is usually applied, enabling to transform a classifier output into a probability. One of the most commonly used calibration is based on logistic regression [8].

Recently, Xu *et al.* [11] proposed a refinement of this calibration within a framework for reasoning under uncertainty called evidence theory [9, 10]. This theory models more precisely the uncertainties inherent to such calibration process and thus enables to prevent an over-fitting issue that may appear, especially when few training data are available. Thus, given a single classifier returning a confidence score after observing a given object, Xu *et al.*'s approach transforms this score into a belief function.

There exists a multivariable version of the logistic regression, called the multiple logistic regression [5], where the technique is defined with more than one input. If we apply this approach to the vector of scores returned by the classifiers for a given object, we can obtain a joint calibration, which returns a probability. Yet, this technique is also prone to the uncertainty problem. Within this scope, we propose to use the evidential extension of calibration proposed by Xu *et al.*, and to apply it to the calibration based on the multiple logistic regression. Thus, for a given object, our proposed approach transforms the vector of scores returned by the classifiers into a belief function.

This paper is organized as follows. First, Section 2 recalls the necessary background on evidence theory. Then, Section 3 exposes the probabilistic calibration based on the multiple logistic regression and the extension to the evidential framework that we propose. In Section 4, the proposed approach and its probabilistic version are compared. Finally, conclusion and perspectives are given in Section 5.

2 Evidence theory

In this section, basic notions of the evidence theory are first exposed in Section 2.1. Applications of this theory to inference and prediction, which are useful to define calibration in the evidential framework, are addressed in Section 2.2.

2.1 Basic notions

The theory of evidence is a framework for reasoning under uncertainty. Let Ω be a finite set called the frame of discernment, which contains all the possible answers to a given question of interest Q. In this theory, uncertainty with respect to the answer to Q is represented using a *Mass Function* (MF) defined as a mapping $m^{\Omega} : 2^{\Omega} \rightarrow [0,1]$ that satisfies $\sum_{A \subseteq \Omega} m^{\Omega}(A) = 1$ and $m^{\Omega}(\emptyset) = 0$. The quantity $m^{\Omega}(A)$ corresponds to the share of belief that supports the claim that the answer is contained in $A \subseteq \Omega$ and nothing more specific. Any subset A of Ω such that $m^{\Omega}(A) > 0$ is called a focal set of m^{Ω}. When the focal sets are nested, m^{Ω} is said to be consonant. A mass function can be equivalently represented by the belief and plausibility functions, respectively defined by

$$Bel^{\Omega}(A) = \sum_{B \subseteq A} m^{\Omega}(B), \quad Pl^{\Omega}(A) = \sum_{B \cap A \neq \emptyset} m^{\Omega}(B), \quad \forall A \subseteq \Omega. \tag{1}$$

The plausibility function restricted to singletons is called the contour function, denoted pl^{Ω} and defined by $pl^{\Omega}(\omega) = Pl^{\Omega}(\{\omega\}), \forall \omega \in \Omega$. When a mass function is consonant, the plausibility function can be recovered from its contour function with $Pl^{\Omega}(A) = \sup_{\omega \in A} pl^{\Omega}(\omega), \quad \forall A \subseteq \Omega$.

Different decision strategies exist to make a decision about the true answer to Q, given a MF m^{Ω} on this answer [4]. In particular, the answer having the smallest so-called *upper* or *lower expected costs* may be selected. When the set of focal elements is reduced to singletons and Ω, and when the costs are taken equal to 0 if the answer is correct and 1 otherwise, the upper and lower expected costs of some answer $\omega \in \Omega$ are respectively defined as $R^*(\omega) = 1 - m^\Omega(\{\omega\})$ and $R_*(\omega) = 1 - m^\Omega(\{\omega\}) - m^\Omega(\Omega)$. Choosing the answer minimizing minimizing the lower (resp. upper) expected costs is called the optimistic (resp. pessimistic) strategy. To avoid making risky decisions, when the expected costs are high, a reject decision can be introduced: we define $R_{rej} \in [0,1]$, and the reject decision is made when R_{rej} is lower than the other expected costs.

2.2 Statistical inference and forecasting

The evidence theory can be used for inference and forecasting. 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 θ after observing the data x. Shafer [9] proposed to represent the knowledge about θ by a consonant belief function Bel_x^Θ based on the likelihood function $L_x : \theta \to f_\theta(x)$, whose contour function is the normalized likelihood function:

$$pl_x^\Theta(\theta) = \frac{L_x(\theta)}{\sup\limits_{\theta' \in \Theta} L_x(\theta')}, \qquad \forall \theta \in \Theta. \tag{2}$$

Suppose now that we have some knowledge about θ after observing some data x, in the form of a contour function pl_x^Θ. The aim of forecasting is to make statements about a not yet observed data $Y \in \mathbb{Y}$, whose conditional distribution given $X = x$ depends on θ. A solution consists in using the sampling model of Dempster [3] to deduce a belief function on $\mathbb{Y}$ [6, 7]. This model proposes to express Y as a function of the parameter θ and some unobserved variable, whose distribution is independent of θ.

Let us consider an important particular case. Assume that $Y \in \mathbb{Y} = \{0, 1\}$ is a random variable with a Bernoulli distribution. In that case, Xu *et al.* [11] showed, by applying inference and forecasting, that we have

$$Bel_x^\mathbb{Y}(\{1\}) = \hat{\theta} - \int_0^{\hat{\theta}} pl_x^\Theta(u)du, \quad Pl_x^\mathbb{Y}(\{1\}) = \hat{\theta} + \int_{\hat{\theta}}^1 pl_x^\Theta(u)du, \tag{3}$$

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

3 An evidential joint calibration approach

Assume that after observing an object which belongs either to class 0 or 1, a SVM classifier returns a confidence score $s \in \mathbb{R}$. To learn how to interpret what this score represents with respect to the true label $y \in \mathbb{Y} = \{0, 1\}$ of the object, a step called calibration may be performed. In particular, the one based on logistic regression is commonly used [8]. It aims to estimate the probability distribution $p^\mathbb{Y}(\cdot|s)$ and relies on a training set. Yet, the less training samples are available, the more the estimated probabilities are uncertain. To manage these uncertainties, Xu *et al.* proposed to refine this calibration using the theory of evidence [11].

We propose to use the multiple version of the logistic regression [5] and to apply it to the outputs of multiple classifiers, *i.e.*, to perform a joint calibration of the scores provided by J binary SVM classifiers. It relies on a training set defined by $\mathcal{X} = \{(S_{11}, ..., S_{J1}, Y_1), ..., (S_{1n}, ..., S_{Jn}, Y_n)\}$, where S_{ji} corresponds to the score given by the j^{th} classifier for the i^{th} test sample, and Y_i its true label. Given a vector of scores $\mathbf{s} = (s_1, ..., s_J)$, with s_j the score returned by the j^{th} classifier, the calibration based on the multiple logistic regression can be defined by

$$P^{\mathbb{Y}}(y = 1|\mathbf{s}) \approx h_{\mathbf{s}}(\sigma) = \frac{1}{1 + \exp(\sigma_0 + \sigma_1 s_1 + \sigma_2 s_2 + ... + \sigma_J s_J)}, \tag{4}$$

where the parameter $\sigma = \{\sigma_0, ..., \sigma_J\} \in \mathbb{R}^{J+1}$ is obtained by maximizing the likelihood function L, defined by

$$L(\sigma) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}, \text{ with } p_i = \frac{1}{1 + \exp(\sigma_0 + \sigma_1 S_1 + ... + \sigma_J S_J)}. \tag{5}$$

To better handle the uncertainties, we propose to extend this approach to the evidential framework by following the same likelihood-based reasoning as in [11]. Calibration of a given vector of scores $\mathbf{s}$ based on logistic regression can be seen as a prediction problem of a Bernoulli variable Y with parameter θ, where $\theta = h_{\mathbf{s}}(\sigma)$. A belief function $Bel^{\mathbb{Y}}(\cdot|\mathbf{s})$ can be derived from the contour function $pl^{\Theta}_{\mathcal{X}}(\cdot|\mathbf{s})$ using Eq. (3). Following Xu *et al.* [11], this contour function can be computed from $Pl^{\Sigma}_{\mathcal{X}}$, which is the plausibility function of $pl^{\Sigma}_{\mathcal{X}}$ defined by

$$pl^{\Sigma}_{\mathcal{X}}(\sigma) = \frac{L(\sigma)}{L(\hat{\sigma})}, \quad \forall \sigma \in \Sigma, \tag{6}$$

with $\hat{\sigma} = (\hat{\sigma}_0, ..., \hat{\sigma}_J)$ the Maximum Likelihood Estimate (MLE) of σ and L the likelihood defined in Eq. (5). As $\theta = h_{\mathbf{s}}(\sigma)$, we have

$$pl^{\Theta}_{\mathcal{X}}(\theta|\mathbf{s}) = \begin{cases} 0 & \text{if } \theta \in \{0, 1\}, \\ Pl^{\Sigma}_{\mathcal{X}}(h_{\mathbf{s}}^{-1}(\theta)) & \text{otherwise,} \end{cases} \tag{7}$$

with

$$h_{\mathbf{s}}^{-1}(\theta) = \left\{ (\sigma_0, \sigma_1, ..., \sigma_J) \in \Sigma | \frac{1}{1 + \exp(\sigma_0 + \sigma_1 s_1 + ... + \sigma_J s_J)} = \theta \right\}, \tag{8}$$

$$= \left\{ (\sigma_0, \sigma_1, ..., \sigma_J) \in \Sigma | \sigma_0 = \ln(\theta^{-1} - 1) - \sigma_1 s_1 - ... - \sigma_J s_J \right\}. \tag{9}$$

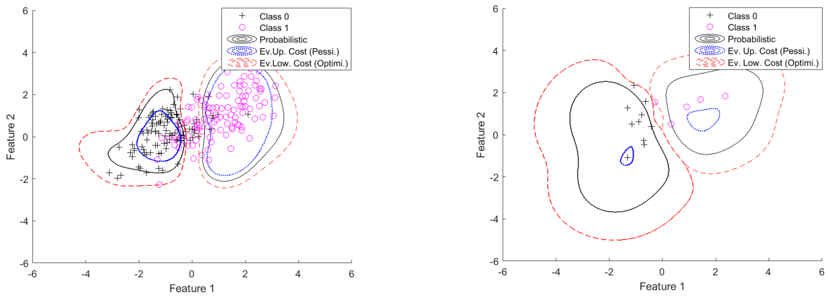
Thus, Eqs. (7) and (9) yield the following contour function

$$pl^{\Theta}_{\mathcal{X}}(\theta|\mathbf{s}) = \sup_{\sigma_1, ..., \sigma_J \in \mathbb{R}} pl^{\Sigma}_{\mathcal{X}'}(\ln(\theta^{-1} - 1) - \sigma_1 s_1 - \sigma_2 s_2 - ... - \sigma_J s_J, \sigma_1, ..., \sigma_J), \tag{10}$$

for all $\theta \in [0, 1]$. The vector of parameters $(\sigma_1, \sigma_2, ..., \sigma_J)$ which maximizes $pl^{\Sigma}_{\mathcal{X}}$ can be approximated using an iterative maximization algorithm. The computational complexity of such algorithm is $O(nJ)$ per iteration.

4 Experiments

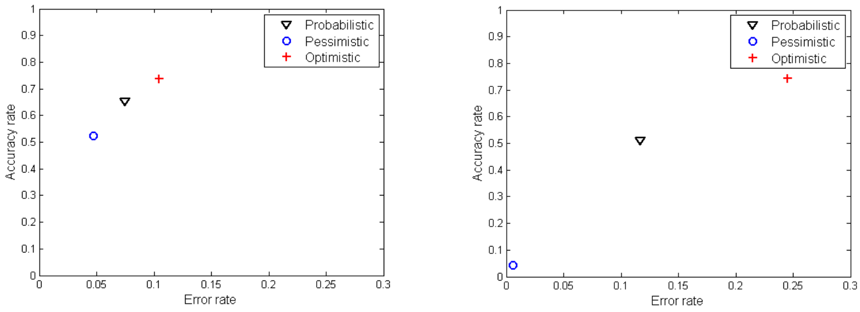
We simulated a binary dataset composed of randomly generated instance vectors from a multivariate normal distribution, composed of two features, with means $\mu_0 = (-1, 0)$ in class 0 and $\mu_1 = (1, 1)$ in class 1, and with a covariance matrix equals to $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ for both classes. The possibility of deciding to reject a test sample was introduced, and we used both pessimistic and optimistic strategies for the evidential approach. We generated a set of 290 training samples: three SVM classifiers were trained, using the LIBSVM library [2], with three non-overlapping subset of 30 training samples of this set, and our evidential joint calibration was trained with the remaining 200 samples. Then, the same experiment was performed but with 15 examples to train the approach. The decision frontiers in both cases are illustrated in Figure 1, for $R_{rej} = 0.2$. As it can be seen, the approach based on the optimistic strategy tends to decide more, hence to reject less, the test samples than the two others and it is the exact opposite for the pessimistic strategy. Furthermore, the frontiers are a lot more distant from each other when there are less examples to train the approach (Figure (1b)), *i.e.*, when there are more uncertainties. The probabilistic calibration only yields one frontier so the impact of the uncertainties is not visible. Thus, evidential joint approaches better reflect the uncertainties than the probabilistic one, and using an evidential approach enables to choose between a strategy which decide more often and reject less test samples, or the opposite.



(a) Approach trained with 200 examples (b) Approach trained with 15 examples

Fig. 1: Decision frontiers in feature space of the joint calibration trained with 200 (1a) and 15 (1b) training examples, for $R_{rej} = 0.2$.

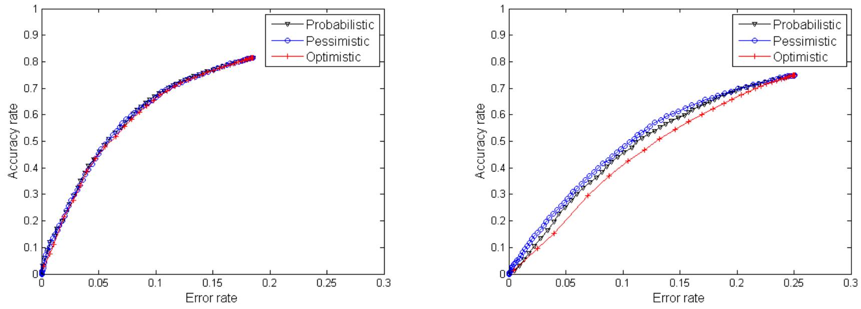
With the same set repartition, we calculated the error rate and accuracy rates for 100 test samples and $R_{rej} = 0.2$. Accuracy rate corresponds to the number of correctly classified objects over the number of classified objects, *i.e.*, not over the total number of test samples as some of them are rejected. The process was repeated for 100 rounds of random partitioning. The obtained points are more distant from each other when few training examples are available (Figure 2). This interval reflects the uncertainties as it is larger when they are more important. This information cannot be obtained with the probabilistic approach, which is represented by only one point.



(a) Approach trained with 200 examples (b) Approach trained with 15 examples

Fig. 2: Error and accuracy rates for $R_{rej} = 0.2$ and with 200 (2a) and 15 (2b) training examples.

Furthermore, we performed the same experiment with R_{rej} varying from 0 to 1, on four datasets (*Australian*, *Diabetes*, *Heart*, *Ionosphere*) of UCI repository [1] and on the simulated dataset. The classifiers were still trained on non-overlapping subsets of 30 examples, either for simulated or real data. Our joint calibration was trained with 45 then 15 samples. Figure 3 shows the results obtained for the simulated dataset; those obtained for the real datasets are similar. For a given error rate, the results obtained with the pessimistic strategy has a higher (or equal) accuracy rate than the probabilistic one when few training examples are available (right column). We may notice that these two points are obtained with



(a) Simulated data – 45 training examples (b) Simulated data – 15 training examples

Fig. 3: Error and accuracy rates with 45 (left) and 15 training examples (right).

different R_{rej}, as seen in the previous experiment. Furthermore, when there are more training examples (left column), the obtained results become similar for the probabilistic and evidential approaches.

Finally, we compared our evidential joint approach to Xu *et al.*'s approach [11], which independently calibrate the scores given by single classifier and combine them with Dempster's rule [9]. We performed the same experiment as the first one detailed in [11], where the training set size for the third classifier was varying. The training of our joint calibration was performed by concatenating the calibration training subsets of the three classifiers. The joint proposed approach presents lower error rates than Xu *et al.*'s approach on the simulated dataset as well as on the real data (results cannot be shown due to space limitations).

5 Conclusion

In this paper, an evidential joint calibration based on logistic regression was proposed. Logistic regression is commonly used to calibrate the scores of a single classifier and we used its multiple version to take into account together the scores returned by multiple classifiers for an object. The application of evidence theory enables to better handle the process uncertainties than the probabilistic version.

We only studied the calibration using logistic regression but the same reasoning can be applied to other calibration techniques. Finally, an extension of our approach to multiclass problem could also be considered in future works.

References

1. K. Bache and M. Lichman. UCI machine learning repository. 2013. http://archive.ics.uci.edu/ml.
2. C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
3. A.P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, 37(2):355–374, 1966.
4. T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.
5. D.W. Hosmer and S. Lemeshow. *Applied logistic regression.* Wiley & Sons, 2004.
6. O. Kanjanatarakul, T. Denœux, and S. Sriboonchitta. Prediction of future observations using belief functions: A likelihood-based approach. *International Journal of Approximate Reasoning*, 72:71–94, 2015.
7. 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.
8. J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. in large margin classifiers*, 10(3):61–74, 1999.
9. G. Shafer. *A mathematical theory of evidence.* Princeton University Press, 1976.
10. Ph. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.
11. P. Xu, F. Davoine, H. Zha, and T. Denœux. Evidential calibration of binary SVM classifiers. *International Journal of Approximate Reasoning*, 72:55–70, 2016.