Active evidential calibration of binary SVM classifiers

Frédéric Pichon, Sébastien Ramel, François Delmotte

Active evidential calibration of binary SVM classifiers

Sébastien Ramel, Frédéric Pichon, and François Delmotte

Univ. Artois, EA 3926, Laboratoire de Génie Informatique et d'Automatique de l'Artois (LGI2A), F-62400 Béthune, France,
firstname.lastname@univ-artois.fr

Abstract. Evidential calibration methods of binary classifiers improve upon probabilistic calibration methods by representing explicitly the calibration uncertainty due to the amount of training (labelled) data. This justified yet undesirable uncertainty can be reduced by adding training data, which are in general costly. Hence the need for strategies that, given a pool of unlabelled data, will point to interesting data to be labelled, *i.e.*, to data inducing a drop in uncertainty greater than a random selection. Two such strategies are considered in this paper and applied to an ensemble of binary SVM classifiers on some classical binary classification datasets. Experimental results show the interest of the approach.

Keywords: belief functions, evidential calibration, active learning.

1 Introduction

Probabilistic calibration methods, such as isotonic and logistic (Platt scaling) regressions, allow to learn from training data how to transform classifier outputs into probabilities that an instance belongs to each of the classes [1]. They are useful for the many applications where it is important to provide such probabilities rather than mere crisp decisions and where the available classifiers output scores, such as SVMs, or inaccurate probabilities, such as Naive Bayes [1,2]. Besides, they have been mainly designed so far for binary classification.

A limitation of these methods is that they do not take into account the uncertainty due to the amount of training data in their probability estimates and, in particular, the less training data, the more uncertain the probability estimates [3]. To address this issue, the calibration problem has been considered recently in the framework of belief function theory, yielding so-called evidential calibration methods (see [3] for the calibration of a single binary classifier and [4] for the calibration of an ensemble of binary classifiers). These latter methods are able to represent explicitly the uncertainty due to the amount of training data, which is important in critical application domains and also leads to better classification performance than probabilistic calibration methods as shown in [3,4].

While it is important to represent the aforementioned uncertainty, it is even better if this uncertainty is as small as possible. In order to reduce it, one needs to bring in additional training (labelled) data, which may be costly and hence

must be done in an efficient manner, *i.e.*, such that for any given number of added labelled data the uncertainty is reduced as much as possible. It is a similar problem to that of active learning [5], except that the primary focus is not on improving accuracy but rather on reducing uncertainty, and it is the problem tackled in this paper.

Specifically, we consider the following setting: we assume an initial set of labelled data from which some classifiers can be evidentially calibrated, and then we consider that it is possible to ask iteratively an oracle to label some data from a pool of data with missing labels. We study two strategies to decide which instances from the pool should be given to the oracle. These strategies are in the spirit of the so-called uncertainty sampling strategy framework from active learning [5], where instances in the pool are ordered according to how much the current classifier is the most unsure about.

This paper is organized as follows. Section 2 recalls the necessary background on the evidential calibration of binary classifiers. Then, Section 3 presents two strategies for the active evidential calibration of such classifiers and reports experimental results when these strategies are applied to binary SVM classifiers. Finally, Section 4 concludes the paper.

2 Evidential calibration

Evidential calibration of binary classifiers, as introduced in [3] for the case of a single classifier and further developed in [4] for an ensemble of classifiers, relies on some recent results by Kanjanatarakul *et al.* [6,7] concerning the prediction of a Bernoulli random variable, which are recalled in the next section.

We will assume that the reader has some basic knowledge of the theory of belief functions (a reminder can be found in [7]).

2.1 Prediction of a Bernoulli random variable

Kanjanatarakul *et al.* [6,7] proposed a general approach which, given some knowledge about some parameter θ obtained by observing a realization x of some random quantity X with distribution $f_\theta(x)$ and represented by a belief function Bel_x^Θ[1], makes it possible to make statements in the form of a belief function $Bel_x^\mathbb{Y}$ about some random quantity $Y \in \mathbb{Y}$ whose conditional distribution $g_{x,\theta}(y)$ given $X = x$ depends on θ.

In particular, if Y is a binary random variable ($\mathbb{Y} = \{0, 1\}$) with associated Bernoulli distribution $\mathcal{B}(\theta)$, $\theta \in [0, 1]$, and if Bel_x^Θ is a consonant belief function whose associated contour function pl_x^Θ is unimodal and continuous, we have [6]:

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

[1] Bel_x^Θ must be *induced by a source* [7]. It may be obtained by a number of evidential methods to statistical inference, and in particular the likelihood-based evidential method [8] in which case Bel_x^Θ is the consonant belief function whose contour function is the normalized likelihood function given the observed data x.

where $\hat{\theta}$ maximizes pl_x^{Θ}. The degree of belief $Bel_x^{\mathbb{Y}}(\{1\})$ represents the amount of evidence strictly supporting $Y = 1$ while the plausibility $Pl_x^{\mathbb{Y}}(\{1\}) = 1 - Bel_x^{\mathbb{Y}}(\{0\})$ is the amount of evidence not contradicting it. Besides, the difference $Pl_x^{\mathbb{Y}}(\{1\}) - Bel_x^{\mathbb{Y}}(\{1\})$, which is equal to the mass $m_x^{\mathbb{Y}}(\{0,1\})$ assigned to the ignorance, is merely the area under the contour function pl_x^{Θ} and the size of this area tends to 0 if, *e.g.*, X follows a binomial distribution with parameters n and θ, Bel_x^{Θ} is obtained using the likelihood-based method and n tends to infinity [6].

2.2 Evidential calibration methods

Let $\mathcal{C} = \{(s_1, y_1), \ldots, (s_n, y_n)\}$ be some training data in a binary classification problem, where $s_i \in \mathbb{S}$ for some domain $\mathbb{S}$ is the output provided by a pre-trained classifier for the i-th training sample with label $y_i \in \{0, 1\}$. For a test sample of output $s \in \mathbb{S}$ and unknown label $y \in \{0, 1\}$, any evidential calibration method proposed in [3] returns two values: the belief $Bel_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})$ and plausibility $Pl_{\mathcal{C},s}^{\mathbb{Y}}(\{1\})$ that $y = 1$. These methods obtain these two values by seeing the label y of the test sample as the realization of a random variable Y with a Bernoulli distribution $\mathcal{B}(\theta)$ given knowledge about θ represented by some consonant belief function $Bel_{\mathcal{C},s}^{\Theta}$ with contour function $pl_{\mathcal{C},s}^{\Theta}$ depending on $\mathcal{C}$ and s, and by applying then to Y the prediction approach recalled in Section 2.1.

The only difference between the evidential calibration methods in [3] is thus the way $pl_{\mathcal{C},s}^{\Theta}$ is defined. There are indeed several ways to define $pl_{\mathcal{C},s}^{\Theta}$: it depends on which probabilistic calibration method is extended and on which evidential approach to statistical inference is used (see [3, Section 4] for details). In this paper, we focus on the evidential calibration methods where $pl_{\mathcal{C},s}^{\Theta}$ is obtained using the likelihood-based evidential approach to statistical inference, as Xu *et al.* [3] showed that this is the approach presenting overall the best performances.

More precisely, let us consider two cases: $\mathbb{S} = \{0, 1\}$ and $\mathbb{S} = \mathbb{R}$. The case $\mathbb{S} = \{0, 1\}$ corresponds to a classifier returning binary outputs and it will allow us to investigate in Section 3 the behaviours of our active evidential calibration strategies in a simple setting. The case $\mathbb{S} = \mathbb{R}$ corresponds to a classifier returning scores, such as a SVM classifier, and it will allow us to recall shortly and progressively the arguably most involved and best evidential calibration method considered so far to deal with an ensemble of classifiers – the behaviours of our active strategies with respect to this latter calibration scheme of an ensemble of classifiers will also be investigated in Section 3.

The case $\mathbb{S} = \{0, 1\}$ can be handled using the likelihood-based evidential extension of the binning calibration method [3], in which case we have[2]:

$$pl_{\mathcal{C},s}^{\Theta}(\theta) = \frac{\theta^{k_s}(1-\theta)^{n_s-k_s}}{\hat{\theta}_s^{k_s}(1-\hat{\theta}_s)^{n_s-k_s}}, \quad \forall s \in \mathbb{S}, \tag{2}$$

with $k_s = |\{(s_i, y_i) \in \mathcal{C} | s_i = s, y_i = 1\}|$, $n_s = |\{(s_i, y_i) \in \mathcal{C} | s_i = s\}|$ and $\hat{\theta}_s = k_s/n_s$.

[2] Eq. (2) corresponds to a degenerate binning approach with only two bins. It can be derived rigorously without referring to the evidential binning calibration, by following a similar reasoning to the one used in [3] to obtain this latter calibration.

The case $\mathbb{S} = \mathbb{R}$ can be handled using the likelihood-based evidential extension of the logistic regression [3], in which case $pl^{\Theta}_{\mathcal{C},s}$ is defined as:

$$pl^{\Theta}_{\mathcal{C},s}(\theta) = \sup_{\sigma_1 \in \mathbb{R}} pl^{\Sigma}_{\mathcal{C}}(\ln(\theta^{-1} - 1) - \sigma_1 s, \sigma_1), \quad \forall s \in \mathbb{S}, \tag{3}$$

with $pl^{\Sigma}_{\mathcal{C}}(\sigma) = \frac{L(\sigma)}{L(\hat{\sigma})}$, $\forall \sigma = (\sigma_0, \sigma_1) \in \Sigma = \mathbb{R}^2$, where $L(\sigma) = \prod_{i=1}^{n} p_i^{t_i}(1-p_i)^{1-t_i}$, with $p_i = \frac{1}{1+\exp(\sigma_0+\sigma_1 s_i)}$ and $t_i = \frac{N_1+1}{N_1+2}$ if $y_i = 1$, $t_i = \frac{1}{N_0+2}$ if $y_i = 0$, with $N_j = |\{(s_i, y_i) \in \mathcal{C} | y_i = j\}|$, and $\hat{\sigma}$ maximizing L.

Of particular interest is that using $pl^{\Theta}_{\mathcal{C},s}$ defined by (2) or by (3) in Eq. (1), $Pl^{\mathbb{Y}}_{\mathcal{C},s}(\{1\}) - Bel^{\mathbb{Y}}_{\mathcal{C},s}(\{1\}) = m^{\mathbb{Y}}_{\mathcal{C},s}(\{0,1\})$ decreases as n increases [3]. In other words, $m^{\mathbb{Y}}_{\mathcal{C},s}(\{0,1\})$ reflects the amount of training data, and in particular the less training data there are, the more ignorance or uncertainty there is.

Let us now consider a somewhat more complex problem, where we have an ensemble of m classifiers such that given a test sample of unknown label $y \in \{0,1\}$, we obtain a vector of outputs $\mathbf{s} = (s^1, ..., s^m) \in \mathbb{R}^m$ with s^j the output of the j-th classifier. In order to be able to interpret $\mathbf{s}$ with respect to y, a solution proposed in [4] consists in calibrating *jointly* the classifiers. A joint calibration proceeds similarly as the calibration of a single classifier: the label y is seen as the realization of a random variable with a Bernoulli distribution $\mathcal{B}(\theta)$ and a belief function $Bel^{\mathbb{Y}}_{\mathcal{C},\mathbf{s}}$ is derived using the prediction approach (1) from knowledge on θ represented by a contour function $pl^{\Theta}_{\mathcal{C},\mathbf{s}}$ depending on $\mathbf{s}$ and a training set $\mathcal{C} = \{(\mathbf{s}_1, y_1), \ldots, (\mathbf{s}_n, y_n)\}$ where $\mathbf{s}_i$ is the output vector provided by the m classifiers for the i-th training sample with label $y_i \in \{0,1\}$. More specifically, Minary *et al.* [4] proposed an evidential joint calibration corresponding to the likelihood-based evidential extension of the multiple logistic regression, which is a generalization of the evidential logistic regression recalled above and, in particular, the definition of $pl^{\Theta}_{\mathcal{C},\mathbf{s}}$ derived in [4] is a straightforward multivariate generalization of (3) (due to lack of space, we refer the reader to [4] for the detailed definition of $pl^{\Theta}_{\mathcal{C},\mathbf{s}}$).

3 Active evidential calibration

As we have seen, evidential calibration methods return for a test sample with classifier output s a degree of belief $Bel^{\mathbb{Y}}_{\mathcal{C},s}(\{1\})$ and a plausibility$Pl^{\mathbb{Y}}_{\mathcal{C},s}(\{1\})$ representing, respectively, the amount of evidence strictly supporting that the label y of the sample is 1 and the amount of evidence not contradicting it. Hence, the greater the interval $[Bel^{\mathbb{Y}}_{\mathcal{C},s}(\{1\}), Pl^{\mathbb{Y}}_{\mathcal{C},s}(\{1\})]$, the more uncertain one is about the actual support that should be given to $y = 1$. It is thus clear that while it is important that uncertainty induced by the training data be represented, this uncertainty should be small enough otherwise no useful conclusion about y may be drawn, that is, the calibrated classifier is not useful.

In order to reduce the uncertainty, one needs to add some training (labelled) data. It is generally possible and relatively easy to obtain some unlabelled data but, depending on the domain, labelling it may be costly. Besides, it may be

the case that not all training data are equivalent with respect to the drop in uncertainty that they induce. Hence, it seems useful to devise some strategies that, given a pool of unlabelled data, will point to interesting data to be labelled, that is, to data that will induce a drop in uncertainty greater than selecting at random data in the pool. We refer to such strategies as active evidential calibration strategies, or active strategies for short, in opposition to the passive strategy, which is the selection at random. We propose two such strategies in Section 3.1, which we then test on a single classifier and on an ensemble of classifiers in Sections 3.2 and 3.3, respectively.

3.1 Active strategies

In pool-based active learning [5], an active learner asks queries in the form of unlabelled instances (taken from the pool) to be labeled by an oracle, and the labeled instances are then moved to the learning set, with the aim that classification accuracy will improve faster than with a random selection strategy. Several query strategy frameworks have been proposed [5]. In particular, *uncertainty sampling* for a classifier with probabilistic outputs selects the unlabelled pool instance for which the classifier output has the greatest (Shannon) entropy.

Since our aim is to reduce the uncertainty represented by the quantity $m^{\mathbb{Y}}_{\mathcal{C},s}(\{0,1\})$ for any given test instance of score $s \in \mathbb{S}$, a natural query strategy is to select from a pool $\mathcal{P} = \{s^{\mathcal{P}}_1, \ldots, s^{\mathcal{P}}_p\}$ of unlabelled instances with classifier outputs $s^{\mathcal{P}}_k$, $k = 1, \ldots, p$, the instance $s^* \in \mathcal{P}$ that has the greatest uncertainty $m^{\mathbb{Y}}_{\mathcal{C},s^*}(\{0,1\})$. We note that an uncertainty measure for a mass function $m^{\mathbb{Y}}$ is the generalized Hartley measure [9], which evaluates its nonspecificity and is defined as $GH(m^{\mathbb{Y}}) := \sum_{A \subseteq \mathbb{Y}} m^{\mathbb{Y}}(A) \log_2 |A|$; if $\mathbb{Y} = \{0,1\}$, we have $GH(m^{\mathbb{Y}}) = m^{\mathbb{Y}}(\{0,1\})$. Hence, this strategy is similar to that of uncertainty sampling in active learning, except that it uses another uncertainty measure (the generalized Hartley measure instead of the Shannon entropy), and may thus be called Hartley Sampling (HS). It selects the instance $s^*_{HS} \in \mathcal{P}$ such that

$$s^*_{HS} = \arg\max_{s^{\mathcal{P}} \in \mathcal{P}} GH(m^{\mathbb{Y}}_{\mathcal{C},s^{\mathcal{P}}}). \tag{4}$$

In addition to the HS strategy, we consider for comparison purposes another query strategy, which is closer to uncertainty sampling of active learning: this second strategy, called Pignistic Sampling (PS), selects the instance $s^*_{PS} \in \mathcal{P}$ whose associated pignistic probability distribution [10] denoted $BetP(m^{\mathbb{Y}}_{\mathcal{C},s^*_{PS}})$ has the greatest (Shannon) entropy:

$$s^*_{PS} = \arg\max_{s^{\mathcal{P}} \in \mathcal{P}} H(BetP(m^{\mathbb{Y}}_{\mathcal{C},s^{\mathcal{P}}})), \tag{5}$$

with $H(P)$ the Shannon entropy of probability distribution P. Note that since uncertainty sampling is designed to improve accuracy, one might expect that PS will improve accuracy, but it is not clear whether it will improve uncertainty.

Let us remark that the generalized Hartley measure and the Shannon entropy of the pignistic transformation have previously shown their interest in improving classification accuracy in the context of active classification [11].

3.2 Active evidential calibration of a classifier with binary outputs

The active strategies described in the previous section are first tested with respect to a single classifier with binary outputs, *i.e.*, $\mathbb{S} = \{0, 1\}$, in which case the classifier is calibrated using (2). The test is conducted using simulated data.

Specifically, let $P(S = s, Y = y)$, $s \in \mathbb{S}$, $y \in \mathbb{Y}$, denote a given bivariate Bernoulli distribution for the pair (S, Y) of binary random variables S and Y, where S represents the classifier output and Y the true class. Such a distribution is completely characterized by the marginal probabilities $P(S = 1)$ and $P(Y = 1)$ and the covariance σ between S and Y [12].

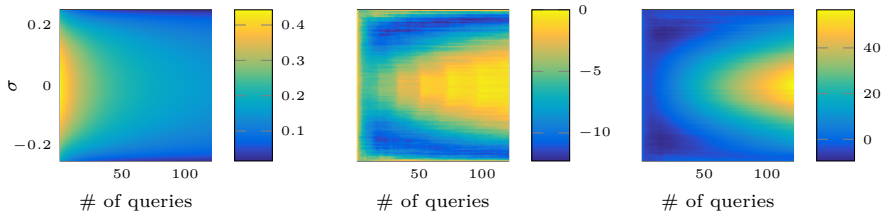
In our experiment, we chose $P(S = 1) = P(Y = 1) = 0.5$ and considered all possible joint distributions $P(S = s, Y = y)$, $s \in \mathbb{S}$, $y \in \mathbb{Y}$, having those marginals: these are all the distributions that are obtained by choosing $\sigma \in [-0.25, 0.25]$, which is the range of possible values for σ given these marginals.

We drew randomly 10^6 samples in each of these joint distributions. We used a 1000-fold cross-validation procedure over these samples: the samples are randomly split into 1000 folds. Each fold (which contains 1000 samples) is in turn considered as the test set, and the other folds are combined to obtain a dataset which is randomly split into two parts: the first part composed of 10 instances is used as initial training data set $\mathcal{C}$ for the evidential calibration of the classifier, and the second part composed of the remaining instances acts as the pool. The maximal number of queries for each query strategy (HS, PS and Random Sampling (RS)) was set to 120. For each fold used as test set and for each query strategy, we computed the average of the uncertainty, *i.e.*, ignorance with respect to the label after calibration, of the test instances as the number of queries increases. Finally, we averaged these latter averages over the 1000 test folds.

Figure 1 shows the performances in terms of uncertainty reduction achieved by the active strategies HS and PS with respect to the passive one (RS) used as reference. HS performs globally better than RS (up to 12% better) – it becomes equivalent to RS when σ gets closer to 0 and the number of queries increases, as well as when σ gets closer to -0.25 and 0.25, which are all extreme dependence situations between S and Y. PS is beneficial with respect to RS for roughly the same zones as HS, albeit to a slightly lesser extent, but clearly detrimental (up to 55% worse) as the number of queries increases and as we get closer to $\sigma = 0$. Let us note that similar figures are obtained when other marginal probabilities $P(S = 1)$ and $P(Y = 1)$ are used (the figures are then somewhat distorted versions of the ones presented here).

3.3 Active evidential joint calibration of binary SVM classifiers

The active strategies are now tested with respect to an ensemble of 3 SVM classifiers (trained with the LIBSVM library), which are jointly calibrated using the evidential multiple logistic regression described in Section 2.2. We used 6 binary classification datasets from the UCI repository: Australian, Heart, Ionosphere, Sonar, WDBC, Diabetes. Each dataset was randomly partitioned into 6 subsets: 3 subsets of 20 instances each to train each SVM, one subset of 100 instances



(a) Average uncertainty of RS over 1000-fold cross validation. (b) Uncertainty change (in %) from RS to HS. (c) Uncertainty change (in %) from RS to PS.

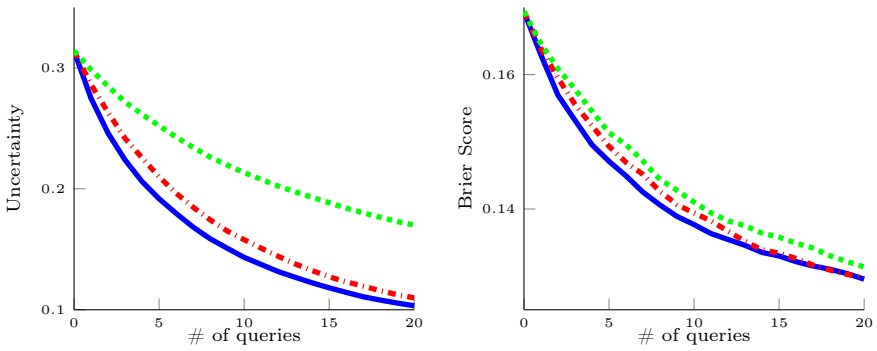
Fig. 1: Comparison of active strategies for a classifier with binary outputs.

to act as test set (except for Sonar, for which we used only 50 test samples due to its relatively small size), one subset of 10 instances to train the initial joint calibration of the classifiers, one subset containing the remaining instances and acting as the pool. Over the test set, we computed the average uncertainty of the strategies RS, HS and PS, as well as their Brier score (mean squared error), which is a standard performance (accuracy-like) measure for probabilistic calibration methods [1,2] (to compute this score, we transformed the belief functions yielded by the evidential calibration into probability distributions using the pignistic transformation). We limited the number of queries to 20. The whole process was repeated for 100 rounds of random partitioning, and the obtained results were averaged over the rounds and then over the 6 datasets. These averages are presented in Figure 2. As in the previous experiment, HS is better than PS to improve uncertainty, and this time PS is always better uncertainty-wise than RS. In addition, HS is better with respect to the Brier score than PS, which in turn improves upon RS. Overall, this experiment indicates that both strategies HS and PS may improve the uncertainty as well as the Brier score in comparison to RS, and that HS may be a better choice than PS.

4 Conclusions

In this paper, the benefits of two active strategies with respect to reducing the uncertainty (and also improving the performance) of the evidential calibration of binary classifiers were investigated. Preliminary experiments showed that while the Pignistic sampling strategy may be beneficial, it may be surpassed by Hartley sampling. Future works include conducting more extensive experiments (with other classifiers, datasets, calibration methods, training sets and pool sizes) to refine these conclusions, finding theoretical explanations for them in the spirit of those existing in active learning [5] and applying the approach to a driver state detection system whose calibration data are costly.

Acknowledgements This work is funded in part by the ELSAT2020 project, which is co-financed by the European Union with the European Regional Development Fund, the French state and the Hauts de France Region Council.



(a) Average uncertainty (b) Brier score

Fig. 2: Comparison of strategies HS (solid blue), PS (dash-dot red) and RS (dotted green) for an ensemble of SVM classifiers.

References

1. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proc. of ICML. (2005) 625–632
2. Zhong, W., Kwok, J.T.: Accurate probability calibration for multiple classifiers. In: Proc. of IJCAI. (2013) 1939–1945
3. Xu, P., Davoine, F., Zha, H., Denoeux, T.: Evidential calibration of binary SVM classifiers. Int. J. Approx. Reason. **72** (2016) 55–70
4. Minary, P., Pichon, F., Mercier, D., Lefevre, E., Droit, B.: Evidential joint calibration of binary SVM classifiers using logistic regression. In: Proc. of SUM. (2017) 405–411
5. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
6. Kanjanatarakul, O., Sriboonchitta, S., Denœux, T.: Forecasting using belief functions: an application to marketing econometrics. Int. J. Approx. Reason. **55**(5) (2014) 1113–1128
7. Kanjanatarakul, O., Denœux, T., Sriboonchitta, S.: Prediction of future observations using belief functions: A likelihood-based approach. Int. J. Approx. Reason. **72** (2016) 71–94
8. Denoeux, T.: Likelihood-based belief function: Justification and some extensions to low-quality data. Int. J. Approx. Reason. **55**(7) (2014) 1535–1547
9. Klir, G.J.: Uncertainty and Information: Foundations of Generalized Information Theory. John Wiley & Sons, Inc (2005)
10. Smets, P., Kennes, R.: The transferable belief model. Artif. Intell. **66**(2) (1994) 191–234
11. Reineking, T.: Active classification using belief functions and information gain maximization. Int. J. Approx. Reason. **72** (2016) 43–54
12. Teugels, J.L.: Some representation of the multivariate Bernoulli and binomial distributions. J. Multivar. Anal. **32** (1990) 256–268