Reliable Multi-class Classification based on Pairwise Epistemic and Aleatoric Uncertainty

Vu-Linh Nguyen[1], **Sébastien Destercke**[1], **Marie-Hélène Masson**[1 2], **Eyke Hüllermeier**[3]
[1] UMR CNRS 7253 Heudiasyc, Sorbonne universités,
Université de technologie de Compiègne CS 60319 - 60203 Compiègne cedex, France
[2] Université de Picardie Jules Verne, France
[3] Department of Computer Science, Paderborn University, 33098 Paderborn, Germany
linh.nguyen@hds.utc.fr, sebastien.destercke@hds.utc.fr, mmasson@hds.utc.fr, eyke@upb.de

Abstract

We propose a method for reliable prediction in multi-class classification, where reliability refers to the possibility of partial abstention in cases of uncertainty. More specifically, we allow for predictions in the form of preorder relations on the set of classes, thereby generalizing the idea of set-valued predictions. Our approach relies on combining learning by pairwise comparison with a recent proposal for modeling uncertainty in classification, in which a distinction is made between reducible (a.k.a. *epistemic*) uncertainty caused by a lack of information and irreducible (a.k.a. *aleatoric*) uncertainty due to intrinsic randomness. The problem of combining uncertain pairwise predictions into a most plausible preorder is then formalized as an integer programming problem. Experimentally, we show that our method is able to appropriately balance reliability and precision of predictions.

1 Introduction

Classification algorithms are usually designed to produce "point predictions" in the form of single class labels. In cases of uncertainty, however, it might be more desirable to provide imprecise (or indeterminate) set-valued predictions, which are reliable in the sense of covering the true class with high probability—very much in the spirit of confidence intervals known from classical statistics, or credible sets in Bayesian analysis. This is especially true in safety-critical applications, such as medical diagnosis. Needless to say, the construction of appropriate imprecise predictions requires a suitable quantification of the underlying uncertainty.

In the literature on uncertainty modeling, a general distinction is made between *epistemic uncertainty*, due to a lack of knowledge, and *aleatoric uncertainty*, due to inherent randomness. Thus, while epistemic uncertainty is in principle reducible, aleatoric uncertainty is not. [Senge *et al.*, 2014] recently argued that a distinction between these two types of uncertainty is useful in machine learning, too, where epistemic uncertainty is typically caused by a limited amount of training data, whereas aleatoric uncertainty is due to overlapping class distributions (leading to almost equal posterior probabilities of several classes).

In this paper, we propose a method for producing reliable or "cautious" predictions in the form of preorder relations R on the set of classes $\Omega = \{\lambda_1, \ldots, \lambda_M\}$, where $(\lambda_i, \lambda_j) \in R$ suggests that, for an instance $\boldsymbol{x}$, λ_i is (weakly) preferred to λ_j as a prediction. To this end, we build on the approach of [Senge *et al.*, 2014], which we lift from binary to multi-class classification using appropriate decomposition techniques [Bishop, 2006; Hüllermeier and Brinker, 2007]. Roughly speaking, preferences and uncertainties are first predicted for each pair of classes, and then combined into an overall relation R, in which aleatoric uncertainty translates into equivalence (indifference) between classes ($(\lambda_i, \lambda_j) \in R$ and $(\lambda_j, \lambda_i) \in R$), and epistemic uncertainty into incomparability (neither $(\lambda_i, \lambda_j) \in R$ nor $(\lambda_j, \lambda_i) \in R$). The approach we use for combining the pairwise information into a most plausible preorder is inspired by [Masson *et al.*, 2016] and formalized as an integer linear programming problem.

As we said, our main goal is to produce predictions in the form of (credal) sets of classes. Hence, in a final step, we map a preorder to its set of maximal elements. Yet, let us emphasize that a preorder provides very rich information, which could also be used for other purposes. In particular, this type of prediction can be seen as a generalization of set-valued predictions (a subset defining a specific partial order), which in turn can be seen as an extension of classification with reject option (where the prediction is either a singleton set or the entire Ω). In fact, a preorder provides a flexible means for expressing preferences for class predictions, and for partially abstaining from a definite decision. Due to the distinction between incomparability and indifference, it also offers an explanation for why a class is present or not in a prediction.

The rest of the paper is organized as follows. In the next section, we recall a method for reliable binary classification, which allows for distinguishing between aleatoric and epistemic uncertainty. In Section 3, we introduce our method for reliable multi-class classification based on preorder prediction. This approach is instantiated for logistic regression as a base learner in Section 4. Related work is briefly addressed in Section 5, and experimental results are presented in Section 6, prior to concluding the paper in Section 7.

2 Epistemic and Aleatoric Uncertainty

A main building block of our method is the assessment of the epistemic and aleatoric uncertainty involved in the discrimination between a *pair* of classes. To this end, we are going to adopt the formal model proposed in [Senge *et al.*, 2014], which is based on the use of relative likelihoods, historically proposed by Birnbaum [Birnbaum, 1962] and then justified in other settings such as possibility theory [Walley and Moral, 1999]. For the sake of completeness, the essence of this approach is briefly recalled in the remainder of this section.

The approach proceeds from an instance space $\mathcal{X}$, an output space $\Omega = \{0, 1\}$ encoding the two classes, and a hypothesis space $\mathcal{H}$ consisting of probabilistic classifiers $h : \mathcal{X} \longrightarrow [0, 1]$. We denote by $p_h(1 \,|\, \boldsymbol{x}) = h(\boldsymbol{x})$ and $p_h(0 \,|\, \boldsymbol{x}) = 1 - h(\boldsymbol{x})$ the (predicted) probability that instance $\boldsymbol{x} \in \mathcal{X}$ belongs to the positive and negative class, respectively. Given a set of training data $\mathbf{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \Omega$, the normalized likelihood of a model h is defined as

$$\pi_{\mathcal{H}}(h) = \frac{L(h)}{L(h^{ml})} = \frac{L(h)}{\max_{h' \in \mathcal{H}} L(h')} \ , \tag{1}$$

where $L(h) = \prod_{i=1}^N p_h(y_i \,|\, \boldsymbol{x})$ is the likelihood of h, and $h^{ml} \in \mathcal{H}$ the maximum likelihood estimation on the training data. For a given instance $\boldsymbol{x}$, the degrees of support (plausibility) of the two classes are defined as follows:

$$\begin{aligned} \pi(1 \,|\, \boldsymbol{x}) &= \sup_{h \in \mathcal{H}} \min \big[\pi_{\mathcal{H}}(h), p_h(1 \,|\, \boldsymbol{x}) - p_h(0 \,|\, \boldsymbol{x})\big], && (2) \\ \pi(0 \,|\, \boldsymbol{x}) &= \sup_{h \in \mathcal{H}} \min \big[\pi_{\mathcal{H}}(h), p_h(0 \,|\, \boldsymbol{x}) - p_h(1 \,|\, \boldsymbol{x})\big]. && (3) \end{aligned}$$

So, $\pi(1 \,|\, \boldsymbol{x})$ is high if and only if a highly plausible model supports the positive class much stronger (in terms of the assigned probability mass) than the negative class (and $\pi(0 \,|\, \boldsymbol{x})$ can be interpreted analogously)[1]. Note that, with $f(a) = 2a - 1$, we can also rewrite (2)-(3) as follows:

$$\begin{aligned} \pi(1 \,|\, \boldsymbol{x}) &= \sup_{h \in \mathcal{H}} \min \big[\pi_{\mathcal{H}}(h), f(h(\boldsymbol{x}))\big], && (4) \\ \pi(0 \,|\, \boldsymbol{x}) &= \sup_{h \in \mathcal{H}} \min \big[\pi_{\mathcal{H}}(h), f(1 - h(\boldsymbol{x}))\big]. && (5) \end{aligned}$$

Given the above degrees of support, the degrees of epistemic uncertainty u_e and aleatoric uncertainty u_a are defined as follows:

$$\begin{aligned} u_e(\boldsymbol{x}) &= \min \big[\pi(1 \,|\, \boldsymbol{x}), \pi(0 \,|\, \boldsymbol{x})\big] \,, && (6) \\ u_a(\boldsymbol{x}) &= 1 - \max \big[\pi(1 \,|\, \boldsymbol{x}), \pi(0 \,|\, \boldsymbol{x})\big] \,. && (7) \end{aligned}$$

Thus, epistemic uncertainty refers to the case where both the positive and the negative class appear to be plausible, while the degree of aleatoric uncertainty (7) is the degree to which none of the classes is supported. These uncertainty degrees are completed with degrees $s_1(\boldsymbol{x})$ and $s_0(\boldsymbol{x})$ of (strict) preference in favor of the positive and negative class, respectively:

$$s_1(\boldsymbol{x}) = \begin{cases} 1 - (u_a(\boldsymbol{x}) + u_e(\boldsymbol{x})) & \text{if } \pi(1 \,|\, \boldsymbol{x}) > \pi(0 \,|\, \boldsymbol{x}), \\ \frac{1 - (u_a(\boldsymbol{x}) + u_e(\boldsymbol{x}))}{2} & \text{if } \pi(1 \,|\, \boldsymbol{x}) = \pi(0 \,|\, \boldsymbol{x}), \\ 0 & \text{if } \pi(1 \,|\, \boldsymbol{x}) < \pi(0 \,|\, \boldsymbol{x}). \end{cases}$$

With an analogous definition for $s_0(\boldsymbol{x})$, we have $s_0(\boldsymbol{x}) + s_1(\boldsymbol{x}) + u_a(\boldsymbol{x}) + u_e(\boldsymbol{x}) \equiv 1$. In addition to defining a partition of unity, the quadruple $(s_1(\boldsymbol{x}), s_0(\boldsymbol{x}), u_e(\boldsymbol{x}), u_a(\boldsymbol{x}))$ has the following properties:

- $s_1(\boldsymbol{x})$ ($s_0(\boldsymbol{x})$) will be high if and only if, for all plausible models, the probability of the positive (negative) class is significantly higher than the one of the negative (positive) class;
- $u_e(\boldsymbol{x})$ will be high if class probabilities strongly vary within the set of plausible models, i.e., if we are unsure how to compare these probabilities. In particular, it will be 1 if and only if we have $h(\boldsymbol{x}) = 1$ and $h'(\boldsymbol{x}) = 0$ for two totally plausible models;
- $u_a(\boldsymbol{x})$ will be high if class probabilities are similar for all plausible models, i.e., if there is strong evidence that $h(\boldsymbol{x}) \approx 0.5$. In particular, it will be close to 1 if all plausible models allocate their probability mass around $h(\boldsymbol{x}) = 0.5$.

3 Reliable Multi-class Prediction

In this section, we present our approach to reliable multi-class prediction, which is based on the idea of binary decomposition and a stepwise simplification (approximation) of the information contained in the set of pairwise comparisons between classes—first in terms of a preorder and then in terms of a set.

3.1 Learning by Pairwise Comparison

In the multi-class setting, we are dealing with a set of $M > 2$ classes $\Omega = \{\lambda_1, \ldots, \lambda_M\}$. Suppose a set of training data $\mathbf{D} = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathcal{X} \times \Omega$ to be given, and denote by $\mathbf{D}_m = \{\boldsymbol{x} \,|\, (\boldsymbol{x}, \lambda_m) \in \mathbf{D}\}$ the observations from class λ_m.

Learning by pairwise comparison (LPC) a.k.a. all-pairs is a decomposition technique that trains one (binary) classifier $h_{i,j}$ for each pair of classes (λ_i, λ_j), $1 \leq i < j \leq M$ [Fürnkranz, 2002]. The task of $h_{i,j}$, which is trained on $\mathbf{D}_{i,j} = \mathbf{D}_i \cup \mathbf{D}_j$, is to separate instances with label λ_i from those having label λ_j. Suppose we solve these problems with the approach described in the previous section, instead of using a standard binary classifier. Then, given a new query instance $\boldsymbol{x} \in \mathcal{X}$, we can produce predictions in the form of a quadruple

$$\mathbf{I}^{i,j}(\boldsymbol{x}) := \Big(s_{\lambda_i}^{i,j}(\boldsymbol{x}), s_{\lambda_j}^{i,j}(\boldsymbol{x}), u_e^{i,j}(\boldsymbol{x}), u_a^{i,j}(\boldsymbol{x})\Big) \,, \tag{8}$$

one for each pair of classes (λ_i, λ_j). These predictions can also be summarized in three $[0, 1]^{M \times M}$ relations, a (strict) preference relation P, an indifference relation A, and an incomparability relation E:

$$P = \big(s_{\lambda_i}^{i,j}(\boldsymbol{x})\big)_{i,j}, \ A = \big(u_a^{i,j}(\boldsymbol{x})\big)_{i,j} \ , E = \big(s_e^{i,j}(\boldsymbol{x})\big)_{i,j}$$

[1] Technically, we assume that, for each $\boldsymbol{x} \in \mathcal{X}$, there are hypotheses $h, h' \in \mathcal{H}$ such that $h(\boldsymbol{x}) \geq 0.5$ and $h'(\boldsymbol{x}) \leq 0.5$, which implies $\pi(1 \,|\, \boldsymbol{x}) \geq 0$ and $\pi(0 \,|\, \boldsymbol{x}) \geq 0$.

Recall that our approach is transductive in the sense that predictions are always derived per instance, i.e., for an individual query instance $\boldsymbol{x}$. Likewise, all subsequent inference steps are tailored for that instance. Keeping this in mind, we will henceforth simplify notation and often omit the dependence of scores and relations on $\boldsymbol{x}$.

3.2 Inferring a Preorder

The structure (P, A, E) provides a rich source of information, which we seek to represent in a condensed form. To this end, we approximate this structure by a preorder R. This approximation may also serve the purpose of correction, since the relational structure (P, A, E) is not necessarily consistent; for example, since all binary classifiers are trained independently of each other, their predictions are not necessarily transitive.

Recall that a preorder is a binary relation $R \subseteq \Omega \times \Omega$ that is reflexive. In the following, we will also use the following notation:

$$\begin{aligned} &\lambda_i \succ_R \lambda_j \text{ (or simply } \lambda_i \succ \lambda_j) && \text{if} \quad r_{i,j} = 1,\, r_{j,i} = 0\,, \\ &\lambda_i \sim_R \lambda_j \text{ (or simply } \lambda_i \sim \lambda_j) && \text{if} \quad r_{i,j} = 1,\, r_{j,i} = 1\,, \\ &\lambda_i \perp_R \lambda_j \text{ (or simply } \lambda_i \perp \lambda_j) && \text{if} \quad r_{i,j} = 0,\, r_{j,i} = 0\,, \end{aligned}$$

where $r_{i,j} = 1$ if $(\lambda_i, \lambda_j) \in R$ and $r_{i,j} = 0$ if $(\lambda_i, \lambda_j) \notin R$. Note that the binary relations $\succ$, $\sim$, $\perp$ are in direct correspondence with the relations P, A, and E, respectively.

How compatible is a relation R with a structure (P, A, E)? Interpreting the scores (8) as probabilities, we could imagine that a relation R is produced by randomly "hardening" the soft (probabilistic) structure (P, A, E), namely by selecting one of the relations $\lambda_i \succ \lambda_j$, $\lambda_j \succ \lambda_i$, $\lambda_i \sim \lambda_j$, $\lambda_i \perp \lambda_j$ with probability $s^{i,j}_{\lambda_i}$, $s^{i,j}_{\lambda_j}$, $u^{i,j}_a$, and $u^{i,j}_e$, respectively. Then, making a simplifying assumption of independence, the probability of ending up with R is given as follows:

$$p(R) = \prod_{\lambda_i \succ_R \lambda_j} s^{i,j}_{\lambda_i} \prod_{\lambda_j \succ_R \lambda_i} s^{i,j}_{\lambda_j} \prod_{\lambda_i \perp_R \lambda_j} u^{i,j}_e \prod_{\lambda_i \sim_R \lambda_j} u^{i,j}_a \qquad (9)$$

The most probable preorder R^* then corresponds to

$$R^* = \arg\max_{R \in \mathbf{R}} p(R)\,, \qquad (10)$$

where $\mathbf{R}$ is the set of all preorders on Ω.

Let us now propose a practical procedure to determine R^*, which is based on representing the optimization problems (10) as a binary linear integer program. To this end, we introduce the following variables:

$$\begin{aligned} X^1_{i,j} &= r_{i,j}(1 - r_{j,i}), & X^2_{i,j} &= r_{j,i}(1 - r_{i,j}), \\ X^3_{i,j} &= (1 - r_{i,j})(1 - r_{j,i}), & X^4_{i,j} &= r_{i,j} r_{j,i}. \end{aligned}$$

Then, by adding the constraints $\sum_{l=1}^4 X^l_{i,j} = 1$ and $X^l_{i,j} \in \{0, 1\}$, we can rewrite the probability (9) as follows:

$$p(R) = \prod_{i<j} \left(s^{i,j}_{\lambda_i}\right)^{X^1_{i,j}} \left(s^{i,j}_{\lambda_j}\right)^{X^2_{i,j}} \left(u^{i,j}_e\right)^{X^3_{i,j}} \left(u^{i,j}_a\right)^{X^4_{i,j}} \qquad (11)$$

Furthermore, the transitivity property

$$r_{i,k} + r_{k,j} - 1 \leq r_{i,j}, \quad \forall\, i \neq j \neq k. \qquad (12)$$

can be easily be encoded by noting that $r_{i,j} = X^1_{i,j} + X^4_{i,j}$ and $r_{i,j} = X^2_{j,i} + X^4_{j,i}$ if $i < j$ and $j < i$, respectively.

Altogether, the most probable preorder $R^* \in \mathbf{R}$ is determined by $X^* = (X^1_{i,j}, \ldots, X^4_{i,j})_{i,j}$, which is the solution of the following optimization problem:

$$\begin{aligned} \max \sum_{i<j} &X^1_{i,j} \ln\left(s^{i,j}_{\lambda_i}\right) + X^2_{i,j} \ln\left(s^{i,j}_{\lambda_j}\right) \qquad (13) \\ &+ X^3_{i,j} \ln\left(u^{i,j}_e\right) + X^4_{i,j} \ln\left(u^{i,j}_a\right) \\ \text{s.t.} \quad &\sum_{l=1}^4 X^l_{i,j} = 1,\ \forall\, 1 \leq i < j \leq M\,, \\ &X^1_{i,j}, X^2_{i,j}, X^3_{i,j}, X^4_{i,j} \in \{0, 1\},\ \forall\, 1 \leq i < j \leq M\,, \\ &r_{i,k} + r_{k,j} - 1 \leq r_{i,j},\ \forall\, i \neq j \neq k\,. \end{aligned}$$

Note that if $u^{i,j}_e = 0$ for all pairs, then the solution will be a complete preorder between class probabilities, which is consistent with our interpretation. Similarly, if $u^{i,j}_a = 0$ and $u^{i,j}_e = 0$ for all pairs, we would obtain a linear ordering, as in [Cheng and Hüllermeier, 2012].

3.3 Obtaining Credible Sets from R^*

Consider the preorder $R^* = R^*(\boldsymbol{x})$ for an unlabelled query instance $\boldsymbol{x}$, and suppose we seek a set-valued prediction $Y(\boldsymbol{x}) \subseteq \Omega$. A reasonable way to obtain such a prediction is to collect all non-dominated classes, i.e., to exclude only those classes λ_j for which $\lambda_i \succ_{R^*} \lambda_j$ for at least one competing class λ_i. A class label of that kind can be seen as a potentially optimal prediction for $\boldsymbol{x}$. Adopting the above notation, the set-valued prediction can also be determined as

$$Y(\boldsymbol{x}) = \left\{ \lambda_i \in \Omega \,\middle|\, \sum_{j<i} X^1_{j,i} + \sum_{i<j} X^2_{i,j} = 0 \right\}, \qquad (14)$$

which means that it can immediately be derived from the solution of (13). Note that full uncertainty, i.e, $Y(\boldsymbol{x}) = \Omega$, only occur if all pairs (λ_i, λ_j) are incomparable or indifferent.

How to obtain a set-valued prediction from the pairwise information is illustrated in the following example.

Example 1 *Assume that we have the output space* $\Omega = \{\lambda_1, \ldots, \lambda_5\}$ *and pairwise information* (8) *for an unlabelled instance* $\boldsymbol{x}$ *given by the following quadruples:*

$$\begin{aligned} &\mathbf{I}^{1,2}(\boldsymbol{x}) = (0, 0.1, 0.6, 0.3), && \mathbf{I}^{1,3}(\boldsymbol{x}) = (0.6, 0, 0.1, 0.2), \\ &\mathbf{I}^{1,4}(\boldsymbol{x}) = (0.9, 0, 0.1, 0), && \mathbf{I}^{1,5}(\boldsymbol{x}) = (0.4, 0, 0.3, 0.3), \\ &\mathbf{I}^{2,3}(\boldsymbol{x}) = (0.6, 0, 0.2, 0.2), && \mathbf{I}^{2,4}(\boldsymbol{x}) = (0.7, 0, 0, 0.3), \\ &\mathbf{I}^{2,5}(\boldsymbol{x}) = (0.9, 0, 0, 0.1), && \mathbf{I}^{3,4}(\boldsymbol{x}) = (0.6, 0, 0.2, 0.2), \\ &\mathbf{I}^{3,5}(\boldsymbol{x}) = (0.9, 0, 0.1, 0), && \mathbf{I}^{4,5}(\boldsymbol{x}) = (0.05, 0.05, 0.4, 0.5). \end{aligned}$$

Solving the optimization problem (13) *gives the most probable preorder* R^* *pictured in Figure 1 with the corresponding value* X^* *s.t.* $X^3_{1,2} = X^1_{1,3} = X^1_{2,3} = X^1_{3,4} = X^1_{3,5} = X^4_{4,5} = 1$. *Finally, from* (14) *we get* $Y(\boldsymbol{x}) = \{1, 2\}$.

4 Instantiation for Logistic Regression

Our approach outlined so far needs to be instantiated with a concrete hypothesis space $\mathcal{H}$, i.e., a class of binary models h

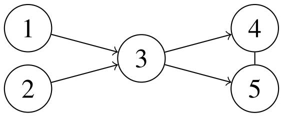


Figure 1: Preorder induced by Example 1 (strict preference symbolized by directed edge, indifference by undirected edge, incomparability by missing edge).

used to derive plausibility degrees (4) and (5), and a related learning algorithm. Here, we present an efficient instantiation for logistic regression. In general, estimating (4) and (5) can be difficult, and may require the use of approximation techniques (e.g., by fitting a surrogate model for $\pi_{\mathcal{H}}(h)$).

Recall that logistic regression assumes posterior probabilities to depend on feature vectors $\boldsymbol{x} = (x^1, \ldots, x^d) \in \mathbb{R}^d$ in the following way:

$$h(\boldsymbol{x}) = p(1 \,|\, \boldsymbol{x}) = \frac{\exp\left(\beta_0 + \sum_{i=1}^{d} \beta_i \, x^i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{d} \beta_i \, x^i\right)} \tag{15}$$

This means that learning the model comes down to estimating a parameter vector $\beta = (\beta_0, \ldots, \beta_d)$, which is commonly done through likelihood maximization [Czepiel, 2002]. To avoid numerical issues (e.g, having to deal with the exponential function for large β) when maximizing the target function, we employ L_2 regularization. The corresponding version of the log-likelihood function (16) is strictly concave [Rennie, 2005]:

$$\begin{aligned} l(\beta) = & \sum_{n=1}^{N} y_n \left(\beta_0 + \sum_{i=1}^{d} \beta_i x_n^i\right) \qquad (16) \\ & - \sum_{n=1}^{N} \ln\left(1 + \exp\left(\beta_0 + \sum_{i=1}^{d} \beta_i x_n^i\right)\right) - \frac{\gamma}{2} \sum_{i=0}^{d} \beta_i^2, \end{aligned}$$

where the regularization term γ will be fixed to 1.

We now focus on determining the degree of support (4) for the positive class, and then summarize the results for the negative class (which can be determined in a similar manner). Associating each hypothesis $h \in \mathcal{H}$ with a vector $\beta \in \mathbb{R}^{d+1}$, the degree of support (4) can be rewritten as follows:

$$\pi(1 \,|\, \boldsymbol{x}) = \sup_{\beta \in \mathbb{R}^{d+1}} \min\left[\pi(\beta), 2h(\boldsymbol{x}) - 1\right] \tag{17}$$

It is easy to see that the target function to be maximized in (17) is not necessarily concave. Therefore, we propose the following approach.

Let us first note that whenever $h(\boldsymbol{x}) < 0.5$, we have $2h(\boldsymbol{x}) - 1 \leq 0$ and $\min\left[\pi_{\mathcal{H}}(h), 2h(\boldsymbol{x}) - 1\right] \leq 0$. Thus the optimal value of the target function (4) can only be achieved for some hypotheses h such that $h(\boldsymbol{x}) \in [0.5, 1]$. For a given value $\alpha \in [0.5, 1]$, the set of hypotheses h such that $h(\boldsymbol{x}) = \alpha$ corresponds to the convex set

$$\boldsymbol{\beta}^{\alpha} = \left\{\beta \,\middle|\, \beta_0 + \sum_{i=1}^{d} \beta_i x^i = \ln\left(\frac{\alpha}{1-\alpha}\right)\right\}. \tag{18}$$

The optimal value $\pi_\alpha^*(1 \,|\, \boldsymbol{x})$ that can be achieved within the region (18) can be determined as follows:

$$\begin{aligned} \pi_\alpha^*(1 \,|\, \boldsymbol{x}) &= \sup_{\beta \in \boldsymbol{\beta}^{\alpha}} \min\left[\pi(\beta), 2\alpha - 1\right] \qquad (19) \\ &= \min\left[\sup_{\beta \in \boldsymbol{\beta}^{\alpha}} \pi(\beta), 2\alpha - 1\right] \end{aligned}$$

Thus, to find this value, we maximize the concave log-likelihood over a convex set:

$$\beta_\alpha^* = \arg \sup_{\beta \in \boldsymbol{\beta}^{\alpha}} l(\beta) \tag{20}$$

As the log-likelihood function (16) is concave and has second-order derivatives, we tackle the problem with a Newton-CG algorithm [Nocedal and Wright, 2006]. Furthermore, the optimization problem (20) can be solved using sequential least squares programming[2] [Philip and Elizabeth, 2010]. Since regions defined in (18) are parallel hyperplanes, the solution of the optimization problem (4) can then be obatined by solving the following problem:

$$\sup_{\alpha \in [0.5,1)} \pi_\alpha^*(1|\boldsymbol{x}) = \sup_{\alpha \in [0.5,1)} \min\left[\pi(\beta_\alpha^*), 2\alpha - 1\right] \tag{21}$$

Following a similar procedure, we can estimate the degree of support for the negative class (5) as follows:

$$\sup_{\alpha \in (0,0.5]} \pi_\alpha^*(0|\boldsymbol{x}) = \sup_{\alpha \in (0,0.5]} \min\left[\pi(\beta_\alpha^*), 1 - 2\alpha\right] \tag{22}$$

Note that limit cases $\alpha = 1$ and $\alpha = 0$ cannot be solved, since the region (18) is then not well-defined (as $\ln(\infty)$ and $\ln(0)$ do not exist). For the purpose of practical implementation, we handle (21) by discretizing the interval over α. That is, we optimize the target function for a given number of values $\alpha \in [0.5, 1)$ and consider the solution corresponding to the α with the highest optimal value of the target function $\pi_\alpha^*(1 \,|\, \boldsymbol{x})$ as the maximum estimator. Similarly, (22) can be handled over the domain $(0, 0.5]$. In practice, we evaluate (21) and (22) on uniform discretizations of cardinality 50 of $[0.5, 1)$ and $(0, 0.5]$, respectively. We can further increase efficiency by avoiding computations for values of α for which we know that $2\alpha - 1$ and $1 - 2\alpha$ are lower than the current highest support value given to class 1 and 0, respectively.

5 Related Work

In the literature, different approaches have been proposed to produce credible (set-valued) predictions in the setting of supervised learning.

In [Coz *et al.*, 2009], the authors proceed from standard probabilistic predictions, i.e., a probability distribution on the set of classes (conditioned on an instance $\boldsymbol{x}$). To produce a set-valued prediction, they invoke the principle of expected cost minimization, where the underlying cost measure combines the precision and correctness of the prediction (based on the F-measure as a performance metric). As an advantage of this approach, note that it can be used with standard methods for probabilistic prediction.

[2] For an implementation in Python, see `https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html`

#	name	# instances	# features	# labels
a	iris	150	4	3
b	wine	178	13	3
c	forest	198	27	4
d	seeds	210	7	3
e	glass	214	9	6
f	ecoli	336	7	8
g	libras	360	91	15
h	dermatology	385	34	6
i	vehicle	846	18	4
j	vowel	990	10	11
k	yeast	1484	8	12
l	wine quality	1599	11	6
m	optdigits	1797	64	10
n	segment	2300	19	7
o	wall-following	5456	24	4

Table 1: Data sets used in the experiments

Closer to our approach are methods based on imprecise probabilities, such as [Corani *et al.*, 2014], which augment probabilistic predictions into probability intervals or sets of probabilities, the size of which reflects the lack of information (reflecting epistemic uncertainty). Similar to this are approaches based on confidence bands in calibration models, for instance [Kull and Flach, 2014; Xu *et al.*, 2016]. They usually control the amount of imprecision by adjusting some certain parameters, e.g., a confidence value.

Conformal prediction [Shafer and Vovk, 2008; Balasubramanian *et al.*, 2014] is a generic approach to reliable (set-valued) prediction that combines ideas from probability theory (specifically the principle of exchangeability), statistics (hypothesis testing, order statistics), and algorithmic complexity. The basic version of conformal prediction is designed for sequential prediction in an online setting, and comes with certain correctness guarantees (predictions are correct with probability $1-\epsilon$, where ϵ is a confidence parameter). Roughly speaking, given an instance $\boldsymbol{x}$, it assigns a *non-conformity* score to each candidate output. Then, considering each of these outcomes as a hypothesis, those outcomes for which the hypothesis can be rejected with high confidence are eliminated. The set-valued prediction is given by the set of those outcomes that cannot be rejected.

6 Experiments

This section presents first experimental results to assess the performance of our approach to reliable classification.

6.1 Data Sets and Experimental Setting

We perform experiments on 15 data sets from the UCI repository (cf. Table 1), following a 10 × 10-fold cross-validation procedure. We compare the performance of our method (referred to as PREORDER) with two competitors. To make the results as comparable as possible, these methods are also implemented with pairwise learning using logistic regression as base learner. Thus, they essentially only differ in how the pairwise information provided by logistic regression is turned into a (reliable) multi-class prediction.

- VOTE: The first method is based on aggregating pairwise predictions via standard voting, which is a common approach in LPC. However, instead of simple weighted voting, we apply the more sophisticated aggregation technique proposed in [Hüllermeier and Vanderlooy, 2010], which shows better performance. Note that, by predicting the winner of the voting procedure, this approach always produces a precise prediction.
- NONDET: As a baseline for set-valued predictions, we use the approach of [Coz *et al.*, 2009], which has been shown to exhibit competitive performance in comparison to other imprecise prediction methods [Zaffalon *et al.*, 2012]. Recall that this approach produces nondeterministic predictions from precise probabilistic assessments. This requires turning pairwise probability estimates into conditional probabilities $(p(\lambda_1 \,|\, \boldsymbol{x}), \ldots, p(\lambda_M \,|\, \boldsymbol{x}))$ on the classes, a problem known as pairwise coupling. To this end, we apply the δ_2 method, which performs best among those investigated in [Wu *et al.*, 2004].

Evaluation metrics for assessing set-valued predictions have to balance correctness (the true class y is an element of the predicted set Y) and precision (size of the predicted set) in an appropriate manner. For example, in [Zaffalon *et al.*, 2012], the authors argue that using the simple discounted accuracy ($1/|Y|$ if $y \in Y$ and 0 otherwise) is equivalent to saying that producing a set-valued prediction is the same as choosing within this set (uniformly) at random. This means that the discounted accuracy does not reward any cautiousness. Also, it can be shown that minimizing the expected discounted accuracy in expectation would never lead to imprecise predictions [Yang *et al.*, 2017]. Here, we therefore adopt the *average utility-discounted accuracy* measure, which has been proposed and formally justified in [Zaffalon *et al.*, 2012]:

$$u(y, Y) = \begin{cases} 0 & \text{if } y \notin Y \\ \dfrac{\alpha}{|Y|} - \dfrac{\beta}{|Y|^2} & \text{otherwise} \end{cases}$$

More specifically, we use the measures u_{65} with $(\alpha, \beta) = (1.6, 0.6)$ and u_{80} with $(\alpha, \beta) = (2.2, 1.2)$. Note that, in the case of precise decisions, both u_{65} and u_{80} reduce to standard accuracy.

6.2 Experimental Results

The average performances in terms of the utility-discounted accuracies are shown in Table 2, with ranks in parenthesis (note that we provide one set of ranks for u_{65}, and another one for u_{80}). Firstly, we notice that PREORDER yields the best average ranks over the 15 data sets, both for u_{80} and u_{65}. Furthermore, a Friedman test [Demšar, 2006] on the ranks yields p-values of 0.0003138 and 0.002319 for u_{80} and u_{65}, respectively, thus strongly suggesting performance differences between the algorithms. The Nemenyi post-hoc test (see Table 3) further indicates that PREORDER is significantly better than VOTE regarding u_{80} and NONDET in the case of u_{65}. Since u_{80} rewards cautious predictions stronger than u_{65} does, it is not surprising that indeterminate classifiers

	VOTE	PREORDER		NONDET	
#	acc.	u_{80}	u_{65}	u_{80}	u_{65}
a	$84.33(3,1)$	$90.45(1)$	$83.29(2)$	$86.71(2)$	$76.88(3)$
b	$96.35(1,1)$	$95.89(2)$	$93.18(2)$	$93.47(3)$	$88.92(3)$
c	$89.76(2,1)$	$92.15(1)$	$88.82(2)$	$88.49(3)$	$81.57(3)$
d	$88.81(3,1)$	$92.15(1)$	$88.16(2)$	$90.03(2)$	$83.60(3)$
e	$47.14(3,3)$	$67.32(1)$	$57.24(1)$	$65.03(2)$	$52.98(2)$
f	$75.57(3,1)$	$80.66(1)$	$75.25(2)$	$77.02(2)$	$68.89(3)$
g	$50.50(3,3)$	$70.51(1)$	$63.91(1)$	$62.50(2)$	$53.02(2)$
h	$96.43(2,2)$	$97.70(1)$	$96.46(1)$	$96.01(3)$	$93.38(3)$
i	$63.99(3,1)$	$71.07(1)$	$62.17(2)$	$68.92(2)$	$57.17(3)$
j	$39.57(3,2)$	$51.10(1)$	$42.57(1)$	$48.22(2)$	$37.27(3)$
k	$49.35(3,2)$	$60.60(2)$	$50.04(1)$	$60.84(1)$	$49.22(3)$
l	$58.10(3,3)$	$69.65(2)$	$59.92(1)$	$71.02(1)$	$59.16(2)$
m	$96.37(3,2)$	$97.67(1)$	$96.81(1)$	$96.85(2)$	$95.46(3)$
n	$84.51(3,3)$	$91.87(1)$	$89.16(1)$	$90.01(2)$	$85.49(2)$
o	$68.69(3,3)$	$76.42(2)$	$70.79(1)$	$77.34(1)$	$70.39(2)$
aver.	(u_{80}, u_{65})	u_{80}	u_{65}	u_{80}	u_{65}
rank	$(2.73, 1.93)$	1.27	1.40	2.00	2.67

Table 2: Average utility-discounted accuracies (%)

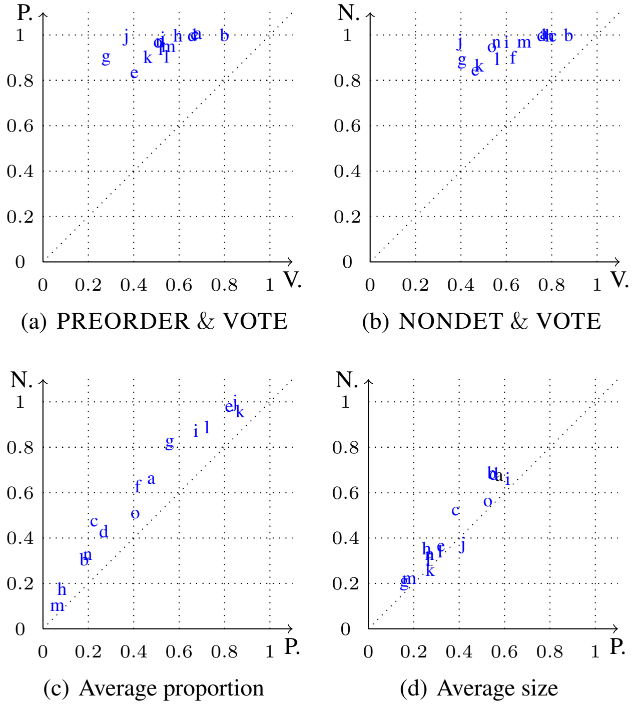
#	H_0	u_{80}	u_{65}
1	$V = P$	**0.00017**	0.3101
2	$V = N$	0.11017	0.1102
3	$P = N$	0.11017	**0.0015**

Table 3: Nemenyi post-hoc test: null hypothesis H_0 and p-value

do better in this case. Yet, even when considering u_{65}, PREORDER remains competitive with VOTE. This suggests that it tends to be more precise than NONDET, while still accurately recognizing those instances for which we have to be cautious.

Ideally, an imprecise classifier should abstain (i.e., provide set-valued predictions) on difficult cases, on which the precise classifier is likely to fail [Yang *et al.*, 2014]. The goal of Figure 2(a,b) is to verify this ability. Figure 2(a) displays, for each data set, the percentage of times the true class is in the prediction of PREORDER, given the prediction was imprecise, versus the accuracy of VOTE on those instances. Figure 2(b) does the same for NONDET. Both imprecise classifiers achieve high percentages (> 80) of correct partial predictions, while the corresponding percentages of VOTE vary in a wider range. Also, the accuracy of the latter significantly drops on those instances (for example, the average accuracy for data set g is 50% in Table 2, but drops to less than 30% in Figure 2(a)), confirming that the imprecise classifiers do indeed abstain on difficult cases. Finally, note that the points in Figure 2(a) are a bit more to the left than those in Figure 2(b), again suggesting that PREORDER is doing slightly better in recognizing difficult instances than NONDET.

For the two imprecise classifiers, we also compare the average proportion of partial predictions and the average (normalized) size of the predictions when at least one method produces a partial prediction. Figures 2(c) and 2(d) indicate that NONDET produces more partial predictions of (slightly) larger size.



(a) PREORDER & VOTE

(b) NONDET & VOTE

(c) Average proportion

(d) Average size

Figure 2: (a) Correctness of the PREORDER in the case of abstention versus accuracy of the VOTE. (b) Correctness of the NONDET in the case of abstention versus accuracy of the VOTE. (c) Proportion of partial predictions when at least one method produces a partial prediction. (d) Average normalized size of the predictions in such cases.

7 Conclusion

This paper introduces an approach to cautious inference and reliable prediction in multi-class classification. The basic idea is to provide predictions in the form of preorder relations, which allow for representing preferences for some candidate classes over others, as well as indifference and incomparability between them; the two latter relations are in direct correspondence with two types of uncertainty, aleatoric and epistemic. This can be seen as a sophisticated way of partial abstention, which generalizes set-valued predictions and classification with reject option. Technically, our approach combines reliable binary classification with pairwise decomposition and approximate inference over preorders.

Practically, by projecting to the set of maximal elements, we only used preorder predictions for the purpose of set-valued classification. Our experiments on this type of problem are quite promising and suggest that our method is highly competitive to existing approaches to reliable prediction.

In future work, we plan to exploit more of the potential of preorder predictions, and to use such predictions in other contexts and problem settings. In active learning, for example, preorder predictions may provide very useful information for guiding the selection of queries. Since our approach applies as soon as a likelihood is defined, we also plan to study its extension to other kinds of likelihood such as evidential ones [Denoeux, 2014].

Acknowledgments

This work was carried out in the framework of Labex MS2T and UML-NET projects, funded by the French Government (Reference ANR-11-IDEX-0004-02, ANR-14-CE24-0026), and was partially supported by the German Research Foundation (DFG, SFB 901).

References

[Balasubramanian *et al.*, 2014] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: Theory, Adaptations and applications*. Newnes, 2014.

[Birnbaum, 1962] Allan Birnbaum. On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306, 1962.

[Bishop, 2006] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[Cheng and Hüllermeier, 2012] Weiwei Cheng and Eyke Hüllermeier. Probability estimation for multi-class classification based on label ranking. In *Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases*, pages 83–98. Springer-Verlag, 2012.

[Corani *et al.*, 2014] Giorgio Corani, Joaquín Abellán, Andrés Masegosa, Serafin Moral, and Marco Zaffalon. *Classification*, pages 230–257. John Wiley & Sons, Ltd, 2014.

[Coz *et al.*, 2009] Juan José del Coz, Jorge Díez, and Antonio Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10(Oct):2273–2293, 2009.

[Czepiel, 2002] Scott A Czepiel. Maximum likelihood estimation of logistic regression models: theory and implementation. *http://czep.net*, 2002.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(Jan):1–30, 2006.

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

[Fürnkranz, 2002] Johannes Fürnkranz. Round robin classification. *Journal of Machine Learning Research*, 2(Mar):721–747, 2002.

[Hüllermeier and Brinker, 2007] Eyke Hüllermeier and Klaus Brinker. Fuzzy-relational classification: Combining pairwise decomposition techniques with fuzzy preference modeling. In *Proceedings of the 5th Conference of the European Society for Fuzzy Logic and Technology*, pages 353–360, 2007.

[Hüllermeier and Vanderlooy, 2010] Eyke Hüllermeier and Stijn Vanderlooy. Combining predictions in pairwise classification: An optimal adaptive voting strategy and its relation to weighted voting. *Pattern Recognition*, 43(1):128–142, 2010.

[Kull and Flach, 2014] Meelis Kull and Peter Flach. Reliability maps: a tool to enhance probability estimates and improve classification accuracy. In *Proceedings of the 2014 European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 18–33. Springer-Verlag, 2014.

[Masson *et al.*, 2016] Marie-Hélène Masson, Sébastien Destercke, and Thierry Denoeux. Modelling and predicting partial orders from pairwise belief functions. *Soft Computing*, 20(3):939–950, 2016.

[Nocedal and Wright, 2006] Jorge Nocedal and S Wright. *Numerical Optimization*. Springer New York, 2006.

[Philip and Elizabeth, 2010] E Philip and WONG Elizabeth. Sequential quadratic programming methods. *UCSD Department of Mathematics Technical Report NA-10-03*, 2010.

[Rennie, 2005] Jason DM Rennie. Regularized logistic regression is strictly convex. *Technical report, MIT*, 2005.

[Senge *et al.*, 2014] Robin Senge, Stefan Bösner, Krzysztof Dembczyński, Jörg Haasenritter, Oliver Hirsch, Norbert Donner-Banzhoff, and Eyke Hüllermeier. Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty. *Information Sciences*, 255:16–29, 2014.

[Shafer and Vovk, 2008] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(Mar):371–421, 2008.

[Walley and Moral, 1999] Peter Walley and Serafin Moral. Upper probabilities based only on the likelihood function. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):831–847, 1999.

[Wu *et al.*, 2004] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.

[Xu *et al.*, 2016] Philippe Xu, Franck Davoine, Hongbin Zha, and Thierry Denoeux. Evidential calibration of binary svm classifiers. *International Journal of Approximate Reasoning*, 72:55–70, 2016.

[Yang *et al.*, 2014] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. Nested dichotomies with probability sets for multi-class classification. In *Proceedings of the Twenty-first European Conference on Artificial Intelligence*, pages 363–368. IOS Press, 2014.

[Yang *et al.*, 2017] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. The costs of indeterminacy: How to determine them? *IEEE Transactions on Cybernetics*, 47(12):4316–4327, 2017.

[Zaffalon *et al.*, 2012] Marco Zaffalon, Giorgio Corani, and Denis Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.