Journal of Classification (2024) 41:2–37
https://doi.org/10.1007/s00357-023-09455-x

ORIGINAL RESEARCH

Classification Under Partial Reject Options

Måns Karlsson[1] · **Ola Hössjer**[1]

Accepted: 23 October 2023 / Published online: 25 November 2023

Abstract
In many applications there is ambiguity about which (if any) of a finite number N of hypotheses that best fits an observation. It is of interest then to possibly output a whole set of categories, that is, a scenario where the size of the classified set of categories ranges from 0 to N. Empty sets correspond to an outlier, sets of size 1 represent a firm decision that singles out one hypothesis, sets of size N correspond to a rejection to classify, whereas sets of sizes $2, \ldots, N-1$ represent a partial rejection to classify, where some hypotheses are excluded from further analysis. In this paper, we review and unify several proposed methods of Bayesian set-valued classification, where the objective is to find the optimal Bayesian classifier that maximizes the expected reward. We study a large class of reward functions with rewards for sets that include the true category, whereas additive or multiplicative penalties are incurred for sets depending on their size. For models with one homogeneous block of hypotheses, we provide general expressions for the accompanying Bayesian classifier, several of which extend previous results in the literature. Then, we derive novel results for the more general setting when hypotheses are partitioned into blocks, where ambiguity within and between blocks are of different severity. We also discuss how well-known methods of classification, such as conformal prediction, indifference zones, and hierarchical classification, fit into our framework. Finally, set-valued classification is illustrated using an ornithological data set, with taxa partitioned into blocks and parameters estimated using MCMC. The associated reward function's tuning parameters are chosen through cross-validation.

Keywords Blockwise cross-validation · Bayesian classification · Conformal prediction · Classes of hypotheses · Indifference zones · Markov Chain Monte Carlo · Reward functions with set-valued inputs · Set-valued classifiers

1 Introduction

Classification of observations among a finite number N of hypotheses or categories is a well-studied problem in statistics, and also a central concept of modern machine learning (Bishop, 2006; Hastie et al., 2009). For the most widely used approach, Bayesian decision theory, it is

✉ Ola Hössjer
ola@math.su.se

Måns Karlsson
mok.fbo@gmail.com

[1] Department of Mathematics, Stockholm University, 106 91 Stockholm, Sweden

well known that the maximum a posteriori classifier maximizes the probability of a correct classification (Bishop, 2006; Berger, 2013). This is a precise classifier, in the sense that it always outputs one single category. For some problems, when there is much ambiguity about the category that fits data the best, it is sensible though to add a reject option to Bayesian classification, where no classification is made (Chow, 1970; Hellman, 1970; Zaffalon, 2002; Freund et al., 2004; Herbei & Wegkamp, 2006; Ripley, 2007), and if such a reject option is followed by additional data collection, this leads to Bayesian sequential analysis (Arrow et al., 1949). These types of classifiers have been generalized and formulated in the context of Bayesian decision theory, where the object is to maximize the posterior expected reward, using reward functions with a set-valued argument (Ha, 1997; del Coz et al., 2009; Zaffalon et al., 2012; Mortier et al., 2021). Each such reward function leads to a classifier that outputs a subset of categories, also referred to as a set-valued classifier (Grycko, 1993), a credal classifier (Zaffalon et al., 2012) a nondeterministic classifier (del Coz et al., 2009) or a partial classifier (Ma & Denoeux, 2021). A number of consistency properties for reward functions with a set-valued input argument have been defined by Yang et al. (2017).

A second frequentistic approach is to regard set-valued classifiers as generalizations of confidence intervals. This includes conformal prediction (Vovk et al., 2005; Shafer & Vovk, 2008; Vovk et al., 2017; Angelopoulos & Bates, 2022), classifiers that guarantee user-specified levels of coverage while minimizing the expected size of the classified set (Sadinle et al., 2019) and classifiers that conversely maximize the expected coverage subject to an upper bound on the expected size of the classified set (Denis & Hebiri, 2017). A third approach is to make use of Dempster-Shafer's theory of belief functions in order to define set-valued classifiers (Ma & Denoeux, 2021).

Regardless of which of these three approaches is used to define the set-valued classifier, the size of the classified set of categories is 0 if the classifier rejects all hypotheses, 1 for a firm decision that singles out one category, between 1 and N for a partial reject option, where ambiguity remains between some but not all categories, and N for a rejection to classify, i.e, when none of the categories are singled out. In any case, the classified set is determinate in the sense that the classifier only outputs one single set of size $0, 1, \ldots, N$. Indeterminate classifiers, with several possible classified sets, are obtained within a Bayesian framework when the posterior distribution is not known exactly but rather belongs to a convex credal set of distributions (Levi, 1983; Walley, 1991; Zaffalon, 2002; Yang et al., 2017).

In a previous article (Karlsson & Hössjer, 2023) we introduced a novel type of outlier detection and automatic choice of tuning parameters for Bayesian set-valued classification between a homogeneous set of categories. Much of our focus in Karlsson and Hössjer (2023) was devoted to computing the posterior probabilities of categories, for data that involve obfuscation and mixtures of different types of measurements. Here we present a theoretical follow-up paper were Bayesian determinate classifiers are studied more generally. In contrast to our previous article, we mostly assume that the posterior probabilities of the categories are given, and we pay less attention to the model for how data were generated. As in the previous paper, we use a decision-theoretic approach, based on reward functions with a set-valued argument. But in this article, we consider a much larger class of reward functions and unify much of previous work on set-valued classification, with general and explicit formulas for the Bayes classifiers that maximize expected rewards. We start by considering a homogeneous collection of categories of the same type and introduce reward functions with an additive or multiplicative penalty for large classified sets. In this context, we find explicit formulas for the Bayes classifier in terms of the ordered a posteriori probabilities of all hypotheses and the penalty terms for the sizes of the classified sets. It is shown that certain instances of conformal prediction, with posterior probabilities used as nonconformity measure, as

well as the minimum coverage approach of Sadinle et al. (2019), are related to a Bayesian classifier whose reward function has an additive and linear penalty for large sets. This close connection between Bayesian and frequentistic set-valued classifiers, when the penalty is a linear function of set size, has also been noted in the review article by Chzhen et al. (2021).

Then, we consider a setting where the categories to choose between are divided into a number of blocks, such that categories (typically) are more similar within than between blocks. It is shown that for reward functions adapted to blocks of categories, the associated Bayes classifiers involve the ordered a posteriori probabilities within each block. In particular, we study reward functions where either the reward of including the correct class in the classified set, or the penalty for large classified sets, is block dependent. In the latter case penalties a and b are incurred when categories from the correct and wrong blocks are classified respectively. We also demonstrate that classification with indifference zones (Bechhofer, 1954; Goldsman, 1986) can be represented as an instance of set-valued classification with two blocks of categories, and that hierarchical classification (Mortier et al., 2022), with three hierarchical levels, is an instance block-based classification as well.

Our framework of set-valued classification, with the set of categories partitioned into blocks, is applied to an ornithological data set. Each observation in data consists of three observed traits for a bird and which taxon the bird belongs to. Based on these data we want to train a classifier of future birds to taxon. The taxa are partitioned into blocks with regard to cross-taxon similarities. The parameters of the underlying statistical model are estimated through a Markov Chain Monte Carlo procedure, and the tuning parameters of the set-valued reward function are estimated through cross-validation.

The article is organized as follows: In Section 2 we introduce the statistical model with N hypotheses, and define the optimal (Bayes) set-valued classifier, with a motivating example in Section 3. Then, we introduce a large class of reward functions for models with one block of categories (Section 4) and several blocks of categories (Section 5) respectively, and give explicit expressions for the corresponding Bayes classifiers. The ornithological data set is analyzed in Section 6, and a discussion in Section 7 concludes the paper.

2 Statistical Model and Optimal Classifiers

Consider a random variable $Z \in \mathcal{Z}$, whose distribution follows one of N possible hypotheses (or categories)

$$H_i : Z \sim f_i, \quad i = 1, \ldots, N, \tag{1}$$

where f_i is the density or probability function of Z under H_i. We will assume a Bayesian framework, and thus the true but unknown hypothesis $I \in \mathcal{N} = \{1, \ldots, N\}$ is a random variable. It is assigned a categorical prior distribution with parameter vector $\pi_i = P(I = i)$, for $i = 1, \ldots, N$, and the corresponding posterior distribution of I, given an observed value z of Z, is

$$p_i = p_i(z) = \mathbb{P}(I = i | Z = z) = \frac{\pi_i f_i(z)}{\sum_{j=1}^{N} \pi_j f_j(z)}. \tag{2}$$

Our objective is to classify z. To this end, a classifier $\hat{I} = \hat{I}(z) \subset \mathcal{N}$ with partial reject options is defined as a subset of categories. We will assume that $\hat{I}$ is a deterministic function of z, and thus we exclude classifiers that involve a randomization procedure on top of data Z. For instance, the random classifier $P(\hat{I} = \{i\} | Z = z) = 1/N$, which randomly and uniformly assigns a category, regardless of data (Zaffalon et al., 2012) is not included in our framework.

When $\hat{I} = \{i\}$, a firm decision of category i is made, also referred to as a determinate prediction (Yang et al., 2017). A precise classifier is one for which

$$\mathbb{P}(|\hat{I}| = 1) = 1, \tag{3}$$

so that a firm decision is made with probability 1 with respect to variations in Z. The case $\hat{I} = \mathcal{N}$ corresponds to a reject option, where no decision is made about which categories that conform with z the most. Note that it is the action "to classify" that is rejected when $\hat{I} = \mathcal{N}$. In particular, a vacuous classifier is one that rejects to classify with probability 1 with respect to variations in Z (Zaffalon et al., 2012). The intermediate case $2 \leq |\hat{I}| = m \leq N - 1$ corresponds to a partial reject option, where the classifier excludes $N - m$ hypotheses but rejects discrimination among the remaining $m = m(z)$ categories. Another possibility, $\hat{I} = \emptyset$, corresponds to a scenario where none of the N categories fit observed data z well, and all hypotheses are excluded. This can be regarded as a safeguard against outliers or otherwise faulty data (Ripley, 2007; Karlsson & Hössjer, 2023).

Let

$$p_{(1)} = p_{(1)}(z) \leq \cdots \leq p_{(N)} = p_{(N)}(z) \tag{4}$$

refer to the ordered posterior category probabilities. For simplicity, we assume that the vector

$$p(Z) = (p_1(Z), \ldots, p_N(Z)) \tag{5}$$

has an absolutely continuous distribution with respect to Lebesgue measure on the standard $(N-1)$-dimensional simplex, so that ties in Eq. 4 occur with probability 0 and hence can be ignored.

The well-known Maximum A Posteriori (MAP) classifier (Berger, 2013)

$$\hat{I} = \hat{I}(z) = \{(N)\} \tag{6}$$

always makes a firm decision, according to Eq. 3. It also maximizes the probability $\mathbb{P}(\hat{I} = \{i\})$ of a correct classification, but has a higher probability of misclassification than correct classification, when $p_{(N)} < 1/2$. The MAP classifier can be formulated as the classifier $\hat{I}$ that maximizes the expected value

$$\bar{V} = \mathbb{E}[R(\hat{I}, I)] \tag{7}$$

of the reward function

$$R(\mathcal{I}, i) = \begin{cases} 1, & \mathcal{I} = \{i\}, \\ 0, & \mathcal{I} \neq \{i\} \end{cases} \tag{8}$$

assigned to a classified subset $\mathcal{I} \subset \mathcal{N}$ of categories when the true category is i. Since $\hat{I} = \hat{I}(Z)$, the expectation in Eq. 7 is with respect to Z and I. Let the value function

$$V(z; \mathcal{I}) = \mathbb{E}\left[R(\mathcal{I}, I) \mid Z = z\right] \tag{9}$$

refer to the conditional expected reward given $Z = z$. It is clear that the optimal Bayes classifier, that maximizes the expected reward, or equivalently maximizes the expected value function $\bar{V} = \mathbb{E}[V(Z; \hat{I}(Z))]$ is obtained as

$$\begin{aligned} \hat{I}(z) &= \arg\max_{\mathcal{I} \subset \mathcal{N}} V(z; \mathcal{I}) \\ &= \arg\max_{\mathcal{I} \subset \mathcal{N}} \mathbb{E}\left[R(\mathcal{I}, I) \mid Z = z\right] \\ &= \arg\max_{\mathcal{I} \subset \mathcal{N}} \sum_{i=1}^{N} R(\mathcal{I}, i)\, p_i(z) \end{aligned} \tag{10}$$

for each $z \in \mathcal{Z}$. Since the reward function Eq. 8 only takes non-zero values for singleton choices of $\mathcal{I} \subset \mathcal{N}$, we only need to consider $\mathcal{I} = \{i\}$ for $i = 1, \ldots, N$, in order to find $\hat{I}$. For each choice $\mathcal{I} = \{i\}$, $V(z; \{i\}) = p_i$. Thus, we choose $i = (N)$, since $p_{(N)}$ for each fixed z is the largest value $V(z, \mathcal{I})$ can return. This shows that the MAP classifier Eq. 6 is the optimal classifier under Eq. 8.

In this paper, we will consider optimal classifiers Eq. 10 for reward functions other than Eq. 8; those that allow not only for single classified categories but also for outliers and (partial) reject options.

3 Motivating Example

Before proceeding with the general theory of set-valued classification, in this section, we will first provide a motivating example. Suppose the participants of a Scandinavian quiz program are asked to classify houses to one of the following $N = 6$ cities,

$$\begin{aligned} 1 &= \text{Stockholm}, \\ 2 &= \text{Gothenburg}, \\ 3 &= \text{Malmö}, \\ 4 &= \text{Oslo}, \\ 5 &= \text{Trondheim}, \\ 6 &= \text{Bergen}, \end{aligned}$$

that naturally divide into two blocks of Swedish (1–3) and Norwegian (4–6) cities. The contestants are exposed to the same data z (a picture of a particular house), but they possibly have different priors π_i as well as different likelihoods f_i for how data was generated. The posterior probabilities Eq. 4 are therefore individual specific. Suppose one contestant comes up with the following posterior probabilities:

Country	Posterior probabilities			Total
Sweden	$p_1 = 0.23$	$p_2 = 0.10$	$p_3 = 0.07$	$P_1 = 0.40$
Norway	$p_4 = 0.22$	$p_5 = 0.20$	$p_6 = 0.18$	$P_2 = 0.60$

These posterior probabilities correspond to an ordering

$$\begin{aligned} (6) &= 1 = \text{Stockholm}, \\ (5) &= 4 = \text{Oslo}, \\ (4) &= 5 = \text{Trondheim}, \\ (3) &= 6 = \text{Bergen}, \\ (2) &= 2 = \text{Gothenburg}, \\ (1) &= 1 = \text{Malmö}, \end{aligned}$$

so that the MAP-classifier $\hat{I} = \{1\}$ outputs the most likely city Stockholm. On the other hand, the three Norwegian cities have posterior probabilities quite close to that of Stockholm, and the overall posterior probability for the three Norwegian cities is larger (0.6) than for the Swedish ones (0.4). Which type of classifier is more preferable? To answer this question we must define how severe it is to output a large set of cities, and also whether cities from the wrong country should incur a smaller reward than cities from the correct country. We will

consider three reward functions I-III that address these questions differently. All of them add a reward of 1 to sets $\mathcal{I} \subset \{1, \ldots, 6\}$ that includes the true city i, but then they add the following types of penalties (which are subtracted from the reward function):

I : Add a penalty $c \geq 0$ for each *extra* city (on top of the first one),
II : Add a penalty $c \geq 0$ for each *extra* city from the correct country and the same penalty $c \geq 0$ for *all* cities from the wrong country,
III : Add *no* penalty for *any* of the cities from the correct country but add a penalty $c \geq 0$ for *all* cities from the wrong country.

The Bayes classifier Eq. 10 for the three reward functions I-III are displayed in the following table, for various choices of c:

Reward function I		Reward function II		Reward function III	
c	$\hat{I}$	c	$\hat{I}$	c	$\hat{I}$
$(0.220, \infty)$	$\{1\}$	$(0.550, \infty)$	$\emptyset$	$(0.550, \infty)$	$\emptyset$
$(0.200, 0.220]$	$\{1, 4\}$	$(0.383, 0.550]$	$\{4\}$	$(0.500, 0.550]$	$\{4\}$
$(0.180, 0.200]$	$\{1, 4, 5\}$	$(0.200, 0.383]$	$\{1, 4\}$	$(0.450, 0.500]$	$\{4, 5\}$
$(0.100, 0.180]$	$\{1, 4, 5, 6\}$	$(0.180, 0.200]$	$\{1, 4, 5\}$	$(0.383, 0.450]$	$\{4, 5, 6\}$
$(0.070, 0.010]$	$\{1, 2, 4, 5, 6\}$	$(0.100, 0.180]$	$\{1, 4, 5, 6\}$	$(0.167, 0.383]$	$\{1, 4, 5, 6\}$
$[0.000, 0.070]$	$\{1, 2, 3, 4, 5, 6\}$	$(0.070, 0.010]$	$\{1, 2, 4, 5, 6\}$	$(0.117, 0.167]$	$\{1, 2, 4, 5, 6\}$
		$[0.000, 0.070]$	$\{1, 2, 3, 4, 5, 6\}$	$[0.000, 0.117]$	$\{1, 2, 3, 4, 5, 6\}$

For all three classifiers, the size of the classified set is a non-decreasing function of the cost parameter c, with all six cities included for $c = 0$. The classifier $\hat{I}$ based on reward function I always includes Stockholm, regardless of c, since it is only extra cities, on top of the first one, that is penalized. On the other hand, the classifiers for reward functions II and III output the empty set for large c. Since the correct country is not known, and all cities from the wrong country are penalized by an amount c for II and III, the best option is to discard all cities for large enough c. Notice also that the classifier based on III tends to include more Norwegian cities than the one based on II, since Norway is the most likely country a posteriori, and only cities from the wrong country are penalized by III.

In the next two sections, we will derive Bayes classifiers for a more general class of reward functions. In Section 4 we consider reward functions that treat all categories homogeneously, with reward function I treated in Example 1. Then, in Section 5 we study reward functions that divide all hypotheses into disjoint blocks, with reward functions II and III treated in Example 7. Finally, in Section 6 we discuss data-driven choices of cost parameters for a class of block-based reward functions that include II and III as special cases.

4 Partial Rejection, One Block of Categories

When $\mathcal{N}$ represents one homogeneous block of categories, the following class of reward functions is natural to use:

Definition 1 *Given a set of categories $\mathcal{N} = \{1, \ldots, N\}$, with $i \in \mathcal{N}$ the true category of an observation and $\mathcal{I} \subset \mathcal{N}$ the classified subset of categories, a reward function*

$$R(\mathcal{I}, i) = R(\tau(\mathcal{I}), \tau(i)) \tag{11}$$

invariant w.r.t. permutations $\tau : \mathcal{N} \to \mathcal{N}$ *of the labels is called an* invariant reward function.

In this section, we will find the optimal Bayes classifier for various invariant reward functions. As we will see, these Bayes classifiers will formulated in terms of the following classifiers:

Definition 2 *A classifier containing the* $m \in \{0, \ldots, N\}$ *categories with the largest posterior probabilities, where*

$$\hat{I}_m = \begin{cases} \{(N+1-m), \ldots, (N)\}, & m \geq 1, \\ \emptyset, & m = 0, \end{cases} \tag{12}$$

is called an m most probable classifier*, abbreviated as an* m*-MP classifier.*

The m-MP classifier is referred to as the pointwise top-m classifier in Chzhen et al. (2021). Obviously, the 1-MP classifier is the MAP classifier, whereas the 0-MP classifier corresponds to the empty set since in this case none of the categories is chosen. Thus, the MAP classifier is a special case of the more general class of m-MP classifiers.

4.1 Additive Penalties for the Sizes of Classified Sets

It is easily seen that

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - g(|\mathcal{I}|) \tag{13}$$

is an invariant reward function. The first term of Eq. 13 corresponds to a reward of 1 for a classified set $\mathcal{I}$ that includes the true category i, whereas the second term is an additive penalty $g(|\mathcal{I}|) \geq 0$ for the size $|\mathcal{I}|$ of the classified set. Proposition 1 links Bayes classifiers of invariant reward functions with additive penalties of set size, to m-MP classifiers with $m = m(z)$ depending on data z.

Proposition 1 *The optimal classifier for an invariant reward function of type Eq. 13, with an additive penalty term* $g(|\mathcal{I}|)$ *for the size of a classified set, is an* m*-MP classifier* $\hat{I}(z) = \hat{I}_{m(z)}$ *with*

$$m(z) = \arg\max_{0 \leq m \leq N} \left[v(m; z) - g(m) \right], \tag{14}$$

where $v(0; z) = v(0) = 0$ *and*

$$v(m; z) = v(m) = \sum_{j=1}^{m} p_{(N+1-j)}, \quad m = 1, \ldots, N. \tag{15}$$

Proof Recall that the optimal classifier $\hat{I}(z)$ maximizes, for each $z \in \mathcal{Z}$, the value function $V(z; \mathcal{I})$ among all nonempty subsets $\mathcal{I}$ of $\mathcal{N}$. For the reward function of Eq. 13 we have that

$$\begin{aligned} V(z; \mathcal{I}) &= \mathbb{E}\left[R(\mathcal{I}, I) \mid Z = z\right] \\ &= \textstyle\sum_{i=1}^{N} \left(1(i \in \mathcal{I}) - g(|\mathcal{I}|)\right) p_i(z) \\ &= \textstyle\sum_{i=1}^{N} 1(i \in \mathcal{I}) p_i(z) - \sum_{i=1}^{N} g(|\mathcal{I}|) p_i(z) \\ &= \textstyle\sum_{i \in \mathcal{I}} p_i - g(|\mathcal{I}|) \\ &\leq \textstyle\sum_{j=1}^{|\mathcal{I}|} p_{(N+1-j)} - g(|\mathcal{I}|) \\ &= v(|\mathcal{I}|; z) - g(|\mathcal{I}|). \end{aligned} \tag{16}$$

Note that the inequality occurs since we go from considering a subset $\mathcal{I} \subseteq \mathcal{N}$ of categories to considering another subset with equally many but the most probable categories. Consequently,

among all $\mathcal{I} \subset \mathcal{N}$ of size $|\mathcal{I}| = m$, the value function $V(z; \mathcal{I})$ is maximized by $\hat{I}_m$ in Eq. 12, for some $m \in \{0, \ldots, N\}$. Among these subsets, the optimal classifier is $\hat{I} = \hat{I}_{m(z)}$, where $m(z)$ is the value of $|\mathcal{I}|$ that maximizes the right-hand side of Eq. 16. Since this value of $m(z)$ is identical to the one in Eq. 14, this finishes the proof. □

Proposition 1 provides an efficient algorithm for computing the Bayesian classifier Eq. 10 for reward functions Eq. 13. Indeed, in order to find the optimal classifier $\hat{I}_{m(z)}$, it follows from the proof of Proposition 1 that after having sorted all posterior probabilities $\{p_i(z)\}_{i=1}^N\}$ (which requires $O(N \log(N))$ operations), we only need to compute the value function $V(z; \mathcal{I})$ in Eq. 10 for the $N + 1$ sets $\mathcal{I} \in \{\hat{I}_0, \ldots, \hat{I}_N\}$ rather than for all 2^N subsets of $\mathcal{N}$. The next result is a corollary of Proposition 1, and it treats an important class of reward functions with a convex additive penalty function:

Corollary 1 *Consider a classifier $\hat{I}(z)$ based on a reward function Eq. 13, for which the additive penalty term $g(m)$ is a convex function of m, with $g(0) = 0$. Then, Eq. 14 simplifies to $\hat{I}(z) = \hat{I}_{m(z)}$, with*

$$m(z) = \max\{m'(z), m''(z)\}, \tag{17}$$

where

$$\begin{aligned} m'(z) &= 0, \\ m''(z) &= \max\{1 \le m \le N;\ p_{(N+1-m)} \ge g(m) - g(m-1)\}, \end{aligned} \tag{18}$$

and $\max \emptyset = -\infty$ in the definition of $m''(z)$.

Proof In order to prove Eq. 18, we first deduce from Eq. 4 that $v(m; z) = v(m)$ is a concave function of m (indeed, the differences $\Delta v(m+1) = v(m+1) - v(m)$ are decreasing as m increases and thus $v(m)$ is concave). Consequently, if $g(m)$ is a convex function of m it follows that $v(m) - g(m)$ is a concave function of m, and it is therefore maximized by

$$m(z) = \max\left\{0, \max_{1 \le m \le N} \{\, v(m) - g(m) \ge v(m-1) - g(m-1)\}\right\}, \tag{19}$$

where $\max \emptyset = -\infty$ in the inner maximization. By the definition of $v(m)$ in Eqs. 15, 19 is equivalent to the expression for $m(z)$ given in Eq. 18. □

Remark 1 *Note that in the second line of Eq. 18, the inequality constitutes an inclusion criterion that the posterior probability of category $(N+1-m)$ needs to fulfill to be included in the classifier. As can be seen from the left and right-hand sides of the inequality, the category with the largest posterior probability $p_{(N+1-m)}$ not yet included, will be included if this posterior probability is larger than the added penalty $g(m) - g(m-1)$ associated with enlarging the size of the classifier from $m-1$ to m. Since $p_{(N+1-m)}$ is a non-increasing function of m, whereas $g(m) - g(m-1)$ is a non-decreasing function of m, it follows that only the sets $\mathcal{I} \in \{\hat{I}_0, \hat{I}_1, \ldots, \hat{I}_{\min(m(z)+1, N)}\}$ need to be considered. This is a computational improvement compared to the $N+1$ sets listed below Proposition 1.*

Example 1 (Proportion-based and linear reward functions.) In order to illustrate Corollary 1 we introduce two closely related reward functions. They both have an additive penalty term $g(m)$ that involves a cost parameter c per classified category, but they differ as to whether singleton sets are penalized or not:

Definition 3 *An invariant reward function of the form*

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - c \max(0, |\mathcal{I}| - 1), \tag{20}$$

with $c \geq 0$, is referred to as a proportion-based reward function*, whereas a reward function*

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - c|\mathcal{I}] \tag{21}$$

with an additive linear penalty term is named a linear reward function*.*

Notice that the proportion-based reward function corresponds to reward function I of Section 3. The framework of penalizing each included category by an amount $c > 0$ is also referred to as class-selective rejection (Ha, 1996) or class-selection (Le Capitaine, 2014). Note that a proportion-based reward function has an additive penalty term that is almost linear in $|\mathcal{I}|$. It is only the maximum operator that distinguishes Eq. 20 from the linear reward function Eq. 21. For a proportion-based reward function, we may interpret c as a cost per *extra* classified category, on top of the first one, whereas the linear reward function has a penalty c for the first classified category as well. For proportion-based and linear reward functions, Corollary 1 simplifies as follows:

Corollary 2 *A proportion-based reward function Eq. 20 gives rise to an m-MP classifier Eq. 12 with*

$$\begin{aligned} m(z) &= \max\{1, \max\{2 \leq m \leq N;\ p_{(N+1-m)} \geq c\}\} \\ &= \max(1, |\{i;\ p_i \geq c\}|) \end{aligned} \tag{22}$$

in Eq. 17, whereas a linear reward function Eq. 21 gives rise to an m-MP classifier with

$$\begin{aligned} m(z) &= \max\{0, \max\{1 \leq m \leq N;\ p_{(N+1-m)} \geq c\}\} \\ &= |\{i;\ p_i \geq c\}|. \end{aligned} \tag{23}$$

Proof Since a proportion-based reward function Eq. 20 has an additive penalty term $g(m) = c \max(0, m-1)$, it follows that

$$g(m) - g(m-1) = \begin{cases} 0, & m = 1, \\ c, & m = 2, \ldots, N. \end{cases} \tag{24}$$

Inserting Eq. 24 into Eqs. 17-18, Eq. 22 follows. The proof of Eq. 23 is analogous. □

The Bayes classifier $\hat{I}_{m(z)}$ of a proportion-based reward function is always non-empty, but this need not be the case for a linear reward function. Ha (1997) considers linear reward functions Eq. 21, but he restricts the first argument of R to the $2^N - 1$ nonempty subsets $\mathcal{I}$ of $\mathcal{N}$. In our context, where the first argument $\mathcal{I}$ of R ranges over all 2^N subsets of $\mathcal{N}$, this is equivalent to using a proportion-based reward function Eq. 20. Ha (1997) proved that the Bayesian classifier $\hat{I}(z) = \hat{I}_{m(z)}$ is given by Eq. 22. Note in particular that the MAP classifier Eq. 6 is obtained for a proportion-based reward function with $c > p_{(N)}$. For this reason, Karlsson and Hössjer (2023) restricted the cost parameter of Eq. 20 to a range $0 \leq c \leq p_{(N)}$ and reparametrized it as

$$c = \rho p_{(N)}, \tag{25}$$

with $0 \leq \rho \leq 1$. The m-MP classifier $\hat{I}_{m(z)}$, with $m(z)$ as in Eq. 22, then takes the form

$$\hat{I} = \{i;\ p_i \geq c\} = \{i;\ p_i \geq \rho p_{(N)}\}. \tag{26}$$

On the other hand, choosing a linear penalty term $g(|\mathcal{I}|) = c|\mathcal{I}|$, as in Eq. 21, and then applying Corollary 2, we find that Eq. 26 is the Bayesian classifier for *any* value $c \geq 0$ of the cost parameter, whereas Eq. 26 is optimal for a proportion-based reward function, only when $0 \leq c \leq p_{(N)}$ (or equivalently, for $0 \leq \rho \leq 1$). In Chzhen et al. (2021), Eq. 26 is referred to as a classifier based on thresholding.

Example 2 (Frequentist classifiers and linear reward functions) It turns out several frequentist methods of set-valued classification are given by Eq. 26, the optimal Bayes classifier for a linear reward function Eq. 21. The associated cost parameter $c = c_n$ depends on the method used, and possibly also on the size n of training data.

Conformal prediction (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2022) is a general method for creating a prediction region $\Gamma^\delta = \Gamma^\delta(z)$ for an observation z with a marginal coverage $(1-\delta)\%$, where $\delta \in (0, 1)$ is chosen freely, typically close to 0. For the special case of classification or categorical prediction, the conformal algorithm (Shafer & Vovk, 2008, Section 4.3) uses as input the new observation z that we want to classify, a labeled training (or calibration) data set of independent and identically distributed (i.i.d.) observations $\mathcal{D} = \{(z_j, i_j)\}_{j=1}^n$ of size n, drawn from the distribution

$$P[(Z, I) = (z, i)] = \pi_i f_i(z). \tag{27}$$

In order to define Γ^δ we need a nonconformity score $A = A(z, i)$, which is larger the more consistent z is with category i. Then, for each possible label $i \in \mathcal{N}$ of z a decision is made as to whether i should be included in the prediction region Γ^δ or not, based on how large the non-conformity score of the new observation is compared to test data.

It turns out that the prediction region Γ^δ is closely related to a Bayesian classifier with an additive and linear reward function Eq. 21, when posterior probabilities $A(z, i) = p_i(z)$ are used as nonconformity measure. We can apply the theory of Shafer and Vovk (2008) and Angelopoulos and Bates (2022) in order to demonstrate this. To this end, let z be the observation we want to classify. For each possible label (or category) $i \in \mathcal{N}$ of z, provisionally set (z, i) as a future observation that is part of training data (although in practice z is rather a future observation that we want to classify). That is, under these assumptions we have

$$\Gamma^\delta(z) = \{i; 1 \le i \le N, p_i(z) \ge c_n(\delta)\}, \tag{28}$$

where $c_n(\delta) = \hat{F}_n^{-1}(1-\delta)$ is the $(1-\delta)$-quantile of the empirical distribution function

$$\hat{F}_n(p) = \frac{1}{n}\sum_{j=1}^n 1(A(z_j, i_j) \le p) = \frac{1}{n}\sum_{j=1}^n 1(p_{i_j}(z_j) \le p)$$

of the nonconformity scores (the posterior probabilities) of test data. If F_i refers to the distribution function of $p_i(Z)$ when $Z \sim f_i$, it follows that each nonconformity score $p_I(Z)$ of test data is drawn from the mixture distribution

$$F = \sum_{i=1}^N \pi_i F_i. \tag{29}$$

Since $\mathcal{D}$ is an i.i.d. sample from Eq. 27, it follows that $\{p_{i_j}(z_j)\}_{j=1}^n$ is an i.i.d. sample from F. If the size n of test data tends to infinity, the Glivenko-Cantelli Theorem implies that $F_n \xrightarrow{\mathcal{L}} F$ converges weakly as $n \to \infty$. Consequently, the threshold of the prediction region Eq. 28 converges as

$$c_n \to c = F^{-1}(\delta) \tag{30}$$

when $n \to \infty$. Sadinle et al. (2019) found that Eq. 26, with a cost parameter $c = c(\delta)$ as in Eq. 30, is also the optimal classifier when minimizing the expected size $E(|\hat{I}(Z)|)$ of the classified set subject to a lower bound $1-\delta$ on the average coverage probability $P(I \in \hat{I}(Z))$.

It was proved in Denis and Hebiri (2017) that Eq. 26 is the optimal classifier when maximizing the expected coverage $P(I \in \hat{I})$, subject to an upper bound k on expected size $E(|\hat{I}|)$

on the size of the classified set. In more detail, they found that the cost parameter of Eq. 26 should be chosen as

$$c = F^{-1}(1 - k/N), \tag{31}$$

where F is the distribution function in Eq. 29, with a uniform prior $\pi_i = 1/N$ on all categories.

4.2 Multiplicative Penalties for the Sizes of Classified Sets

In this section, we consider reward functions

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I})g(|\mathcal{I}|) \tag{32}$$

with a multiplicative penalty term $g(|\mathcal{I}) \geq 0$ for the size $|\mathcal{I}|$ of a classified set $\mathcal{I}$. Mortier et al. (2021) provide general results for Bayesian set-valued classifiers based on a multiplicative reward function Eq. 32. Theorem 1 of their paper (cf. in particular (5) of its proof) is equivalent to the following result:

Proposition 2 *The optimal classifier for an invariant reward function of type Eq. 32, with a multiplicative penalty term $g(|\mathcal{I}|)$ for the size of a classified set, is an m-MP classifier $\hat{I}(z) = \hat{I}_{m(z)}$ with*

$$\begin{aligned} m(z) &= \arg\max_{0 \leq m \leq N} [v(m; z)g(m)] \\ &= \arg\max_{1 \leq m \leq N} [v(m; z)g(m)], \end{aligned} \tag{33}$$

where $v(m; z)$ is defined in Proposition 1.

Proof The first equality of Eq. 33 is proved in the same way as Proposition 1. The second equality follows from the fact that $v(0; z) = v(0; z)g(0) = 0$ regardless of the value of $g(0)$, whereas $v(m; z)g(m) \geq 0$ for $m = 1, \ldots, N$. □

Proposition 2 provides a computationally efficient algorithm for finding the Bayesian classifier $\hat{I}_{m(z)}$, in that the value function $V(z; \mathcal{I})$ in Eq. 10 only need to be computed for the N sets $\mathcal{I} \in \{\hat{I}_1, \ldots, \hat{I}_N\}$ rather than for all 2^N subsets of $\mathcal{N}$.

Remark 2 *Proposition 2 implies that the Bayesian classifier $\hat{I}_{m(z)}$ of a multiplicative reward function Eq. 32 is non-empty with probability 1, regardless of the value of $g(0)$. This is not surprising, since $R(\emptyset, i) = 0$ for such a reward function, whatever the value of $g(0)$ is. Consequently, there is no need to specify $g(0)$ for a multiplicative reward function.*

Example 3 (Classification with reject options.) Ripley (2007) introduced a reward function with a reject option. In our notation, it corresponds to

$$R(\mathcal{I}, i) = \begin{cases} 1(i = j), & \mathcal{I} = \{j\}, \\ 0, & |\mathcal{I}| \in \{0, 2, 3, \ldots, N - 1\}, \\ r, & \mathcal{I} = \mathcal{N}, \end{cases} \tag{34}$$

with a reward of $1/N < r < 1$ assigned to the reject option. The special case of Eq. 34 for $N = 2$ categories was treated by Chow (1970); Herbei and Wegkamp (2006). Note that Eq. 34 is equivalent to a reward function Eq. 13, with a multiplicative penalty term

$$g(m) = \begin{cases} 1, & m = 1, \\ 0, & m = 2, \ldots, N - 1, \\ r, & m = N. \end{cases} \tag{35}$$

The corresponding classifier

$$\hat{I} = \begin{cases} (N), & p_{(N)} > r, \\ \mathcal{N}, & p_{(N)} \le r \end{cases} \tag{36}$$

is deduced from Proposition 2. This classifier is of interest for Bayesian sequential analysis (Arrow et al., 1949; Berger, 2013) and sequential clinical trials (Carlin et al., 1998), where the reject option $\hat{I} = \mathcal{N}$ corresponds to delaying the decision and collecting more data before the sequential procedure is stopped and a specific category is chosen.

Classification with a reject option can also be viewed as the Bayes classifier for a reward function Eq. 13 with an additive penalty

$$g_{\text{add}} = \begin{cases} 0, & m = 0, 1, \\ 1, & m = 2, \ldots, N-1, \\ 1-r, & m = N. \end{cases}$$

Indeed, it follows from Proposition 1 that the optimal classifier for such an additive reward function is Eq. 36.

Example 4 (Classifiers with a fixed or maximal size.) It has been suggested to fix the size of the classified set in advance (Russakovsky et al, 2015). Choosing $m_0 \in \{1, \ldots, N\}$ as the apriori size of the classified set corresponds to using a reward function Eq. 32 with a multiplicative penalty function

$$g(m) = \begin{cases} 1, & m = m_0, \\ 0, & m \ne m_0. \end{cases} \tag{37}$$

It follows from Eq. 33 that $m(z) = m_0$, so that the m_0-MP classifier $\hat{I}_{m_0}$ is optimal for such a reward function. In particular, the reward function Eq. 8, which only gives a positive reward to firm decisions ($|\hat{I}| = 1$), corresponds to using a multiplicative penalty Eq. 37 with $m_0 = 1$.

Note that the pointwise size strategy of Chzhen et al. (2021), where the probability $P(i \in \hat{I})$ of not committing a classification error, is maximized subject to a constraint $|\hat{I}| \le m_0$, corresponds to using a multiplicative penalty

$$g(m) = \begin{cases} 1, & m = 1, \ldots, m_0, \\ 0, & m = m_0 + 1, \ldots, N, \end{cases} \tag{38}$$

and it also leads to a Bayesian classifier $\hat{I}_{m_0}$.

The formula for the Bayes classifier simplifies when the argmax of $v(m; z)g(m)$ is a connected subset of $\{1, \ldots, N\}$. This is summarized in the following corollary of Proposition 2:

Corollary 3 *Consider a reward function Eq. 32, with a multiplicative penalty function g, for which*

$$\begin{aligned} w(m) &= w(m; z) \\ &= v(m; z)g(m) \text{ is maximized over a connected} \\ &\quad \text{set with upper bound } m(z). \end{aligned} \tag{39}$$

The Bayes classifier is then given by $\hat{I}(z) = \hat{I}_{m(z)}$, *with*

$$\begin{aligned} m(z) = \max\{1, \max\{1 \le m \le N;\ & g(m)p_{(N+1-m)} \\ & \ge v(m-1; z)[g(m-1) - g(m)]\}\}. \end{aligned} \tag{40}$$

Proof For ease of notation, put $w(m) = w(m; z)$ and $v(m) = v(m; z)$. Since the argmax of $w(m) = v(m)g(m)$, according to Eq. 39, is a connected subset of $\{0, 1, \ldots, N\}$, Proposition 2 implies that the Bayes classifier for a multiplicative reward function Eq. 32 is an $m(z)$-MAP classifier $\hat{I}_{m(z)}$ with

$$\begin{aligned} m(z) &= \max\left\{1, \max_{1\le m\le N}\{\, w(m) \ge w(m-1)\}\right\} \\ &= \max\left\{1, \max_{1\le m\le N}\{\, v(m)g(m) \ge v(m-1)g(m-1)\}\right\}. \end{aligned} \tag{41}$$

But $v(m)g(m) \ge v(m-1)g(m-1)$ is equivalent to $g(m)p_{(N+1-m)} \ge v(m-1)[g(m-1) - g(m)]$, and therefore $m(z)$ is given by Eq. 40. □

Corollary 3 provides a computationally efficient algorithm for finding the optimal Bayesian classifier $\hat{I}_{m(z)}$ that is faster than the one in Proposition 2. In order to find $\hat{I}_{m(z)}$ it suffices to compute the value function $V(\mathcal{I}; z)$ in Eq. 10 for the $\min(m(z) + 1, N)$ sets $\mathcal{I} \in \{\hat{I}_1, \ldots, \hat{I}_{\min(m(z)+1,N)}\}$, rather than for all $2^N - 1$ non-empty subsets of $\mathcal{N}$.

Corollary 4 provides a sufficient condition for Corollary 3 to hold.

Corollary 4 *Consider a multiplicative reward function Eq. 32, with a penalty function $g(m) > 0$ such that $1/g(m)$ is a convex function of $m = 1, \ldots, N$. Then, Eq. 39 holds, so that the Bayes classifier $\hat{I}_{m(z)}$ is given by Eq. 40.*

Proof In order to prove Eq. 39, it suffices to verify that $w(m)$ is non-decreasing in m for $m = 1, \ldots, m(z)$ and strictly decreasing in m for $m = m(z) + 1, \ldots, N$. This is equivalent to showing that $\Delta w(m) = w(m) - w(m-1)$ satisfies

$$\Delta w(m) < 0 \Longrightarrow \Delta w(m+1) < 0, \quad m = 2, \ldots, N-1. \tag{42}$$

It is convenient to introduce the function $s(m) = 1/g(m)$. Since

$$\Delta w(m) = \frac{s(m-1)v(m) - s(m)v(m-1)}{s(m-1)s(m)},$$

it follows that $\Delta w(m) < 0$ is equivalent to

$$l(m) := \frac{v(m) - v(m-1)}{v(m-1)} < \frac{s(m) - s(m-1)}{s(m-1)} =: u(m).$$

Recall that v is increasing and concave, because of Eq. 4, whereas s is convex, by assumption. In addition, if $\Delta w(m) < 0$, then necessarily $s(m) > s(m-1)$ and because of the convexity of $s(\cdot)$, this function must be strictly increasing on $\{m-1, \ldots, N\}$. After some computations, it can be seen that the concavity of v and the convexity of s lead to

$$\begin{aligned} l(m+1) &\le l(m)/[1 + l(m)], \\ u(m+1) &\ge u(m)/[1 + u(m)]. \end{aligned}$$

Since $x \to x/(1+x)$ is strictly increasing it follows that $l(m) < u(m)$ implies $l(m+1) < u(m+1)$, which is equivalent to Eq. 42. □

Remark 3 *Corollary 4 is a generalization of Theorem 2 of Mortier et al. (2021), where the authors prove that the Bayes classifier $\hat{I}_{m(z)}$ is given by Eq. 40, whenever $1/g(m)$ is convex and non-decreasing. In Corollary 4 we dropped the assumption that $1/g(m)$ is non-decreasing.*

We will now present examples of reward functions, with a multiplicative reward function, for which Corollary 4 applies:

Example 5 (Reward functions with a multiplicative, rational penalty.) Suppose the reward function Eq. 32 has a multiplicative penalty that is a rational function

$$g(m) = \frac{a}{b+m} \tag{43}$$

of $m = 1, \ldots, N$ for some constants $a > 0$ and $b \geq 0$. This type of reward function has been studied by del Coz et al. (2009). They viewed set-valued classifiers $\hat{I}(z)$ as a way of solving an information retrieval task for each input z. A multiplicative reward function with penalty Eq. 43 and $a = 1 + \beta^2$, $b = \beta^2$ was referred to as an F_β-metric. At one extreme, the case $\beta = 0$ ($a = 1, b = 0$) gives rise to

$$R(\mathcal{I}, i) = \begin{cases} 0, & \mathcal{I} = \emptyset, \\ 1(i \in \mathcal{I})/|\mathcal{I}|, & |\mathcal{I}| = 1, \ldots, N, \end{cases} \tag{44}$$

which was named a discounted accuracy reward by Zaffalon et al. (2012) and a precision reward by del Coz et al. (2009). It can be shown that the MAP classifier Eq. 6 is optimal for Eq. 44 (cf. Proposition 2 of Mortier et al. (2021)). At the other extreme, the limit $\beta^2 \to \infty$ of the F_β-metric corresponds to a reward function

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}), \tag{45}$$

referred to as recall reward by del Coz et al. (2009). It does not penalize the size of $\mathcal{I}$, and therefore it leads to an optimal vacuous classifier ($\mathcal{I}(z) = \mathcal{N}$, regardless of z). Note that the F_β-reward is a weighted harmonic average of the two reward functions Eqs. 44 and 45, and the larger β^2 is the less it penalizes large sets. del Coz et al. (2009) showed that the Bayes classifier $\hat{I}_{m(z)}$ of a multiplicative reward function with penalty Eq. 43 is given by Eq. 40. This result is a special case of Corollary 4, since $1/g(m) = (b+m)/a$ is a linear and hence a convex function of m. Zaffalon et al. (2012) studied a multiplicative reward function with a penalty

$$g(m) = \frac{a}{m} - \frac{b}{m^2} \tag{46}$$

for $m = 1, \ldots, N$ and some conveniently chosen parameters $a > b \geq 0$ (with particular focus on $a = 1.6$, $b = 0.6$). In order to show that Eq. 46 satisfies Corollary 4, we notice that

$$s(m) = \frac{1}{g(m)} = \frac{m^2}{a(m-c)},$$

with $0 \leq c = b/a < 1$. Viewing $m > c$ as a real-valued argument of s it can be seen that the second derivative

$$s''(m) = \frac{2c^2}{a(m-c)^3}$$

of s is strictly positive for all $m > c$. This implies that s is convex on $\{1, \ldots, m\}$, since the second-order difference of s satisfies

$$\begin{aligned} \Delta^2 s(m) &= s(m) - 2s(m-1) + s(m-2) \\ &= \int_{m-2}^{m} (1 - |x - m + 1|)_+ \, s''(x) dx > 0 \end{aligned} \tag{47}$$

for $m = 3, \ldots, N$, with $y_+ = \max(0, y)$ the positive part of y. Corollary 4 thus applies to the penalty function Eq. 46, in spite of the fact that $s(m)$ need not be increasing.

Corollary 5 provides another sufficient condition for Corollary 3 to hold.

Corollary 5 *Consider a multiplicative reward function Eq. 32, with a penalty function* $g(m) = h(1/m)$ *such that* h *is a concave increasing function on* $[0, 1]$ *with* $h(0) = 0$ *and* $h(1) > 0$. *Then, Eq. 39 holds, so that the Bayes classifier* $\hat{I}_{m(z)}$ *is given by Eq. 40.*

Proof Our proof of Corollary 5 parallels that of Corollary 4. Define $w(m) = v(m)g(m) = v(m)h(1/m)$ for $m = 1, \ldots, N$. As in the proof of Corollary 4 it suffices to establish

$$\Delta w(m) < 0 \Longrightarrow \Delta w(m+1) < 0 \tag{48}$$

for $m = 2, \ldots, N-1$, where

$$\Delta w(m) = w(m) - w(m-1) = \Delta v(m) h(\frac{1}{m}) + v(m-1)\left[h(\frac{1}{m-1}) - h(\frac{1}{m})\right].$$

The left hand side of Eq. 48, $\Delta w(m) < 0$, is equivalent to

$$l(m) := \frac{\Delta v(m)}{v(m-1)} < \frac{h(\frac{1}{m-1}) - h(\frac{1}{m})}{h(\frac{1}{m})} =: u(m).$$

It follows from Eq. 4 that $v(m)$ is a concave and non-decreasing function of m with $v(0) = 0$ and $v(m) > 0$ for $m > 0$. From this, it follows that

$$l(m+1) \leq \frac{l(m)}{1 + l(m)}. \tag{49}$$

Using the fact that $x \to x/(1+x)$ is a strictly increasing function of x, Eq. 48 will follow from Eq. 49 if we establish that

$$u(m+1) \geq \frac{u(m)}{1 + u(m)}. \tag{50}$$

Since h is increasing and concave, with $h(0) = 0$, we have that

$$u(m) \leq \frac{\frac{m+1}{m-1}[h(\frac{1}{m}) - h(\frac{1}{m+1})]}{h(\frac{1}{m+1}) + [h(\frac{1}{m}) - h(\frac{1}{m+1})]} = \frac{m+1}{m-1} \cdot \frac{u(m+1)}{1 + u(m+1)},$$

which is equivalent to

$$u(m+1) \geq \frac{\frac{m-1}{m+1} u(m)}{1 - \frac{m-1}{m+1} u(m)}. \tag{51}$$

Thus, Eq. 50 follows if we can prove that the right-hand side of Eq. 51 is at least as large as the right-hand side of Eq. 50. With $x = u(m)$ and $a = (m-1)/(m+1)$, this is equivalent to establishing that

$$\frac{a}{1 - ax} \geq \frac{x}{1 + x} \Longleftrightarrow x^2 + \frac{a-1}{a} x + 1 \geq 0.$$

But the last inequality follows from the fact that $-2 \leq (a-1)/a = -2/(m-1) < 0$ for $m = 2, \ldots, N-1$. This finishes the proof of Eq. 50, and hence of Eq. 48. □

Example 6 (Reward functions with a multiplicative, rational penalty, contd.) Zaffalon et al. (2012) have studied multiplicative reward functions, with a penalty function $g(m) = h(1/m)$ such that h is increasing and concave on $[0, 1]$ with $h(0) = 0$ and $h(1) > 0$. It follows from Corollary 5 that the Bayes classifier $\hat{I}_{m(z)}$ for such reward functions can be obtained as in Corollary 3. In particular, the penalty function Eq. 46 corresponds to choosing $h(x) = ax - bx^2$, and it satisfies the requirements of Corollary 5 if $a > b > 0$. In Example 5 we found that a reward function with such a penalty also satisfies the requirements of Corollary 4.

Proposition 3 shows that none of the conditions on the penalty function g, in Corollaries 4 and 5, are contained in one another.

Proposition 3 *It is possible to find a penalty function g, of a multiplicative reward function Eq. 32, that satisfies the conditions of Corollary 5 but not the conditions of Corollary 4, or the other way around.*

Proof Put $x = 1/m$, for real-valued arguments $m \in [1, \infty)$ and $x \in [0, 1]$. Assume further that the multiplicative reward function has a penalty function $g(m) = 1/s(m) = h(x)$.

Suppose h satisfies the conditions of Corollary 5. That is, h is non-decreasing and concave with $h(0) = 0$ and $h(x) > 0$ for $x \in (0, 1]$. Differentiating the relation $s(m) = 1/h(1/m)$ twice with respect to m we find that

$$s''(m) = \frac{h'(\frac{1}{m})}{h^2(\frac{1}{m})m^3}\left[\frac{2h'(\frac{1}{m})}{h(\frac{1}{m})m} - \frac{h''(\frac{1}{m})}{h'(\frac{1}{m})m} - 2\right]. \tag{52}$$

Consider in particular the function

$$h(x) = \begin{cases} 3x/2, & 0 \le x \le 1/3, \\ 1/2 + 3(x - 1/3)/4, & 1/3 \le x \le 1. \end{cases} \tag{53}$$

Since $h \ge 0$ is concave with $h(0) = 0$ it follows that $h(x) \ge xh'(x)$. In particular, for h in Eq. 53 we have strict inequality, i.e., $h(x) > xh'(x)$, for $1/3 < x \le 1$. Moreover, for this choice of h we also have $h''(x) = 0$ for $x > 1/3$ or $1 \le m < 3$. Insertion into Eq. 52 yields $s''(m) < 0$ for $1 \le m < 3$. It follows from Eq. 47 that $\Delta^2 s(3) < 0$. Consequently, s does not satisfy the conditions of Corollary 4.

Suppose conversely that s satisfies the conditions of Corollary 4. That is, $s(m) > 0$ is convex on $m = 1, \ldots, N$. We will additionally assume that $s'(m) > 0$ and $s''(m) \ge 0$ for the real-valued argument $m \in [1, \infty)$. Differentiating the relation $h(x) = 1/s(1/x)$ twice with respect to x, we find, in analogy with Eq. 52, that

$$h''(x) = \frac{s'(\frac{1}{x})}{s^2(\frac{1}{x})x^3}\left[\frac{2s'(\frac{1}{x})}{s(\frac{1}{x})x} - \frac{s''(\frac{1}{x})}{s'(\frac{1}{x})x} - 2\right]. \tag{54}$$

Putting $s(m) = e^{f(m)}$, we have that $s'(m) = f'(m)s(m)$ and $s''(m) = [f'(m)^2 + f''(m)]s(m)$. Inserting these relations into Eq. 54 we find that

$$h''(x) = \frac{s'(m)m^3}{s^2(m)}\left[m(f'(m) - \frac{f''(m)}{f'(m)}) - 2\right]. \tag{55}$$

For instance, with $s(m) = e^m$ and $f(m) = m$, formula Eq. 55 simplifies to

$$h''(x) = \frac{m^3}{s(m)}(m - 2). \tag{56}$$

This implies $h''(x) > 0$ for $0 < x < 1/2$, proving that h is not concave and hence does not satisfy the conditions of Corollary 5. □

5 Several Blocks of Categories

This section will cover an extension of the classical classification problem, where observations belong to a category and categories belong to supercategories, or blocks, as we will call them.

Thus, we partition the N categories into K blocks of sizes $N_1, \ldots, N_K$, with $\sum_{k=1}^{K} N_k = N$. Without loss of generality, the labels i are defined so that each block

$$\mathcal{N}_k = \left\{ i; \sum_{l=1}^{k-1} N_l + 1 \leq i \leq \sum_{l=1}^{k} N_l \right\} \tag{57}$$

consists of adjacent categories. The different scenarios when it is worse to misclassify within a block than between blocks, and vice versa, will be a subject of study later on in this section. In order to define classifiers that take this into account, we introduce a new type of reward functions.

Definition 4 *For a set of N categories, partitioned into blocks $\mathcal{N}_k$, $k = 1, \ldots, K$ with N_k categories in each block,* block invariant reward functions *satisfy Eq. 11 only for block preserving permutations $\tau(\mathcal{N}_k) = \mathcal{N}_k$, $k = 1, \ldots, K$.*

In this section, we will find Bayes classifiers of block invariant reward functions. To this end, it will be helpful to order the posterior probabilities $p_i, i \in \mathcal{N}_k$ within each block as

$$p_{(k1)} \leq \cdots \leq p_{(kN_k)}.$$

Definition 5 *For an integer vector*

$$\boldsymbol{m} = (m_1, \ldots, m_K) \in \bigotimes_{k=1}^{K} \{0, \ldots, N_k\} \tag{58}$$

let

$$\hat{I}_{\boldsymbol{m}} = \{(ki);\ 1 \leq k \leq K, N_k + 1 - m_k \leq i \leq N_k\} \tag{59}$$

be a classifier that includes the m_k categories with the largest posterior probabilities from block k. We call this an $\boldsymbol{m}$-MP classifier.

Note that the $\boldsymbol{m}$-MP classifier is a subset of $\mathcal{N}$ of size

$$m = \sum_{k=1}^{K} m_k. \tag{60}$$

In particular, the two extreme scenarios with no categories classified or a rejection to classify, correspond to $\hat{I}_{(0,\ldots,0)} = \emptyset$ and $\hat{I}_{(N_1,\ldots,N_K)} = \mathcal{N}$ respectively. Moreover, Eq. 60 reduces to an m_k-MP classifier Eq. 12 for a particular block k if $\boldsymbol{m}$ has only one non-zero element m_k.

5.1 Block-Dependent Rewards of Including the True Category

A class of block-invariant reward functions, with an additive penalty, is

$$R(\mathcal{I}, i) = d_{k(i)} 1(i \in \mathcal{I}) - g(|\mathcal{I}|), \tag{61}$$

with $k(i)$ the block to which i belongs, and d_k the reward of including category $i \in \mathcal{N}_k$ in $\mathcal{I}$, when i is true. The corresponding block invariant reward function, with multiplicative penalties, is

$$R(\mathcal{I}, i) = d_{k(i)} 1(i \in \mathcal{I}) g(|\mathcal{I}|). \tag{62}$$

Ha (1997) studied a reward function Eq. 61 with an additive proportion-based penalty term Eq. 20, in the special case when all categories belong to different blocks, i.e., $K = N$ and $\mathcal{N}_k = \{k\}$. Introduce the reward weighted posterior probabilities $q_i = d_{k(i)} p_i = q_i(z)$, and notice that the value function Eq. 9 simplifies to

$$V(z; \mathcal{I}) = \begin{cases} \sum_{i \in \mathcal{I}} q_i - g(|\mathcal{I}|), & \text{for reward function Eq.61,} \\ g(|\mathcal{I}|) \sum_{i \in \mathcal{I}} q_i, & \text{for reward function Eq.62.} \end{cases} \tag{63}$$

In conjunction with Eq. 10, this implies that the value functions, for rewards Eqs. 61 and 62, are obtained by replacing p_i with q_i for the value functions of Sects. 4.1 and 4.2 respectively. Consequently, all results of Sects. 4.1 and 4.2 remain valid for reward functions Eqs. 61 and 62 respectively, if $p_{(1)} \leq \cdots p_{(N)}$ are replaced $q_{[1]} \leq \cdots \leq q_{[N]}$, the ordered $\{q_i;\ i = 1, \ldots, N\}$. Therefore, the Bayes classifier of a reward function Eqs. 61 or 62, is of the form

$$\hat{I}_m^{(d)} = \{[N + m - 1], \ldots, [N]\} \tag{64}$$

for some $m = m(z)$, with subscript (d) indicating that this classifier involves the vector $d = (d_1, \ldots, d_K)$ of multiplicative rewards of including the true category in each block. It follows that Eq. 64 is the analog of the m-MP classifier Eq. 12, with p_i replaced by q_i. It is also an $\boldsymbol{m}$-MP classifier Eq. 59, since

$$\hat{I}_m^{(d)} = \hat{I}_{\boldsymbol{m}},$$

with $\boldsymbol{m} = \boldsymbol{m}(z) = (m_1, \ldots, m_N)$ defined by $m_k = |\{i \in \mathcal{N}_k; q_i \geq q_{[N-m+1]}\}|$ for $k = 1, \ldots, N$.

5.2 Additive Penalties of Sizes of Classified Sets

In this section, we consider block-invariant reward functions for which the reward of including the true category in $\mathcal{I}$ is the same regardless of the block to which this category belongs. On the other hand, the penalty for the size of classified sets is block-dependent. A class of such block invariant reward functions, with additive penalty, is

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - g_{k(i)}(|\mathcal{I}_{k(i)}|, |\mathcal{I}| - |\mathcal{I}_{k(i)}|), \tag{65}$$

where

$$\mathcal{I}_k = \mathcal{I} \cap \mathcal{N}_k \tag{66}$$

contains the categories of the classified set $\mathcal{I}$ that belong to block k, whereas $k(i)$ is the block to which i belongs, i.e., $i \in \mathcal{N}_{k(i)}$. Moreover, g_k is a penalty term for misclassification, when the true category i belongs to $\mathcal{N}_k$. This term is a function of the number of categories $|\mathcal{I}_{k(i)}|$ in the classified set $\mathcal{I}$ that belong to the correct block as well as the number of classified categories $|\mathcal{I}| - |\mathcal{I}_{k(i)}|$ that belong to any of the wrong blocks.

Proposition 4 links $\boldsymbol{m}$-MP-classifiers to Bayes classifiers of block invariant reward functions of type Eq. 65.

Proposition 4 *The optimal classifier, for an additive reward function Eq. 65, is an $\boldsymbol{m}(z)$-MP classifier $\hat{I}(z) = \hat{I}_{\boldsymbol{m}(z)}$ in Eq. 59, with*

$$\boldsymbol{m}(z) = \arg\max_{\boldsymbol{m}=(m_1,\ldots,m_K)} \sum_{k=1}^{K} \left[v_k(m_k; z) - P_k g_k(m_k, m - m_k) \right], \tag{67}$$

where

$$\begin{aligned} P_k = P_k(z) &= \sum_{i \in \mathcal{N}_k} p_i, \\ m &= \sum_{k=1}^{K} m_k, \\ v_k(m_k; z) &= \sum_{j=1}^{m_k} p_{(k,N_k+1-j)} \end{aligned} \tag{68}$$

when $m_k > 0$ *and* $v_k(0; z) = 0$ *for* $k = 1, \ldots, K$.

Proof The proof mimics that of Proposition 1. We start by finding the value function $V(z; \mathcal{I})$ in Eq. 10 for the block invariant reward function Eq. 65. It is given by

$$\begin{aligned} V(z; \mathcal{I}) &= \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} p_i - \sum_{k=1}^{K} P_k g_k(|\mathcal{I}_k|, |\mathcal{I}| - |\mathcal{I}_k|) \\ &\leq \sum_{k=1}^{K} \left[\sum_{j=1}^{|\mathcal{I}_k|} p_{(k,N_k+1-j)} - P_k g_k(|\mathcal{I}_k|, |\mathcal{I}| - |\mathcal{I}_k|) \right] \\ &= \sum_{k=1}^{K} \left[v_k(|\mathcal{I}_k|; z) - P_k g_k(|\mathcal{I}_k|, |\mathcal{I}| - |\mathcal{I}_k|) \right]. \end{aligned} \tag{69}$$

From this it follows that $V(z; \mathcal{I})$ is maximized, among all $\mathcal{I}$ with $|\mathcal{I}_k| = m_k$ for $k = 1, \ldots, K$, by $\hat{I}_{(m_1,\ldots,m_K)}$ in Eq. 59. The optimal classifier is therefore $\hat{I} = \hat{I}_{\boldsymbol{m}(z)}$, where $\boldsymbol{m}(z) = (m_1(z), \ldots, m_K(z))$ is the value of $(|\mathcal{I}_1|, \ldots, |\mathcal{I}_K|)$ that maximizes the right hand side of Eq. 69. Hence $\boldsymbol{m}(z)$ is given by Eq. 67. □

Example 7 (Composite proportion-based reward functions.) We will consider a class of reward functions that are special cases Eq. 65. These reward functions involve two cost parameters a and b:

Definition 6 *A block invariant reward function of the form*

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - a \max\left(|\mathcal{I}_{(i)}| - 1, 0\right) - b\left(|\mathcal{I}| - |\mathcal{I}_{(i)}|\right), \tag{70}$$

where $0 \leq a \leq b$ *are fixed constants, is called a* composite proportion-based reward function.

Notice that reward functions II and III of Section 3 correspond to a composite proportion-based reward function with $a = b = c$ and $a = 0$, $b = c$ respectively. It follows from Proposition 4 that composite proportion-based reward functions have very explicit optimal classifiers:

Corollary 6 *A composite proportion-based reward function (70), gives rise to an optimal classifier that is an* $\boldsymbol{m}(z)$*-MP classifier* $\hat{I}(z) = \hat{I}_{\boldsymbol{m}(z)}$ *in Eq. 59 with*

$$\begin{aligned} m_k(z) &= \arg\max_{0 \leq m_k \leq N_k} \left[v_k(m_k, z) - a P_k \max(m_k - 1, 0) - b(1 - P_k) m_k \right] \\ &= \max\{m'_k(z), m''_k(z)\}, \end{aligned} \tag{71}$$

for $k = 1, \ldots, K$, *where*

$$\begin{aligned} m'_k(z) &= 1\left(p_{(kN_k)} \geq (1 - P_k)b\right), \\ m''_k(z) &= \max\{2 \leq m_k \leq N_k;\ p_{(k,N_k+1-m_k)} \geq P_k a + (1 - P_k)b\}, \end{aligned} \tag{72}$$

and with $\max \emptyset = -\infty$ *used in the definition of* $m''_k(z)$.

Proof In view of Proposition 4, it suffices to prove that for a composite proportion-based reward function Eq. 70, the optimal classifier is of type $\hat{I}(z) = \hat{I}_{\boldsymbol{m}(z)}$, where $\boldsymbol{m}(z) = (m_1(z), \ldots, n_K(z))$ is given by Eq. 71. To this end, we first notice, from the right-hand side of Eq. 69, that the value function of a classified set $\hat{I}_{\boldsymbol{m}}$ is

$$\begin{aligned} V(z; \hat{I}_{\boldsymbol{m}}) &= \sum_{k=1}^{K} \left[v_k(m_k; z) - P_k \left(a \max\{m_k - 1, 0\} + b \sum_{l; l \neq k} m_l \right) \right] \\ &= \sum_{k=1}^{K} \left[v_k(m_k; z) - (P_k a \max\{m_k - 1, 0\} + (1 - P_k) b m_k) \right]. \end{aligned} \tag{73}$$

Since $V(z; \hat{I}_{\boldsymbol{m}})$ splits into a sum of K terms that are functions of $m_1, \ldots, m_K$ respectively, it follows that $V(z; \hat{I}_{\boldsymbol{m}})$ is maximized, as a function of $\boldsymbol{m}$, by maximizing each term separately with respect to m_k. The maximum for term k, on the right hand side of Eq. 73, is attained for

$$m_k(z) = \arg\max_{0 \leq m_k \leq N_k} \left[v_k(m_k; z) - (P_k a \max\{m_k - 1, 0\} + (1 - P_k) b m_k) \right], \tag{74}$$

in accordance with the first identity of Eq. 71. The second identity of Eq. 71 follows from the fact that the function being maximized in Eq. 74 is concave in m_k.

To give some more intuition to the choice of a and b for the reward function Eq. 70, we will look at the penalty term

$$g_k(m_k, m - m_k) = a \max(m_k - 1, 0) + b(m - m_k)$$

and the optimal classifier $\hat{I}(z) = \hat{I}_{\boldsymbol{m}(z)}$ defined in Eqs. 71-72 of Corollary 6. Note that a cost of a is incurred per *extra* category from the correct block in the classified set $\hat{I}$, whereas a cost of b is added for each category in the classified set originating from the wrong block. These costs are chosen and can be interpreted as threshold values for including categories in the classifier, especially when looking at Eq. 72. From the first row of this equation, we notice that a low value on b means that we are more prone to include the most probable category from each block, whereas the second row implies that with a low value of a we are more prone to include several categories from the same block. However, since b occurs in the second row of Eq. 72 as well, a small a might not have any effect if b is large. On the other hand, by combining a small b with a large a, we get a classifier that is composed of categories from many blocks, but few categories from each block. Such a classifier might not include the correct category, but will to a large extent include some category from the correct block, and might be suitable when we want to safeguard in particular against erroneous superclassification. Finally, if we want to ensure that $\hat{I} \neq \emptyset$ in composite classification, we have to choose

$$b \leq \max\left(\frac{p_{(1N_1)}}{1 - P_1}, \ldots, \frac{p_{(KN_K)}}{1 - P_K} \right). \tag{75}$$

Let us end Example 7 by considering two special cases of the proportion-based reward function Eq. 70. The first one occurs when $a = b = c$, and it corresponds to a reward function

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - c\left[|\mathcal{I}| - 1(|\mathcal{I}_{k(i)}| > 0) \right] \tag{76}$$

that differs slightly from Eq. 20 in that all categories in the classified set $\mathcal{I}$ are penalized by c when none of them belong to the correct block $k(i)$, that is, when $\mathcal{I}_{k(i)} = \emptyset$.

The second special case of Eq. 70 occurs when it is known that observation z belongs to block k, i.e., $P_k(z) = 1$. The classifier $\hat{I}_{\boldsymbol{m}(z)}$ in Eq. 59, with $\boldsymbol{m}(z) = (m_1(z), \ldots, m_K(z))$ as in Eq. 71, then simplifies to

$$\begin{aligned} m_k(z) &= \max\left\{1, \max\{2 \le m_k \le N_k;\ p_{(k,N_k+1-m_k)} \ge a\}\right\}, \\ m_l(z) &= 0,\ l \ne k. \end{aligned}$$

This corresponds to an m-MP classifier Eq. 12, with $m = m(z)$ as in Eqs. 17 and 22, when classification is restricted to categories within class k, and a penalty $c = a$ is incurred per extra classified category.

5.3 Multiplicative Penalties of Sizes of Classified Sets

We will now consider block-invariant reward functions

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) g_{k(i)}(|\mathcal{I}_{k(i)}|, |\mathcal{I}| - |\mathcal{I}_{k(i)}|) \tag{77}$$

for which the penalty $g_{k(i)}(\cdot, \cdot) \ge 0$ for the size of the classified set is multiplicative, non-negative, and block-dependent. Since $R(\emptyset, i) = 0$ for all $i = 1, \ldots, N$, it clearly suffices to consider nonempty sets $\mathcal{I} \ne \emptyset$. Proposition 5 links Bayes classifiers of reward functions Eq. 77 to $\boldsymbol{m}$-MP-classifiers. It is proved in the same way as Proposition 4.

Proposition 5 *The optimal classifier, for a multiplicative reward function Eq. 77, is an $\boldsymbol{m}(z)$-MP classifier $\hat{I}(z) = \hat{I}_{\boldsymbol{m}(z)}$ in Eq. 59, with*

$$\boldsymbol{m}(z) = \arg\max_{\boldsymbol{m}=(m_1,\ldots,m_K)} \sum_{k=1}^{K} v_k(m_k; z) g_k(m_k, m - m_k), \tag{78}$$

where $m = \sum_{k=1}^{K} m_k$, $v_k(m_k; z) = \sum_{j=1}^{m_k} p_{(k,N_k+1-j)}$ for $m_k > 0$ and $v_k(0; z) = 0$.

Example 8 (Indifference zones.) Suppose we want to know which of $N - 1$ normally distributed populations with unit variances and expected values $\theta_1, \ldots, \theta_{N-1}$ has the largest expected value. Let

$$Z = (Z_{ij};\ 1 \le i \le N - 1, 1 \le j \le n_i)$$

be a sample of independent random variables, with $Z_{ij} \sim N(\theta_i, 1)$. Letting ϕ be the density function of a standard normal distribution, this gives a likelihood

$$f(z; \theta) = \prod_{i=1}^{N-1} \prod_{j=1}^{n_i} \phi(z_{ij} - \theta_i)$$

with a parameter vector $\theta = (\theta_1, \ldots, \theta_{N-1})$ that is assumed to have a prior density $P(\theta)$. Divide the parameter space $\Theta = \mathbb{R}^{N-1}$ into a disjoint union

$$\Theta = \Theta_1 \cup \ldots \cup \Theta_N \tag{79}$$

of N regions, where

$$\begin{aligned} \Theta_i &= \{\theta;\ \theta_i = \theta' \ge \theta'' + \epsilon\}, \quad i = 1, \ldots, N - 1, \\ \Theta_N &= \Theta \setminus (\Theta_1 \cup \ldots \cup \Theta_{N-1}), \end{aligned} \tag{80}$$

θ' and θ'' are the largest and second-largest expected values respectively, with $\epsilon > 0$ a small number. Whereas $\Theta_1, \ldots, \Theta_{N-1}$ correspond to all hypotheses where some parameter θ_i is largest by a margin of at least ϵ, Θ_N is the *indifference zone*, where none of the populations has an expected value that is at least ϵ units larger than all the others (Bechhofer, 1954; Goldsman, 1986).

It is possible to put this model into the framework of Section 2, with

$$\begin{aligned} \pi_i &= \textstyle\int_{\Theta_i} P(\theta) d\theta, \\ f_i(z) &= \textstyle\int_{\Theta_i} f(z;\theta) P(\theta) d\theta / \int_{\Theta_i} P(\theta) d\theta, \end{aligned} \tag{81}$$

for $i = 1, \ldots, N$, where the first $N-1$ categories represent a population, whereas category N represents the indifference zone. We will consider reward functions

$$R(\mathcal{I}, i) = \begin{cases} 1(\mathcal{I} = \{i\}), & i = 1, \ldots, N-1, \\ r1(\mathcal{I} = \{N\}), & i = N, \end{cases} \tag{82}$$

where $r > 0$ is the reward of not selecting any population as having the largest mean when the parameter vector belongs to the indifference zone. This corresponds to a block invariant reward function with the two blocks

$$\begin{aligned} \mathcal{N}_1 &= \{1, \ldots, N-1\}, \\ \mathcal{N}_2 &= \{N\} \end{aligned} \tag{83}$$

of categories, although misclassification in this example is more serious *within* than between blocks. Note that Eq. 82, with blocks as in Eq. 83, is an instance of a multiplicative reward functions in Eq. 77, with penalty

$$g_k(m_k, m - m_k) = \begin{cases} 1, & k = 1, m_1 = m = 1, \\ r, & k = 2, m_2 = m = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{84}$$

It follows from Proposition 4 that the optimal classifier is

$$\hat{I} = \begin{cases} \{(1, N-1)\}, & p_{(1,N-1)} \geq r p_N, \\ \{N\}, & p_{(1,N-1)} < r p_N. \end{cases}$$

Note that Eq. 82 is also an instance of Eq. 62, with a multiplicative penalty $g(m) = 1(m = 1)$, and block dependent rewards $d_1 = 1$ and $d_2 = r$.

Example 9 (Hierarchical classification.) Suppose the N categories have a hierarchical structure, as leaves of a tree with L levels. The inner nodes of this tree consist of multiple categories, and the levels of the tree are ordered so that the leaves are at level $l = 0$ and the inner nodes at levels $l = 1, \ldots, L-1$. Each inner node has a number of child nodes that contain disjoint sets of categories, with a union that equals the categories of the parent node. The level of a node is the longest ancestral path directed from this node to a leaf descendant. In particular, the root of the tree, at level $L-1$, corresponds to all nodes $\mathcal{N}$. A recent review of hierarchical classification is provided in Mortier et al. (2022). A typical hierarchical classifier enforces $\hat{I}$ to be a node of the tree. With this restriction, classification with a reject option (Example 3) corresponds to a hierarchical classifier with $L = 2$ levels, where $\hat{I}$ either equals $\mathcal{N}$ (the root of the tree) or a single category (the leaves). Set-valued classification with blocks of categories corresponds to hierarchical classification with $L = 3$ levels, where the K nodes of the intermediate level 1 correspond to the blocks $\mathcal{N}_1, \ldots, \mathcal{N}_K$ of categories. A hierarchical classifier with $L = 3$ levels, where each block is penalized in the same way, is obtained

from a reward function that assigns a reward of 1 to the true category $\{i\}$, a reward of r_1 to the true block $\mathcal{N}_{k(i)}$ of categories, and a reward of r_2 to the whole set $\mathcal{N}$ of categories, for some numbers $0 < r_2 < r_1 < 1$, whereas all other sets are assigned a reward of 0. This corresponds to a multiplicative reward function Eq. 77 with penalty

$$g_k(m_k, m - m_k) = \begin{cases} 1; & \text{if } m_k = m = 1, \\ 0, & \text{if } m_k = m \in \{0, 2, 3, \ldots, N_k - 1\}, \\ r_1, & \text{if } m_k = m = N_k, \\ 0, & \text{if } m_k < m < N, \\ r_2, & \text{if } m = N. \end{cases} \tag{85}$$

Let $k(z)$ refer to the block that maximizes $P_k(z)$ in Eq. 68. It follows from Proposition 5 that the Bayes estimator for a reward function with multiplicative penalty Eq. 85 equals

$$\hat{I}(z) = \begin{cases} (N), & p_{(N)} > \max(r_1 P_{k(z)}, r_2), \\ \mathcal{N}_{k(z)}, & r_1 P_{k(z)} > \max(p_{(N)}, r_2), \\ \mathcal{N}, & r_2 > \max(p_{(N)}, r_1 P_{k(z)}). \end{cases}$$

Mortier et al. (2022) consider more general hierarchical classifiers that are not restricted to output any of the nodes of the hierarchical tree. In particular, they propose a classifier $\hat{I}$ that maximizes the expectation of the recall reward in Eq. 45, subject to upper bounds on its size $|\hat{I}|$ and representation complexity $n(\hat{I})$ (the minimal number of nodes, of the hierarchical tree, sufficient for describing $\hat{I}$). A closely related classifier can also be obtain by maximizing the expectation of the multiplicative reward

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) a^{|\mathcal{I}|-1} b^{n(\mathcal{I})-1}, \tag{86}$$

for some constants $0 < a < 1$ and $0 \leq b \leq 1$, with $0^0 = 1$ when $b = 0$. Note that the penalty term $g(\mathcal{I}) = a^{|\mathcal{I}|-1} b^{n(\mathcal{I})-1}$ does not conform with Eq. 65, since it is not a function of $|\mathcal{I}_{k(i)}|$ and $|\mathcal{I}| - |\mathcal{I}_{k(i)}|$. The case $b = 1$ treats all categories as one single block, and the resulting multiplicative reward function is a special case of Eq. 32, and consequently it belongs to the framework of Section 4.2. At the other extreme, $b = 0$ gives zero reward to any set that comprises more than one node of the tree. Hence it gives rise to a hierarchical classifier that enforces single nodes of the tree as outputs. In particular, when all block sizes are equal ($N_k = N/k$), the reward function Eq. 86 is a special case of Eq. 85 with $b = 0$, $r_1 = a^{N/k-1}$ and $r_2 = a^{N-1}$.

6 Illustration for An Ornithological Data Set

In this section taxon identification will be used as a case study in order to illustrate the use of the composite proportion-based reward function Eq. 70 for classification, with a data-driven choice of the cost parameters a and b. To this end, we will look at four bird species that are morphologically similar, but share three measurable traits. In Karlsson and Hössjer (2023) we treat the underlying fitting problem in detail for the same four bird species, assuming that they form one homogeneous block of categories, with covariates, heteroscedasticity, missing values, and imperfect observations. Since this paper has a stronger emphasis on developing a theory of classification, we use a subset of data from Karlsson and Hössjer (2023) for the purpose of illustration, and moreover we divide the four species into three blocks. The reduced data set that we focus on here includes complete observations only from a certain stratum of the population, eliminating the need for covariates. To a large extent, our data is

the same as in Malmhagen et al. (2013), where the same classification problem is treated with mostly descriptive statistics.

6.1 Data Set

The four species considered are *Reed warbler*, *Marsh warbler*, *Blyth's reed warbler* and *Paddyfield warbler*, and the three shared traits are *wing length*, *notch length* and *notch position*. For details on the measurements of the traits, see for instance Svensson (1992); Malmhagen et al. (2013) and Karlsson and Hössjer (2023). In total, we have 882 complete observations of juvenile birds, with the number n_i of birds of each species given in Table 1. This gives rise to a training data set

$$\mathcal{D}_i = \{z_{ij};\ j = 1, \ldots, n_i\}$$

for each species $i = 1, \ldots, 4$.

The species were partitioned as follows: *Reed warbler* and *Marsh Warbler* constitute the block "common breeders", *Blyth's reed warbler* constitutes the block "rare breeder" and *Paddyfield warbler* constitutes the block "rare vagrant". We acknowledge that the grouping is quite arbitrary and that the block names are inaccurate in most places of the world, but here it will mainly illustrate classification with a partitioned label space.

6.2 Model

We assume that the parameters associated with each category are independent. If $\theta_i \in \Theta_i$ is the parameter vector associated with category i, with a prior distribution $P(\theta_i)$, and if $f(\mathcal{D}_i; \theta_i)$ refers to the likelihood of training data for taxon i, the posterior distribution of θ_i is

$$P(\theta_i|\mathcal{D}_i) = \frac{P(\theta_i) f(\mathcal{D}_i|\theta_i)}{P(\mathcal{D}_i)}. \tag{87}$$

Let z be an observed new data point that we wish to classify, based on training data. If $f(z; \theta_i)$ is the likelihood of the test data point for taxon i, we integrate over the posterior distribution Eq. 87 in order to obtain the corresponding likelihood

$$f_i(z) = \int_{\Theta_i} f(z; \theta_i) P(\theta_i|\mathcal{D}_i) d\theta_i \tag{88}$$

of z for the hypothesis H_i that corresponds to this taxon. Then, we insert Eq. 88 into Eq. 2 in order to obtain the posterior probabilities $p_i(z)$ of all categories. In our example we assume that $z_{ij} \sim \text{MVN}(\mu_i, \Sigma_i)$ has a multivariate normal distribution $f(\cdot; \theta_i)$, with $\theta_i = (\mu_i, \Sigma_i)$.

Table 1 The number of observations of each species

Species	n_i
Reed warbler	409
Blyth's reed warbler	41
Paddyfield warbler	18
Marsh warbler	414

We estimate Eq. 88 using Monte Carlo simulation, i.e., we simulate L realizations of $\theta_{il} = (\mu_{il}, \Sigma_{il})$ from $P(\theta_i \mid \mathcal{D}_i)$ for each i, compute

$$\hat{f}_i(z) = \frac{1}{L}\sum_{l=1}^{L} f(z; \theta_{il}) \tag{89}$$

and then plug Eq. 89 into Eq. 2. For a more detailed model setup, see Section 3.1 and Appendix A of Karlsson and Hössjer (2023). All implementation was done in R (R Core Team, 2021), using the package `mvtnorm` (Genz et al., 2021).

6.3 Classifiers Based on Composite Proportion-Based Re-ward Functions

As mentioned above, will derive Bayesian classifiers from the composite proportion-based reward function Eq. 70. This reward function involves the two constants $a \geq 0$ and $b > 0$, and the corresponding classifier of z is denoted $\hat{I}_{(a,b)}(z)$. We will regard $0 \leq \varepsilon = a/b$ as a fixed parameter that quantifies how much more severe it is to misclassify a category outside a block than inside it, with severity inversely proportional to ε. If $0 \leq \varepsilon < 1$, it is more severe to misclassify between blocks than within, whereas the opposite is true when $\varepsilon > 1$. Note in particular the reward functions II and III of Section 3 are composite-based reward functions with $\varepsilon = 1$ and $\varepsilon = 0$ respectively.

The parameter b will be chosen through leave-one-out cross-validation. To this end, let $\tilde{R}(\mathcal{I}, i)$ be a binary-valued reward function (to be chosen below) without any penalty term, and let

$$R_{ab}^{\mathrm{cv}}(\tilde{R}) = \sum_{i=1}^{N} \frac{w_i}{n_i} \sum_{j=1}^{n_i} \tilde{R}(\hat{I}_{(a,b)}(z_{ij}), i) \tag{90}$$

refer to the fraction of observations z_{ij} in the training data set $\mathcal{D} = \mathcal{D}_1 \cup \ldots \cup \mathcal{D}_N$ that return a reward in the cross-validation procedure. That is, $R_{ab}^{\mathrm{cv}}(\tilde{R})$ equals the fraction of observations z_{ij} with $\tilde{R}(\hat{I}_{(a,b)}(z_{ij}), i) = 1$, where $\hat{I}_{(a,b)}(z_{ij})$ is the classifier of z_{ij} based on the rest of the data. It is further assumed that w_i are non-negative weights that sum to 1, such as $w_i = 1/N$ or $w_i = n_i / \sum_{j=1}^{N} n_j$.

The choice of b will depend on which reward function $\tilde{R}$ that is used in Eq. 90. Some possible choices of the binary-valued reward function are given in Table 2. For two of them ($\tilde{R}_3$ and $\tilde{R}_4$) the reward $\tilde{R}(\mathcal{I}, i)$ is a non-decreasing function of $\mathcal{I}$, and therefore the non-reward rate $1 - R_{\varepsilon b,b}^{\mathrm{cv}}$, obtained from the cross validation procedure Eq. 90, is a non-decreasing function of b. We will therefore choose b as the largest cost parameter for which the non-reward rate is at most $\delta > 0$, i.e.,

$$b_{\varepsilon\delta} = \max\{b \geq 0;\ 1 - R_{\varepsilon b,b}^{\mathrm{cv}} \leq \delta\}. \tag{91}$$

In particular, it can be seen that when reward functions $\tilde{R}_3$ is used in Eq. 91, $b_{\varepsilon\delta}$ corresponds to the largest choice of b such that the sets $\hat{I}_{(\varepsilon b,b)}(z_{ij})$ are still large enough for a fraction $1-\delta$ of them to cover the true categories of the training dataset. Consequently, this generalizes an instance of the conformal algorithm (Shafer & Vovk, 2008, Section 4.3) from the case of one block ($K = 1$) to several blocks ($K > 1$), although we use cross-validation from training data rather than a prediction of new observations, as in Shafer and Vovk (2008).

Table 2 The four binary reward functions used in the case study

Binary reward function	*Reward criteria*
$\tilde{R}_1(\mathcal{I}, i) = 1(\mathcal{I} = \{i\})$	Correct (point) classification
$\tilde{R}_2(\mathcal{I}, i) = 1(i \in \mathcal{I} \wedge \mathcal{I} \subseteq \mathcal{N}_{k(i)})$	Correct category is in the classifier, and no category from an incorrect block is in the classifier. The analogy of $\tilde{R}_1$ for block prediction.
$\tilde{R}_3(\mathcal{I}, i) = 1(i \in \mathcal{I})$	Correct category in classifier
$\tilde{R}_4(\mathcal{I}, i) = 1(\mathcal{I} \cap \mathcal{N}_{k(i)} \neq \emptyset)$	Some category from the correct block in the classifier. The analogy of $\tilde{R}_3$ for block prediction.

Since $\tilde{R}_1(\mathcal{I}, i) \leq \tilde{R}_2(\mathcal{I}, i) \leq \tilde{R}_3(\mathcal{I}, i) \leq \tilde{R}_4(\mathcal{I}, i)$, it follows that $\tilde{R}_1$ is the least generous reward function and $\tilde{R}_4$ the most generous one. Note that $R^{\text{cv}}_{\varepsilon b,b}(\tilde{R}_3)$ and $R^{\text{cv}}_{\varepsilon b,b}(\tilde{R}_4)$ are both decreasing functions of b, so that Eq. 91 makes sense for choosing b for any of these two choices of $\tilde{R}$, whereas Eq. 92 is more appropriate for choosing b for the other two reward functions

For the other two reward functions $\tilde{R}_1$ or $\tilde{R}_2$ of Table 2, it is no longer the case that estimated non-reward rate $1 - R^{\text{cv}}_{\varepsilon b,b}$ in Eq. 90 is monotonic in b. For these two choices of $\tilde{R}$ we rather choose the cost parameter

$$b_\varepsilon = \arg\min_{b \geq 0}(1 - R^{\text{cv}}_{\varepsilon b,b}) \tag{92}$$

in order to minimize the estimated non-reward rate. The rationale for Eq. 92 is that more classified sets $\hat{I}_{(\varepsilon b,b)}(z_{ij})$ of training data will be empty for larger b, whereas more of them will include several species/groups of species for smaller b. And both of these features will increase the non-reward rate $1 - R^{\text{cv}}_{\varepsilon b,b}$ when $\tilde{R}_1$ or $\tilde{R}_2$ is used to quantify rewards.

We will look at three different prior distributions, namely a uniform prior ($\pi^{(\text{flat})}$), a prior proportional to the number of observations n_i from each category in training data ($\pi^{(\text{prop})}$), and a prior proportional to the number of registered birds of each species at the Falsterbo Bird observatory throughout its operational history ($\pi^{(\text{real})}$). The purpose of $\pi^{(\text{flat})}$ is to represent a situation of no prior knowledge of how likely any of the categories is to occur. The prior $\pi^{(\text{real})}$ is supposed to exemplify a real-world situation of having some commonly occurring species and some very rare ones, whereas $\pi^{(\text{prop})}$ is a middle ground between these two extremes that fits the data set $\mathcal{D}$ very well. As weights we choose

$$\begin{aligned} w_i^{(\text{bird})} &= n_i / \sum_{j=1}^{4} n_j, \\ w_i^{(\text{spec})} &= 1/4, \\ w_i^{(\text{rare})} &= 1 / \sum_{j=1}^{4} (\pi_i^{(\text{real})} / \pi_j^{(\text{real})}), \end{aligned} \tag{93}$$

for $i = 1, 2, 3, 4$. We weight each observation equally with $w^{(\text{bird})}$, each species equally with $w^{(\text{spec})}$, whereas the species are weighted higher the less expected they are with $w^{(\text{rare})}$. Since the number of observations is not balanced across species, $w^{(\text{bird})}$ will weight species higher the more common they are, i.e., the more observations we have of them. Using $w^{(\text{spec})}$, birds will be weighted unequally due to the same imbalance (the less birds of a given species there are, the higher weights are assigned to these birds). Finally, as mentioned above, $w^{(\text{rare})}$

weights less frequently observed species more heavily; the rationale is that we value observations of rarely occurring species, as data on these are scarce.

6.4 Results

We will analyze the estimated reward rate Eq. 90 for the four choices of $\tilde{R}$ that are listed in Table 2. In Tables 3 and 4, we present the automatic choice of the cost parameter b (cf. Eqs. 91 and 92) for two ratios $\varepsilon = 1/2$ and $\varepsilon = 2$ of a and b, and for all nine combinations priors and weights. As can be seen from Table 4, the same optimal b-values are found for $\tilde{R}_1$ and $\tilde{R}_2$, with the same non-reward rates. This is mostly due to the small block sizes, meaning that it sometimes is equivalent to picking the correct species, as to pick the correct block. Comparing the three prior distributions, we see from Table 4 that $\pi^{(\text{prop})}$ overall has the smallest non-reward rates for $\tilde{R}_1$ and $\tilde{R}_2$ from training data, followed by $\pi^{(\text{flat})}$ and $\pi^{(\text{real})}$. This is consistent with Table 3 where $\pi^{(\text{prop})}$ overall gives the largest values of b, and hence the smallest classified sets $\hat{I}_{(\varepsilon b,b)}(z_{ij})$, that are sufficient to guarantee a reward rate of at least $1-\delta$. Note also that the non-reward rates of $\tilde{R}_1$ and $\tilde{R}_2$ are very large when rare species are assigned high weights apriori ($\pi^{(\text{real})}$), more so the higher weights these rare species are given in the cross-validation scheme. This is expected, because of a lack of data to classify and validate the rare species well.

In Figs. 1 and 2, we plot the value of estimated non-reward rate $1 - R^{\text{cv}}_{\varepsilon b,b}$ for a grid of b-values, for $\varepsilon = 1/2$ and $\varepsilon = 2$ respectively. Note the monotone decrease of $\tilde{R}_3$ and $\tilde{R}_4$ as b decreases. Also, notice the minimums of $\tilde{R}_1$ and $\tilde{R}_2$ in the graphs.

Finally we refer to Appendix A for further visualizations of $\tilde{R}_1$ and $\tilde{R}_2$, evaluated over a lattice of a and b-values. It can seen that for $\tilde{R}_2$ the optimal non-reward rate is achieved by choosing $a = 0$. This is straightforward to explain, as $\tilde{R}_2$ does not punish the inclusion of several categories from the correct block. For this reason the classifier $\hat{I}_{(a,b)}$ with minimal

Table 3 The table specifies the estimated values of the cost parameter b, using Eq. (91) with $\delta = 0.05$

ε	Prior	$\tilde{R}$	$b_{\varepsilon,0.05}$		
			$w^{(\text{bird})}$	$w^{(\text{spec})}$	$w^{(\text{rare})}$
1/2	$\pi^{(\text{flat})}$	$\tilde{R}_3$	≥ 20.00	2.29	2.11
		$\tilde{R}_4$	≥ 20.00	3.05	2.11
	$\pi^{(\text{prop})}$	$\tilde{R}_3$	≥ 20.00	5.23	4.81
		$\tilde{R}_4$	≥ 20.00	6.96	4.81
	$\pi^{(\text{real})}$	$\tilde{R}_3$	1.07	0.43	0.18
		$\tilde{R}_4$	≥ 20.00	0.52	0.18
2	$\pi^{(\text{flat})}$	$\tilde{R}_3$	≥ 20.00	2.29	2.11
		$\tilde{R}_4$	≥ 20.00	3.05	2.11
	$\pi^{(\text{prop})}$	$\tilde{R}_3$	≥ 20.00	5.23	4.81
		$\tilde{R}_4$	≥ 20.00	6.96	4.81
	$\pi^{(\text{real})}$	$\tilde{R}_3$	1.07	0.21	0.18
		$\tilde{R}_4$	≥ 20.00	0.52	0.18

These estimates of b are computed for each combination of ε, prior π_i, and weights w_i. We evaluated Eq. (90) for $0.01 \leq b \leq 20$ with a resolution of 0.01

Table 4 The table specifies the estimated values of the cost parameter b, using Eq. 92. These estimates of b are computed for each combination of ε, prior π_i, and weights w_i. They were found using the `optimise`-function in R

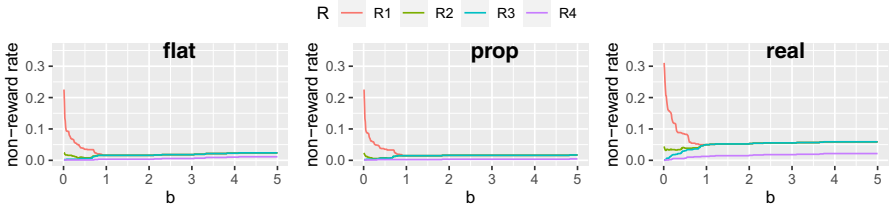
ε	Prior	$\tilde{R}$	$w^{(\text{bird})}$		$w^{(\text{spec})}$		$w^{(\text{rare})}$	
			b_ε	non-reward rate	b_ε	non-reward rate	b_ε	non-reward rate
$\frac{1}{2}$	$\pi^{(\text{flat})}$	$\tilde{R}_1$	1.24	1.59%	17.16	9.12%	1.16	4.06 %
		$\tilde{R}_2$	1.24	1.59%	17.16	9.12%	1.16	4.06%
	$\pi^{(\text{prop})}$	$\tilde{R}_1$	2.63	1.59%	2.24	4.06%	2.63	4.06%
		$\tilde{R}_2$	2.63	1.59%	2.24	4.06%	2.63	4.06%
	$\pi^{(\text{real})}$	$\tilde{R}_1$	10.28	6.69%	10.28	19.41%	10.28	49.62%
		$\tilde{R}_2$	10.28	6.69%	10.28	19.41%	10.28	49.62%
2	$\pi^{(\text{flat})}$	$\tilde{R}_1$	1.24	1.59%	17.16	9.12%	1.16	4.06%
		$\tilde{R}_2$	1.24	1.59%	17.16	9.12%	1.16	4.06%
	$\pi^{(\text{prop})}$	$\tilde{R}_1$	2.63	1.59%	2.24	4.06%	2.63	4.06%
		$\tilde{R}_2$	2.63	1.59%	2.24	4.06%	2.63	4.06%
	$\pi^{(\text{real})}$	$\tilde{R}_1$	10.28	6.69%	10.28	19.41%	10.28	49.62%
		$\tilde{R}_2$	10.28	6.69%	10.28	19.41%	10.28	49.62%

non-reward rate includes as many categories as possible from each block with at least one classified member, corresponding to $a = 0$. Notice also that there are large regions of values of a and b that attain the minimum non-reward rate.

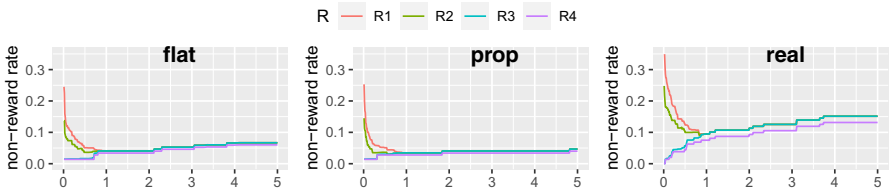
7 Discussion

In this article, we introduce a general framework of set-valued classification of data that originates from one of a finite number of possible hypotheses. Using reward functions with a set-valued input argument, we investigate the properties of the optimal (Bayes) classifier by maximizing the expected reward. Explicit formulas for the Bayes classifier are derived for a large class of reward functions, many of which either extend or unify previous work on set-valued classification. Our work includes scenarios where hypotheses either constitute one homogeneous block or can be divided into several blocks, such that ambiguity within blocks of hypotheses is (typically) less serious than ambiguity between these blocks. We illustrate the latter type of model with an ornithological data set, where taxa (hypotheses) are divided into blocks. In particular, a cross-validation-based algorithm is introduced for estimating a cost parameter of the reward function from training data, in order to guarantee a minimal fraction of observations from training data that are included in the corresponding classified sets.

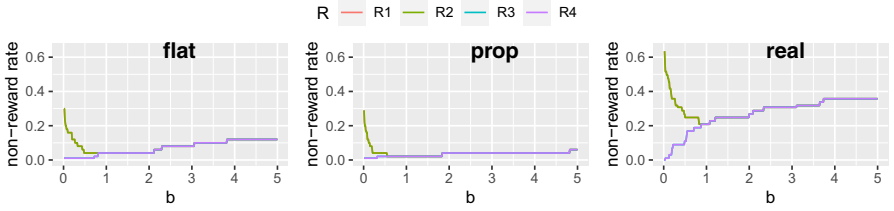
As mentioned in Ripley (2007), a possible reason for including reject options is to obtain classifiers that are more reliable but also less expensive to use than a precise classifier Eq. 3 that always outputs singleton sets. In our case study of Section 6, for instance, a possible option when $|\hat{I}| > 1$ is to consult an expert who would be able to identify the bird species morphologically, without using the measured traits. Although expertise does not come cheap, this could still be an alternative when the expected cost of precise classification exceeds the



(a) Using $w^{(\text{bird})}$ we observe similar curves for $\pi^{(\text{flat})}$ and $\pi^{(\text{prop})}$, whereas $\pi^{(\text{real})}$ gives overall higher non-reward rates.



(b) Using $w^{(\text{spec})}$, the curves for $\tilde{R}_1$ and $\tilde{R}_2$ look similar for $\pi^{(\text{flat})}$ and $\pi^{(\text{prop})}$, whereas the small values for $\tilde{R}_3$ and $\tilde{R}_4$ occur either for small values of b ($\pi^{(\text{flat})}$) or for all values of b ($\pi^{(\text{prop})}$). Again, the non-reward rates are overall higher using $\pi^{(\text{real})}$.

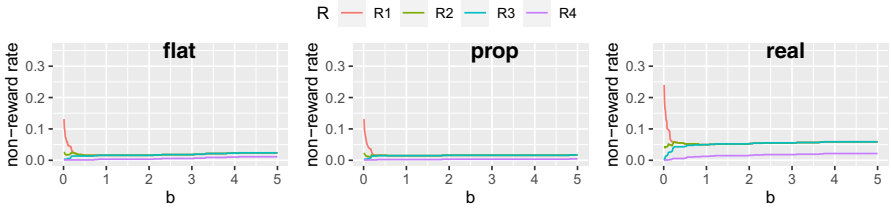


(c) The scale along the vertical axis (for the non-reward rate) is different in these subplots compared to (a) and (b). Using $w^{(\text{rare})}$, with a very uneven weighting of species, the curves for $\tilde{R}_1$ and $\tilde{R}_2$, as well as for $\tilde{R}_3$ and $\tilde{R}_4$, take values very close to each other. Thus it seems like these curves overlap, when in reality they do not. This is just a consequence of the extreme amount of up-weighting of rare species, which are both in singleton blocks.

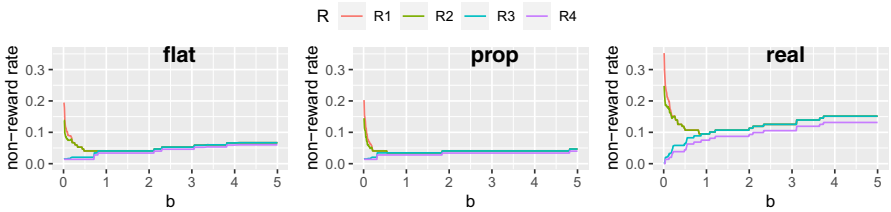
Fig. 1 This figure represents the case $\varepsilon = 1/2$. The prior is specified in the title of each graph, whereas the weights are explained in the subcaptions. Each color of the functions in the graphs corresponds to one of the four reward functions $\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \tilde{R}_4$, given by the legends above each subfigure. For all priors and weights observe that the non-reward rates of $\tilde{R}_3$ and $\tilde{R}_4$ decrease monotonically as b decreases, whereas those of $\tilde{R}_1$ and $\tilde{R}_2$ have a global minimum

cost of consulting an expert. The latter cost might be independent or a function of the number of hypotheses she needs to consider. In the former case, the reject option of Ripley (2007) would suffice, and in the latter case a partial rejection to classify could be beneficial.

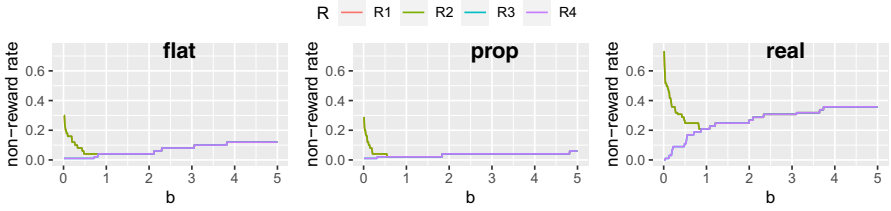
A number of generalizations of our work are possible, which we divide into four subsections.



(a) Using $w^{(\text{bird})}$ we observe similar curves for $\pi^{(\text{flat})}$ and $\pi^{(\text{prop})}$, whereas $\pi^{(\text{real})}$ gives overall higher non-reward rates.



(b) Using $w^{(\text{spec})}$, the curves for $\tilde{R}_1$ and $\tilde{R}_2$ look similar for $\pi^{(\text{flat})}$ and $\pi^{(\text{prop})}$, whereas the small values for $\tilde{R}_3$ and $\tilde{R}_4$ occur either for small values of b ($\pi^{(\text{flat})}$) or for all values of b ($\pi^{(\text{prop})}$). Again, the non-reward rates are overall higher using $\pi^{(\text{real})}$.



(c) The scale along the vertical axis (for the non-reward rate) is different in these subplots compared to (a) and (b). Using $w^{(\text{rare})}$, with a very uneven weighting of species, the curves for $\tilde{R}_1$ and $\tilde{R}_2$, as well as for $\tilde{R}_3$ and $\tilde{R}_4$, take values very close to each other. Thus it seems like these curves overlap, when in reality they do not. This is just a consequence of the extreme amount of up-weighting of rare species, which are both in singleton blocks.

Fig. 2 This figure represents the case $\varepsilon = 2$. The prior is specified in the title of each graph, whereas the weights are explained in the subcaptions. Each color of the functions in the graphs corresponds to one of the four reward functions $\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \tilde{R}_4$, given by the legends above each subfigure. For all priors and weights we observe that the non-reward rates of $\tilde{R}_3$ and $\tilde{R}_4$ decrease monotonically as b decreases, whereas those of $\tilde{R}_1$ and $\tilde{R}_2$ have a global minimum

7.1 Reward Functions Generated from Single Set Costs

Suppose the reward function satisfies $0 \leq R(\mathcal{I}, i) = 1 - C(\mathcal{I}, i) \leq 1$, with $C(\mathcal{I}, i)$ the cost of classifying $\mathcal{I}$ when i is true. Yang et al. (2017) started with a cost function $C(\{j\}, i)$ for singleton sets, and then defined p-discounted costs

$$1 - R(\mathcal{I}, i) = C(\mathcal{I}, i) = \left[\frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} C(\{j\}, i)^p \right]^{1/p} \tag{94}$$

for all $\mathcal{I} \neq \emptyset$, with $0 < p \leq 1$ a fixed parameter. When all categories belong to one homogeneous block (Section 4), it is natural to make use of the 0–1 loss $C(\{j\}, i) = 1(j \neq i)$.

Then, Eq. 94 simplifies to a reward function

$$R(\mathcal{I}, i) = 1 - \left[\frac{1}{|\mathcal{I}|} \sum_{j \in \mathcal{I}} 1(j \neq i) \right]^{1/p} = 1 - \left[1 - \frac{I(i \in \mathcal{I})}{|\mathcal{I}|} \right]^{1/p} \tag{95}$$

that is multiplicative Eq. 32, with a penalty

$$g(m) = 1 - (1 - \frac{1}{m})^{1/p}.$$

Note that $p = 1$ corresponds to the discounted accuracy reward Eq. 44, and that $g(m) = h(1/m)$ satisfies the properties of Corollary 5 for any value of p, since $h(x) = 1-(1-x)^{1/p}$ is concave on $[0, 1]$. It is of interest to study properties of Bayesian classifiers based on p-discounted reward functions Eq. 94 more generally, for other choices of $C(\{j\}, i)$ that correspond to blocks of categories.

7.2 Conditional Reward Functions

7.2.1 Regression Models

Suppose for instance that the new observation z that we want to classify, by means of the optimal classifier $\hat{I} = \hat{I}(z)$ in Eq. 10, involves a covariate vector x and a response variable y. Following Karlsson and Hössjer (2023), the most straightforward approach is to include covariates into the observation $z = (x, y)$ that is to be classified. The covariate information will then be included in the category distributions f_i, in the posterior probabilities p_i of all categories $i = 1, \ldots, N$, and in the resulting Bayes classifiers. However, if we also want the ambiguity of the classifier to depend on covariate information, it is possible to consider a class of reward functions $R(\mathcal{I}, i, x)$ that not only depend on the classified set $\mathcal{I}$ and the true category i, but also on the covariate x of a new observation that is to be classified. For instance, in the case of one homogeneous block of categories (Section 4),

$$R(\mathcal{I}, i, x) = 1(i \in \mathcal{I}) - c(x)|\mathcal{I}|$$

is a version of the linear reward function Eq. 21 where the cost parameter $c = c(x)$ is covariate dependent. The conformal prediction algorithm of Example 2 involves choosing the cost $c(\delta) = F^{-1}(\delta) = Q(\delta)$ of including more labels in the classifier, as a quantile $Q(\delta)$ of the distribution F of posterior probabilities (cf. Eq. 30). In this context, it is possible to use quantile regression (Koenker, 2005; Bottai et al., 2010) and rather choose the cost parameter

$$c(\delta; x) = F^{-1}(\delta|x) = Q(\delta|x) = g^{-1}[x^\top \beta(\delta)]$$

as a conditional quantile function corresponding to $F(t|x) = P(p_I(x, Y) \leq t)$. This is a regression model where the parameter vector $\beta(\delta)$ is a function of the quantile δ, whereas g is a link function. This approach might be particularly helpful for models with heteroscedasticity, or if the cost of imprecise classification is covariate dependent.

7.2.2 Conditional Coverage

Another possibility is to consider reward functions $R(\mathcal{I}, p(z))$ that involve the classified set $\mathcal{I}$ and the vector $p(z)$ of a posteriori probabilities in Eq. 5. For instance, a reward function

$$R(\mathcal{I}, p(z)) = 1\left(\sum_{i\in\mathcal{I}} p_i(z) \geq 1-\delta\right)/|\mathcal{I}| \tag{96}$$

is defined for all non-empty subsets $\mathcal{I}$ of $\mathcal{N}$. The objective of this reward function is to guarantee a lower bound $1-\delta$ on the conditional coverage probability $P(I \in \hat{I}|z)$, and subject to this constraint minimize the size $|\mathcal{I}|$ of the classified set. Chzhen et al. (2021) prove that the optimal Bayes classifier is $\hat{I}(z) = \hat{I}_{m(z)}$, with

$$m(z) = \min\{m;\ 1 \leq m \leq N, v(m; z) \geq 1-\delta\}.$$

Note that Eq. 96 can be viewed as a conditional version of the precision reward in Eq. 44.

7.3 Other Reward Functions for Models with Blocks of Categories

When the N categories can be divided into blocks, it is possible to consider reward functions with other additive penalty terms than those of Section 5.2. Instead of penalizing the number of classified categories within the correct and wrong blocks respectively, as in Eq. 65, one penalizes the number of *wrongly classified categories* within the correct and wrong blocks. This corresponds to a reward function

$$R(\mathcal{I}, i) = 1(i \in \mathcal{I}) - g_{k(i)}(|\mathcal{I}_{k(i),(-i)}|, |\mathcal{I}| - |\mathcal{I}_{k(i)}|, |\mathcal{I}|), \tag{97}$$

where $\mathcal{I}_{k(i),(-i)} = \mathcal{I}_{k(i)} \setminus \{i\}$ is the number of wrong categories of $\mathcal{I}$ that belong to block $k(i)$ when category i is true.

For instance, the reward function Eq. 82 for indifference zones (Example 8) was formulated in terms of a multiplicative reward function with penalty term Eq. 84. However, it can also be formulated in accordance with a reward function Eq. 97 with an additive penalty. In order to see this let m_1 and m_2 refer to the number of wrongly classified categories of blocks 1 and 2, whereas m is the size of the classified set (hence $m-(m_1+m_2)$ equals 1 or 0 depending on whether $i \in \mathcal{I}$ or not). Then, Eq. 82 corresponds to having penalty terms

$$\begin{aligned} g_1(m_1, m_2, m) &= 1(m = m_1 + m_2 + 1 \text{ and } (m_1, m_2) \neq (0, 0)), \\ g_2(0, m_1, m) &= 1(m = m_1 + 1) - r1(m = 1 \text{ and } m_1 = 0) \end{aligned}$$

in Eq. 97 when the true category belongs to blocks 1 and 2, respectively.

7.4 Multi-Label Classification

Suppose an observation z belongs to *several* categories $I \subset \mathcal{N}$, such as when I represents the properties associated with z. The task of predicting I from data is referred to as multi-label classification (Lewis, 1995; Tsoumakas & Katakis, 2007; Dembczyński et al., 2012; Zhang & Zhou, 2013; Nguyen & Hullermeier, 2020). It is appropriate, in the context of multi-label

classification, to represent a classified set $\mathcal{I} \subset \mathcal{N}$ as a binary vector $\boldsymbol{\iota} = (\iota_1, \ldots, \iota_N)$ of length N, where $\iota_j = 1$ if $j \in \mathcal{I}$ and $\iota_j = 0$ if $j \notin \mathcal{I}$. Likewise, we let $\boldsymbol{i} = (i_1, \ldots, i_N)$ represent an assumed value of I, with $i_j = 1$ if $j \in I$ and $i_j = 0$ if $j \notin I$. Denote by $R(\boldsymbol{\iota}, \boldsymbol{i})$ the reward for a classified set $\boldsymbol{\iota}$ when $\boldsymbol{i}$ is the true set of categories. The MAP classifier Eq. 8, for instance, corresponds, in the context of multi-label classification, to a reward function

$$R(\boldsymbol{\iota}, \boldsymbol{i}) = 1(\boldsymbol{\iota} = \boldsymbol{i}). \tag{98}$$

A frequently used reward function, which in contrast to Eq. 98 allows for some misclassified categories, is

$$R(\boldsymbol{\iota}, \boldsymbol{i}) = 1 - g(\boldsymbol{\iota}, \boldsymbol{i}) = 1 - \frac{|\boldsymbol{\iota} - \boldsymbol{i}|}{N} = 1 - \frac{1}{N} \sum_{j=1}^{N} |\iota_j - i_j|. \tag{99}$$

Its additive penalty $g(\boldsymbol{\iota}, \boldsymbol{i})$ is the normalized Hamming distance between $\boldsymbol{\iota}$ and $\boldsymbol{i}$. This is an instance of a decomposable loss/penalty, with $g(\boldsymbol{\iota}, \boldsymbol{i}) = \sum_j g_j(\iota_j, i_j)$ a sum of componentwise losses. It is known (Dembczyński et al., 2012) that such penalties make it tractable to compute the Bayes classifier

$$\hat{I}(z) = \arg\max_{\boldsymbol{\iota}} E[R(\boldsymbol{\iota}, I)|z].$$

The F-measure (Lewis, 1995)

$$R(\boldsymbol{\iota}, \boldsymbol{i}) = \frac{2 \sum_{j=1}^{N} \iota_j i_j}{\sum_{j=1}^{N} (\iota_j + i_j)} \tag{100}$$

simplifies to Eq. 43, with $a = 2$ and $b = 1$, when there is only one true category ($|\boldsymbol{i}| = 1$). As a complement to Eqs. 99 and 100, it would be of interest to consider reward functions that penalize type I errors ($\mathcal{I} \setminus I \neq \emptyset$) and type II errors ($I \setminus \mathcal{I} \neq \emptyset$) differently. For instance, if all categories belong to one homogeneous block and type II errors are regarded as more serious than type I errors, a natural extension of an additive penalty reward Eq. 13, to the multilabel classification context, is a function

$$R(\boldsymbol{\iota}, \boldsymbol{i}) = 1(\boldsymbol{i} \leq \boldsymbol{\iota}) - g(|\boldsymbol{\iota}|), \tag{101}$$

where the first term gives a unit reward if all true categories are included in the classifier (no type II errors), whereas the second term penalizes the size $|\mathcal{I}| = |\boldsymbol{\iota}|$ of the classified set $\mathcal{I}$. Analogously, it is possible to combine a reward for no type II errors with a multiplicative penalty. The resulting reward function

$$R(\boldsymbol{\iota}, \boldsymbol{i}) = 1(\boldsymbol{i} \leq \boldsymbol{\iota}) g(|\boldsymbol{\iota}|), \tag{102}$$

naturally extends Eq. 32 to a setting of multi-label classification. Another option is to generalize reward functions Eqs. 101 and 102 in order to include abstention to classify some categories (Nguyen & Hullermeier, 2020).

Appendix A: Optimizing $\tilde{R}_1$ and $\tilde{R}_2$

In Fig. 3, the non-reward rate $1 - R_{ab}^{\text{cv}}(\tilde{R})$ (cf. Eq. 90) is plotted as a function of the two cost parameters a and b of the classifier $\hat{I}_{(a,b)}$ for the two reward functions $\tilde{R}_1$ and $\tilde{R}_2$ of Table 2. The objective function is not smooth and thus it can be hard to optimize. However, we obtained

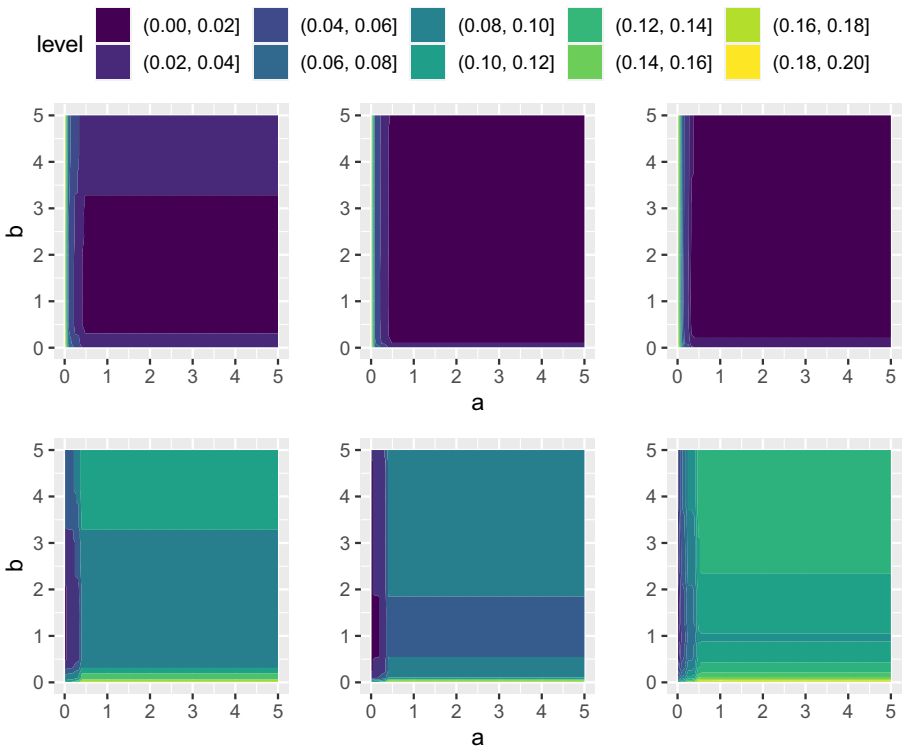


Fig. 3 The figure contains filled contour plots of the estimated non-reward rate $1 - R^{\text{cv}}_{ab}(\tilde{R})$ (cf. Eq. 90), as a function of the two cost parameters a and b of the classifier $\hat{I}_{(a,b)}$. These estimated non-reward rates make use of weights $w_i^{(\text{bird})}$ (cf. Eq. 93), the reward function $\tilde{R}_1$ (top row) and $\tilde{R}_2$ (bottom row). The columns, from left to right, correspond to the priors $\pi^{(\text{flat})}$, $\pi^{(\text{prop})}$ and $\pi^{(\text{real})}$ (cf. Section 6.3). The levels of the contour plots are crudely drawn, but it can still seen that $\tilde{R}_1$ attains a low non-reward rate for a large set of (a, b), whereas $\tilde{R}_2$ attains its lowest non-reward rates over a small region where a is close to 0

good results with Nelder-Mead optimization (Nelder & Mead, 1965), as implemented in the `optim`-function in R, with a starting value of (a, b) that corresponds to a small non-reward rate.

Acknowledgements The authors would like to thank Vilhelm Niklasson at Stockholm University for suggesting the topic of conformal prediction, and the Falsterbo Bird Observatory for providing the data set. This data will be made available upon reasonable request. We also thank two anonymous reviewers for very helpful comments that considerably improved the manuscript.

Funding Open access funding provided by Stockholm University.

References

Angelopoulos, A. N. & Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107:07511v6.

Arrow, K. J., Blackwell, D., & Girshick, M.A. (1949). Bayes and minimax solutions of sequential decision problems. *Econometrica, Journal of the Econometric Society*, 213–244.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 16–39.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bottai, M., Cai, B., & McKeown, R. E. (2010). Logistic quantile regression for bounded outcomes. *Statistical Medicine, 29*, 309–317.

Carlin, B.P., Kadane, J.B., & Gelfand, A.E. (1998). Approaches for optimal sequential decision analysis in clinical trials. *Biometrics*, 964–975.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory, 16*(1), 41–46.

Chzhen, E., Denis, C., Hebiri, M., & Lorieul, T. (2021). Set-valued classification-overview via a unified framework. arXiv:2102:12318v1.

del Coz, J. J., Díez, J., & Bahamonde, A. (2009). Learning nondeterministic classifiers. *Journal of Machine Learning Research, 10*, 2273–2293.

Dembczyński, K., Waegeman, W., Cheng, W., & Hullermeier, E. (2012). On label dependence and loss minimization in multi-label classification. *Machine Learning, 88*, 5–45.

Denis, C., & Hebiri, M. (2017). Confidence sets with expected sizes for multiclass classification. *Journal of Machine Learning Research, 18*, 1–28.

Freund, Y., Mansour, Y., & Schapire, R. E. (2004). Generalization bounds for averaged classifiers. *The Annals of Statistics, 32*(4), 1698–1722.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-2.

Goldsman, D. (1986). Tutorial on indifference-zone normal means ranking and selection procedures. In *Proceedings of the 18th conference on Winter simulation*, pp. 370–375.

Grycko, E. (1993). Classification with set-valued decision functions. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and classification, studies in classification, data analysis and knowledge organization* (pp. 218–224). Berlin, Heidelberg: Springer.

Ha, T. M. (1996). An optimum class-selective rejection fule for pattern recognition. In *Proceedings of the 13th International Conference on Pattern Recognition*, Volume 2, pp. 75–80. IEEE.

Ha, T. M. (1997). The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(6), 608–615.

Hastie, T., Tibshirani, R., & Friedman, J., et al. (2009). *The elements of statistical learning* (Second ed.). Number 10. Springer series in statistics New York.

Hellman, M. E. (1970). The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics, 6*(3), 179–185.

Herbei, R. & Wegkamp M.H. (2006). Classification with reject option. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, 709–721.

Karlsson, M., & Hössjer, O. (2023). Identification of taxon through classification with partial reject options. *Journal of the Royal Statistical Society, Series C, 72*(4), 937–975.

Koenker (2005). *Quantile Regression (Econometric Society monographs; no. 38)*. Cambridge University Press.

Le Capitaine, H. (2014). A unified view of class-selection with probabilistic classifiers. *Pattern Recognition, 47*, 843–853.

Levi, I. (1983). *The enterprise of knowledge: An assay on knowledge, credal probability, and chance*. MIT Press.

Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, pp. 246–254. ACM.

Ma, L., & Denoeux, T. (2021). Partial classification in the belief function framework. *Knowledge-Based Systems, 114*, 106742.

Malmhagen, B., Karlsson, M., & Menzie, S. (2013). Using wing morphology to separate four species of Acrocephalus warblers in Scandinavia. *Ringing & Migration, 28*(July), 63–68.

Mortier, T., Hullermeyer, E. Dembczyński, K., & Waegeman W. (2022). Set-valued prediction in hierarchical classification with constrained representation complexity. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). PLMR*, *Vol. 180*, pp 1392–1401.

Mortier, T., Wydmuch, M., Dembczyński, K., Hullermeier, E., & Waegeman, W. (2021). Efficient set-valued prediction in multi-class classification. *Data Mining and Knowledge Discovery, 35*(4), 1435–1469.
Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308–313.
Nguyen, V.-L. & E. Hullermeier (2020). Reliable multilabel classification: Prediction with partial abstention. In *Proceedings of the 34th AAAI Conference on Artificial Intelligene (AAAI-20)*, *vol. 34*, pp. 5264–5271.
R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge University Press.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Set-valued classification - overview via a unified framework. *International Journal of Computer Vision, 115*, 211–252.
Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association, 114*(525), 223–234.
Shafer, G. & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research 9*(3).
Svensson, L. (1992). *Identification guide to European passerines*.
Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining, 3*(3), 1–13.
Vovk, V., A. Gammerman, & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.
Vovk, V., Nouretdinov, I., Federova, V., Petej, I., & Gammerman, A. (2017). Criteria of efficiency for set-valued classification. *Annals of Mathematics and Artificial Intelligence, 81*, 21–47.
Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall.
Yang, G., Destercke, S., & Marie-Hélène, M. (2017). The cost of indeterminacy: How to determine them? *IEEE Transactions on Cybernetics, 47*(12), 4316–4327.
Zaffalon, M. (2002). The naive credal classifier. *Journal of Statistiical Planning and Inference, 105*(1), 5–21.
Zaffalon, M., Corani, G., & Mauá, D. (2012). Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning, 53*, 1282–1301.
Zhang, M.-L., & Zhou, Z.-H. (2013). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering, 26*, 1819–1837.