Reliable Multilabel Classification: Prediction with Partial Abstention

Vu-Linh Nguyen, Eyke Hüllermeier
Heinz Nixdorf Institute and Department of Computer Science, Paderborn University, Germany
vu.linh.nguyen@uni-paderborn.de, eyke@upb.de

Abstract

In contrast to conventional (single-label) classification, the setting of *multilabel classification* (MLC) allows an instance to belong to several classes simultaneously. Thus, instead of selecting a single class label, predictions take the form of a subset of all labels. In this paper, we study an extension of the setting of MLC, in which the learner is allowed to partially abstain from a prediction, that is, to deliver predictions on some but not necessarily all class labels. We propose a formalization of MLC with abstention in terms of a generalized loss minimization problem and present first results for the case of the Hamming loss, rank loss, and F-measure, both theoretical and experimental.

1 Introduction

In statistics and machine learning, classification with abstention, also known as classification with a reject option, is an extension of the standard setting of classification, in which the learner is allowed to refuse a prediction for a given query instance; research on this setting dates back to early work by Chow (1970) and Hellman (1970) and remains to be an important topic till today, most notably for binary classification (Bartlett and Wegkamp 2008; Cortes, DeSalvo, and Mohri 2016; Franc and Prusa 2019; Grandvalet et al. 2008). For the learner, the main reason to abstain is a lack of certainty about the corresponding outcome—refusing or at least deferring a decision might then be better than taking a high risk of a wrong decision.

Nowadays, there are many machine learning problems in which complex, structured predictions are sought (instead of scalar values, like in classification and regression). For such problems, the idea of abstaining from a prediction can be generalized toward *partial abstention*: Instead of predicting the entire structure, the learner predicts only parts of it, namely those for which it is certain enough. This idea has already been realized, e.g., for the case where predictions are rankings (Cheng et al. 2010; 2012).

Another important example is *multilabel classification* (MLC), in which an outcome associated with an instance is a labeling in the form of a subset of an underlying reference set of class labels (Dembczyński et al. 2012; Tsoumakas, Katakis, and Vlahavas 2009; Zhang and Zhou 2014). In this paper, we study an extension of the setting of MLC, in which the learner is allowed to partially abstain from a prediction, that is, to deliver predictions on some but not necessarily all class labels (or, more generally, to refuse committing to a single complete prediction). Although MLC has been studied extensively in the machine learning literature in the recent past, there is surprisingly little work on MLC with abstention so far—a notable exception is (Pillai, Fumera, and Roli 2013), to which we will return in the Section 7.

Prediction with abstention is typically realized as a two-stage approach. First, the learner delivers a prediction that provides information about its uncertainty. Then, taking this uncertainty into account, a decision is made about whether or not to predict, or on which parts. In binary classification, for example, a typical approach is to produce probabilistic predictions and to abstain whenever the probability is close to $1/2$. We adopt a similar approach, in which we rely on probabilistic MLC, i.e., probabilistic predictions of labelings.

In the next section, we briefly recall the setting of multilabel classification. The generalization toward MLC with (partial) abstention is then introduced and formalized in Section 3. Instantiations of the setting of MLC with abstention for the specific cases of the Hamming loss, rank loss, and F-measure are studied in Sections 4–6, respectively. Related work is discussed in Section 7. Finally, experimental results are presented in Section 8, prior to concluding the paper in Section 9. All formal results in this paper (propositions, remarks, corollaries) are stated without proofs, which are deferred to the supplementary material.

2 Multilabel Classification

In this section, we describe the MLC problem in more detail and formalize it within a probabilistic setting. Along the way, we introduce the notation used throughout the paper.

Let $\mathcal{X}$ denote an instance space, and let $\mathcal{L} = \{\lambda_1, \ldots, \lambda_m\}$ be a finite set of class labels. We assume that an instance $\boldsymbol{x} \in \mathcal{X}$ is (probabilistically) associated with a subset of labels $\Lambda = \Lambda(\boldsymbol{x}) \in 2^{\mathcal{L}}$; this subset is often called the set of relevant labels, while the complement $\mathcal{L} \setminus \Lambda$ is considered as irrelevant for $\boldsymbol{x}$. We identify a set Λ of relevant labels with a

binary vector $\boldsymbol{y} = (y_1, \ldots, y_m)$, where $y_i = [\![\lambda_i \in \Lambda]\!]$.[1] By $\mathcal{Y} = \{0, 1\}^m$ we denote the set of possible labelings.

We assume observations to be realizations of random variables generated independently and identically (i.i.d.) according to a probability distribution $\boldsymbol{p}$ on $\mathcal{X} \times \mathcal{Y}$, i.e., an observation $\boldsymbol{y} = (y_1, \ldots, y_m)$ is the realization of a corresponding random vector $\mathbf{Y} = (Y_1, \ldots, Y_m)$. We denote by $\boldsymbol{p}(\mathbf{Y} \mid \boldsymbol{x})$ the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \boldsymbol{x}$, and by $\boldsymbol{p}_i(Y_i \mid \boldsymbol{x})$ the corresponding marginal distribution of Y_i:

$$\boldsymbol{p}_i(b \mid \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y} : y_i = b} \boldsymbol{p}(\boldsymbol{y} \mid \boldsymbol{x}) \,. \tag{1}$$

A multilabel classifier $\mathbf{h}$ is a mapping $\mathcal{X} \longrightarrow \mathcal{Y}$ that assigns a (predicted) label subset to each instance $\boldsymbol{x} \in \mathcal{X}$. Thus, the output of a classifier $\mathbf{h}$ is a vector

$$\hat{\boldsymbol{y}} = \mathbf{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_m(\boldsymbol{x})) \,.$$

Given training data in the form of a finite set of observations $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$, drawn independently from $\Pr(\mathbf{X}, \mathbf{Y})$, the goal in MLC is to learn a classifier $\mathbf{h} : \mathcal{X} \longrightarrow \mathcal{Y}$ that generalizes well beyond these observations in the sense of minimizing the expected risk with respect to a specific loss function.

In the literature, various MLC loss functions have been proposed, including the Hamming loss, the subset 0/1 loss, the F-measure, the Jaccard measure, and the rank loss. The Hamming loss is given by

$$\ell_H(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{m} [\![y_i \neq \hat{y}_i]\!] \;, \tag{2}$$

and the subset 0/1 loss by $\ell_S(\boldsymbol{y}, \hat{\boldsymbol{y}}) = [\![\boldsymbol{y} \neq \hat{\boldsymbol{y}}]\!]$. Thus, both losses generalize the standard 0/1 loss commonly used in classification, but in a very different way. Hamming and subset 0/1 are prototypical examples of what is called a (label-wise) *decomposable* and *non-decomposable* loss, respectively (Dembczyński et al. 2012). A decomposable loss can be expressed in the form

$$\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \sum_{i=1}^{m} \ell_i(y_i, \hat{y}_i) \tag{3}$$

with suitable binary loss functions $\ell_i : \{0, 1\}^2 \longrightarrow \mathbb{R}$, whereas a non-decomposable loss does not permit such a representation. It can be shown that, to produce optimal predictions $\hat{\boldsymbol{y}} = \mathbf{h}(\boldsymbol{x})$ minimizing expected loss, knowledge about the marginals $\boldsymbol{p}_i(Y_i \mid \boldsymbol{x})$ is enough in the case of a decomposable loss, but not in the case of a non-decomposable loss (Dembczyński et al. 2012). Instead, if a loss is non-decomposable, high-order probabilities are needed, and in the extreme case even the entire distribution $\boldsymbol{p}(\mathbf{Y} \mid \boldsymbol{x})$ (like in the case of the subset 0/1 loss). On an algorithmic level, this means that MLC with a decomposable loss can be tackled by what is commonly called binary relevance (BR) learning (i.e., learning one binary classifier for each label individually), whereas non-decomposable losses call for more sophisticated learning methods that are able to take label-dependencies into account.

[1] $[\![\cdot]\!]$ is the indicator function, i.e., $[\![A]\!] = 1$ if the predicate A is true and $= 0$ otherwise.

3 MLC with Abstention

In our generalized setting of MLC with abstention, which is introduced in this section, the classifier is allowed to produce *partial predictions*

$$\hat{\boldsymbol{y}} = \mathbf{h}(\boldsymbol{x}) \in \mathcal{Y}_{pa} \coloneqq \{0, \bot, 1\}^m \,, \tag{4}$$

where $\hat{y}_i = \bot$ indicates an abstention on the label λ_i; we denote by $A(\hat{\boldsymbol{y}}) \subseteq [m] \coloneqq \{1, \ldots, m\}$ and $D(\hat{\boldsymbol{y}}) \coloneqq [m] \setminus A(\hat{\boldsymbol{y}})$ the set of indices i for which $\hat{y}_i = \bot$ and $\hat{y}_i \in \{0, 1\}$, respectively, that is, the indices on which the learner abstains and decides to predict.

3.1 Risk Minimization

To evaluate a reliable multilabel classifier, a generalized loss function

$$L : \mathcal{Y} \times \mathcal{Y}_{pa} \longrightarrow \mathbb{R}_+ \tag{5}$$

is needed, which compares a partial prediction $\hat{\boldsymbol{y}}$ with a ground-truth labeling $\boldsymbol{y}$. Given such a loss, and assuming a probabilistic prediction for a query instance $\boldsymbol{x}$, i.e., a probability $\boldsymbol{p}(\cdot \mid \boldsymbol{x})$ on the set of labelings (or at least an estimation thereof), the problem of risk minimization comes down to finding

$$\begin{aligned} \hat{\boldsymbol{y}} &\in \operatorname*{argmin}_{\hat{\boldsymbol{y}} \in \mathcal{Y}_{pa}} \mathbf{E}\big(L(\boldsymbol{y}, \hat{\boldsymbol{y}})\big) \\ &= \operatorname*{argmin}_{\hat{\boldsymbol{y}} \in \mathcal{Y}_{pa}} \sum_{\boldsymbol{y} \in \mathcal{Y}} L(\boldsymbol{y}, \hat{\boldsymbol{y}}) \cdot \boldsymbol{p}(\boldsymbol{y} \mid \boldsymbol{x}) \,. \end{aligned} \tag{6}$$

The concrete form of this optimization problem as well as its difficulty depend on several choices, including the underlying MLC loss function ℓ and its extension L.

3.2 Generalized Loss Functions

On the basis of a standard MLC loss ℓ, a generalized loss function (5) can be derived in different ways, also depending on how to penalize the abstention. Further below, we propose a generalization based on an additive penalty. Before doing so, we discuss some general properties that might be of interest for generalized losses.

As a first property, we should expect a generalized loss L to reduce to its conventional version ℓ in the case of no abstention. In other words,

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \ell(\boldsymbol{y}, \hat{\boldsymbol{y}}) \,,$$

whenever $\hat{\boldsymbol{y}}$ is a precise prediction $\hat{\boldsymbol{y}} \in \mathcal{Y}$. Needless to say, this is a property that every generalized loss should obey.

Monotonicity. Another reasonable property is *monotonicity*: The loss should only increase (or at least not decrease) when (i) turning a correct prediction on a label λ_i into an abstention or an incorrect prediction, (ii) or turning an abstention into an incorrect prediction. This reflects the following chain of preferences: a correct prediction is better than an abstention, which in turn is better than an incorrect prediction. More formally, for a ground-truth labeling $\boldsymbol{y}$ and a partial prediction $\hat{\boldsymbol{y}}_1$, let $C_1, A_1 \subseteq \mathcal{L}$ denote the subset of labels on which the prediction is correct and on which the learner

abstains, respectively, and define $C_2, A_2 \subseteq \mathcal{L}$ analogously for a prediction $\hat{\boldsymbol{y}}_2$. Then

$$(C_2 \subseteq C_1) \ \wedge \ \Big((C_2 \cup A_2) \subseteq (C_1 \cup A_1)\Big) \qquad (7)$$
$$\Rightarrow L(\boldsymbol{y}, \hat{\boldsymbol{y}}_1) \leq L(\boldsymbol{y}, \hat{\boldsymbol{y}}_2)\,.$$

Uncertainty-alignment. Intuitively, when producing a partial prediction, an optimal prediction rule is supposed to abstain on the most uncertain labels. More formally, consider a generalized loss function L and a prediction $\hat{\boldsymbol{y}}$ which, for a query $\boldsymbol{x} \in \mathcal{X}$, is a risk-minimizer (6). Moreover, denoting by $p_i = \boldsymbol{p}_i(1 \,|\, \boldsymbol{x})$ the (marginal) probability that label λ_i is relevant for $\boldsymbol{x}$, it is natural to quantify the degree of uncertainty on this label in terms of

$$u_i = 1 - 2|p_i - 1/2| = 2\min(p_i, 1 - p_i)\,, \qquad (8)$$

or any other function symmetric around $1/2$. We say that $\hat{\boldsymbol{y}}$ is *uncertainty-aligned* if

$$\forall\, y_i \in A(\hat{\boldsymbol{y}}), y_j \in D(\hat{\boldsymbol{y}}) :\ u_i \geq u_j\,.$$

Thus, a prediction is uncertainty-aligned if the following holds: Whenever the learner decides to abstain on label λ_i and to predict on label λ_j, the uncertainty on λ_j cannot exceed the uncertainty on λ_i. We then call a loss function L uncertainty-aligned if it guarantees the existence of an uncertainty-aligned risk-minimizer, regardless of the probability $\boldsymbol{p} = \boldsymbol{p}(\cdot \,|\, \boldsymbol{x})$.

Additive Penalty for Abstention Consider the case of a partial prediction $\hat{\boldsymbol{y}}$ and denote by $\hat{\boldsymbol{y}}_D$ and $\hat{\boldsymbol{y}}_A$ the projections of $\hat{\boldsymbol{y}}$ to the entries in $D(\hat{\boldsymbol{y}})$ and $A(\hat{\boldsymbol{y}})$, respectively. As a natural extension of the original loss ℓ, we propose a generalized loss of the form

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \ell(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D) + f(A(\hat{\boldsymbol{y}}))\,, \qquad (9)$$

with $\ell(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D)$ the original loss on that part on which the learner predicts and $f(A(\hat{\boldsymbol{y}}))$ a penalty for abstaining on $A(\hat{\boldsymbol{y}})$. The latter can be seen as a measure of the loss of usefulness of the prediction $\hat{\boldsymbol{y}}$ due to its partiality, i.e., due to having no predictions on $A(\hat{\boldsymbol{y}})$.

An important instantiation of (9) is the case where the penalty is a counting measure, i.e., where f only depends on the number of abstentions:

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \ell(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D) + f\big(|A(\hat{\boldsymbol{y}})|\big)\,. \qquad (10)$$

A special case of (10) is to penalize each abstention $\hat{y}_i = \perp$ with the same constant $c \in [0, 1]$, which yields

$$L(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \ell(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D) + |A(\hat{\boldsymbol{y}})| \cdot c\,. \qquad (11)$$

Of course, instead of a linear function f, more general penalty functions are conceivable. For example, a practically relevant penalty is a concave function of the number of abstentions: Each additional abstention causes a cost, so f is monotone increasing in $|A(\hat{\boldsymbol{y}})|$, but the marginal cost of abstention is decreasing.

Proposition 1. *Let the loss ℓ be decomposable in the sense of* (3)*, and let $\hat{\boldsymbol{y}}$ be a risk-minimizing prediction (for a query instance $\boldsymbol{x}$). The minimization of the expected loss* (10) *is then accomplished by*

$$\hat{\boldsymbol{y}} = \operatorname*{argmin}_{1 \leq d \leq m} \mathbf{E}\left(\ell(\boldsymbol{y}, \hat{\boldsymbol{y}}_d)\right) + f(m - d)\,, \qquad (12)$$

where the prediction $\hat{\boldsymbol{y}}_d$ is specified by the index set

$$D_d(\hat{\boldsymbol{y}}_d) := \{\pi(1), \ldots, \pi(d)\}\,, \qquad (13)$$

and the permutation π sorts the labels in increasing order of the label-wise expected losses

$$s_i = \min_{\hat{y}_i \in \{0,1\}} \mathbf{E}(\ell_i(y_i, \hat{y}_i))\,,$$

i.e., $s_{\pi(1)} \leq \cdots \leq s_{\pi(m)}$.

As shown by the previous proposition, a risk-minimizing prediction for a decomposable loss can easily be found in time $O(m \log(m))$, simply by sorting the labels according to their contribution to the expected loss, and then finding the optimal size d of the prediction according to (12).

4 The Case of Hamming Loss

This section presents first results for the case of the Hamming loss function (2). In particular, we analyze extensions of the Hamming loss according to (10) and address the corresponding problem of risk minimization.

Given a query instance $\boldsymbol{x}$, assume conditional probabilities $p_i = \boldsymbol{p}(y_i = 1 \,|\, \boldsymbol{x})$ are given or made available by an MLC predictor $\mathbf{h}$. In the case of Hamming, the expected loss of a prediction $\hat{\boldsymbol{y}}$ is then given by

$$\mathbf{E}\left(\ell_H(\boldsymbol{y}, \hat{\boldsymbol{y}})\right) = \sum_{i:\hat{y}_i=1} 1 - p_i + \sum_{i:\hat{y}_i=0} p_i$$

and minimized by $\hat{\boldsymbol{y}}$ such that $\hat{y}_i = 0$ if $p_i \leq 1/2$ and $\hat{y}_i = 1$ otherwise.

In the setting of abstention, we call a prediction $\hat{\boldsymbol{y}}$ a d-prediction if $|D(\hat{\boldsymbol{y}})| = d$. Let π be a permutation of $[m]$ that sorts labels according to the uncertainty degrees (8), i.e., such that $u_{\pi(1)} \leq u_{\pi(2)} \leq \cdots \leq u_{\pi(m)}$. As a consequence of Proposition 1, we obtain the following result.

Corollary 1. *In the case of Hamming loss, let $\hat{\boldsymbol{y}}$ be a risk-minimizing prediction (for a query instance $\boldsymbol{x}$). The minimization of the expected loss* (10) *is then accomplished by*

$$\hat{\boldsymbol{y}} = \operatorname*{argmin}_{1 \leq d \leq m} \mathbf{E}\left(\ell_H(\boldsymbol{y}, \hat{\boldsymbol{y}}_d)\right) + f(m - d)\,, \qquad (14)$$

where the prediction $\hat{\boldsymbol{y}}_d$ is specified by the index set

$$D_d(\hat{\boldsymbol{y}}_d) = \{\pi(1), \ldots, \pi(d)\}\,. \qquad (15)$$

Corollary 2. *The extension* (10) *of the Hamming loss is uncertainty-aligned. In the case of the extension* (11) *of the Hamming loss, the optimal prediction is given by* (15) *with*

$$d = |\{i \in [m] \mid \min(p_i, 1 - p_i) \leq c\}|\,.$$

Remark 1. *The extension* (10) *of the Hamming loss is monotonic, provided f is non-decreasing and such that $f(k + 1) - f(k) \leq 1$ for all $k \in [m - 1]$.*

5 The Case of Rank Loss

In the case of the rank loss, we assume predictions in the form of rankings instead of labelings. Ignoring the possibility of ties, such a ranking can be represented in the form of a permutation π of $[m]$, where $\pi(i)$ is the index j of the label λ_j on position i (and $\pi^{-1}(j)$ the position of label λ_j). The rank loss then counts the number of incorrectly ordered label-pairs, that is, the number of pairs λ_i, λ_j such that λ_i is ranked worse than λ_j although λ_i is relevant while λ_j is irrelevant:

$$\ell_R(\boldsymbol{y}, \pi) = \sum_{(i,j):y_i > y_j} [\![\pi^{-1}(i) > \pi^{-1}(j)]\!] \ ,$$

or equivalently,

$$\ell_R(\boldsymbol{y}, \pi) = \sum_{1 \leq i < j \leq m} [\![y_{\pi(i)} = 0 \wedge y_{\pi(j)} = 1]\!] \, . \tag{16}$$

Thus, given that the ground-truth labeling is distributed according to the probability $\boldsymbol{p}(\cdot \,|\, \boldsymbol{x})$, the expected loss of a ranking π is

$$\mathbf{E}(\pi) := \mathbf{E}\left(\ell_R(\boldsymbol{y}, \pi)\right) = \sum_{1 \leq i < j \leq m} \boldsymbol{p}_{\pi(i),\pi(j)}(0, 1 \,|\, \boldsymbol{x}) \, , \tag{17}$$

where $\boldsymbol{p}_{u,v}$ is the pairwise marginal

$$\boldsymbol{p}_{u,v}(a, b \,|\, \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y} : y_u = a, y_v = b} \boldsymbol{p}(\boldsymbol{y} \,|\, \boldsymbol{x}) \, . \tag{18}$$

In the following, we first recall the risk-minimizer for the rank loss as introduced above and then generalize it to the case of partial predictions. To simplify notation, we omit the dependence of probabilities on $\boldsymbol{x}$ (for example, we write $\boldsymbol{p}_{u,v}(a, b)$ instead of $\boldsymbol{p}_{u,v}(a, b \,|\, \boldsymbol{x})$), and write (i) as indices of permuted labels instead of $\pi(i)$. We also use the following notation: For a labeling $\boldsymbol{y}$, let $r(\boldsymbol{y}) = \sum_{i=1}^{m} y_i$ be the number of relevant labels, and $c(\boldsymbol{y}) = r(\boldsymbol{y})(m - r(\boldsymbol{y}))$ the number of relevant/irrelevant label pairs (and hence an upper bound on the rank loss).

A risk-minimizing ranking π, i.e., a ranking minimizing (17), is provably obtained by sorting the labels λ_i in decreasing order of the probabilities $p_i = \boldsymbol{p}_i(1 \,|\, \boldsymbol{x})$, i.e., according to their probability of being relevant (Dembczyński et al. 2012). Thus, an optimal prediction π is such that

$$p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(m)} \, . \tag{19}$$

To show this result, let $\bar{\pi}$ denote the reversal of π, i.e., the ranking that reverses the order of the labels. Then, for each pair (i, j) such that $y_i > y_j$, either π or $\bar{\pi}$ incurs an error, but not both. Therefore, $\ell_R(\boldsymbol{y}, \pi) + \ell_R(\boldsymbol{y}, \bar{\pi}) = c(\boldsymbol{y})$, and

$$\ell_R(\boldsymbol{y}, \pi) - \ell_R(\boldsymbol{y}, \bar{\pi}) = 2\ell_R(\boldsymbol{y}, \pi) - c(\boldsymbol{y}) \, . \tag{20}$$

Since $c(\boldsymbol{y})$ is a constant that does not depend on π, minimizing $\ell_R(\boldsymbol{y}, \pi)$ (in expectation) is equivalent to minimizing the difference $\ell_R(\boldsymbol{y}, \pi) - \ell_R(\boldsymbol{y}, \bar{\pi})$. For the latter, the expectation (17) becomes

$$\begin{aligned} \mathbf{E}'(\pi) &= \sum_{1 \leq i < j \leq m} \boldsymbol{p}_{(i),(j)}(0, 1) - \boldsymbol{p}_{(i),(j)}(1, 0) \qquad (21) \\ &= \sum_{1 \leq i < j \leq m} p_{(j)} - p_{(i)} \\ &= \sum_{1 \leq i \leq m} (2i - (m+1)) p_{(i)} \, , \end{aligned}$$

where the transition from the first to the second sum is valid because (Dembczyński, Cheng, and Hüllermeier 2010)

$$\begin{aligned} &\boldsymbol{p}_{u,v}(0, 1) - \boldsymbol{p}_{u,v}(1, 0) \\ &= \boldsymbol{p}_{u,v}(0, 1) + \boldsymbol{p}_{u,v}(1, 1) - \boldsymbol{p}_{u,v}(1, 1) - \boldsymbol{p}_{u,v}(1, 0) \\ &= \boldsymbol{p}_v(1) - \boldsymbol{p}_u(1) = p_v - p_u \, . \end{aligned}$$

From (21), it is clear that a risk-minimizing ranking π is defined by (19).

To generalize this result, let us look at the rank loss of a partial prediction of size $d \in [m]$, i.e., a ranking of a subset of d labels. To simplify notation, we identify such a prediction, not with the original set of indices of the labels, but the positions of the corresponding labels in the sorting (19). Thus, a partial prediction of size d is identified by a set of indices $K = \{k_1, \ldots, k_d\}$ such that $k_1 < k_2 < \cdots < k_d$, where $k \in K$ means that the label $\lambda_{(k)}$ with the kth largest probability $p_{(k)}$ in (19) is included. According to the above result, the optimal ranking π_K on these labels is the identity, and the expected loss of this ranking is given by

$$\mathbf{E}(\pi_K) = \sum_{1 \leq i < j \leq d} \boldsymbol{p}_{(k_i),(k_j)}(0, 1) \, . \tag{22}$$

Lemma 1. *Assuming (conditional) independence of label probabilities in the sense that $\boldsymbol{p}_{i,j}(y_i, y_j) = \boldsymbol{p}_i(y_i)\boldsymbol{p}_j(y_j)$, the generalized loss (10) is minimized in expectation by a partial prediction with decision set of the form*

$$K_d = \langle\!\langle a, b \rangle\!\rangle := \{1, \ldots, a\} \cup \{b, \ldots, m\} \, , \tag{23}$$

with $1 \leq a < b \leq m$ and $m + a - b + 1 = d$.

According to the previous lemma, an optimal d-selection K_d leading to an optimal (partial) ranking of length d is always a "boundary set" of positions in the ranking (19). The next lemma establishes an important relationship between optimal selections of increasing length.

Lemma 2. *Let $K_d = \langle\!\langle a, b \rangle\!\rangle$ be an optimal d-selection (23) for $d \geq 2$. At least one of the extensions $\langle\!\langle a+1, b \rangle\!\rangle$ or $\langle\!\langle a, b-1 \rangle\!\rangle$ of K_d is an optimal $(d+1)$-selection.*

Thanks to the previous lemma, a risk-minimizing partial ranking can be constructed quite easily (in time $O(m \log(m))$. First, the labels are sorted according to (19). Then, an optimal decision set is produced by starting with the boundary set $\langle\!\langle 1, m \rangle\!\rangle$ and increasing this set in a greedy manner (a concrete algorithm is given in the supplementary material).

Proposition 2. *Given a query instance $\boldsymbol{x}$, assume conditional probabilities $\boldsymbol{p}(y_i = 1 \,|\, \boldsymbol{x}) = h_i(\boldsymbol{x})$ are made available by an MLC predictor $\mathbf{h}$. A risk-minimizing partial ranking can be constructed in time $O(m \log(m))$.*

Remark 2. *The extension* (10) *of the rank loss is not uncertainty-aligned.*

Since a prediction is a (partial) ranking instead of a (partial) labeling, the property of monotonicity as defined in Section 3.2 does not apply in the case of rank loss. Although it would be possible to generalize this property, for example by looking at (in)correctly sorted label pairs instead of (in)correct labels, we refrain from a closer analysis here.

6 The Case of F-measure

The F-measure is the harmonic mean of precision and recall and can be expressed as follows:

$$F(\boldsymbol{y}, \hat{\boldsymbol{y}}) := \frac{2\sum_{i=1}^{m} y_i \hat{y}_i}{\sum_{i=1}^{m}(y_i + \hat{y}_i)} \,. \tag{24}$$

The problem of finding the expected F-maximizer

$$\begin{aligned} \hat{\boldsymbol{y}} &= \arg\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}} \mathbf{E}\left(F(\boldsymbol{y}, \hat{\boldsymbol{y}})\right) \\ &= \arg\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}} \sum_{\boldsymbol{y} \in \mathcal{Y}} F(\boldsymbol{y}, \hat{\boldsymbol{y}}) \cdot \boldsymbol{p}(\boldsymbol{y} \,|\, \boldsymbol{x}) \end{aligned} \tag{25}$$

has been studied quite extensively in the literature (Chai 2005; Dembczyński et al. 2011; Decubber et al. 2018; Jansche 2007; Lewis 1995; Quevedo, Luaces, and Bahamonde 2012; Waegeman et al. 2014; Ye et al. 2012). Obviously, the optimization problem (25) can be decomposed into an inner and an outer maximization as follows:

$$\hat{\boldsymbol{y}}^k := \arg\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}_k} \mathbf{E}\left(F(\boldsymbol{y}, \hat{\boldsymbol{y}})\right) \,, \tag{26}$$

$$\hat{\boldsymbol{y}} := \arg\max_{k \in \{0,\ldots,m\}} \mathbf{E}\left(F(\boldsymbol{y}, \hat{\boldsymbol{y}}^k)\right) \,, \tag{27}$$

where $\mathcal{Y}_k := \{\hat{\boldsymbol{y}} \in \mathcal{Y} | \sum_{i=1}^{m} \hat{y}_i = k\}$ denotes the set of all predictions with exactly k positive labels.

Lewis (1995) showed that, under the assumption of conditional independence, the F-maximizer has always a specific form: it predicts the k labels with the highest marginal probabilities p_i as relevant, and the other $m - k$ labels as irrelevant. More specifically, for any number $k = 0, \ldots, m$, the solution of the optimization problem (26), namely a k-optimal solution $\hat{\boldsymbol{y}}^k$, is obtained by setting $\hat{y}_i = 1$ for the k labels with the highest marginal probabilities p_i, and $\hat{y}_i = 0$ for the remaining ones. Thus, the F-maximizer (27) can be found as follows:

- The labels λ_i are sorted in decreasing order of their (predicted) probabilities p_i.
- For every $k \in \{0, \ldots, m\}$, the optimal prediction $\hat{\boldsymbol{y}}^k$ is defined as described above.
- For each of these $\hat{\boldsymbol{y}}^k$, the expected F-measure is computed.
- As an F-measure maximizer $\hat{\boldsymbol{y}}$, the k-optimal prediction $\hat{\boldsymbol{y}}^k$ with the highest expected F-measure is adopted.

Overall, the computation of $\hat{\boldsymbol{y}}$ can be done in time $O(m^2)$ (Decubber et al. 2018; Ye et al. 2012).

To define the generalization of the F-measure, we first turn it into the loss function $\ell_F(\boldsymbol{y}, \hat{\boldsymbol{y}}) := 1 - F(\boldsymbol{y}, \hat{\boldsymbol{y}})$. The generalized loss is then given by

$$L_F(\boldsymbol{y}, \hat{\boldsymbol{y}}) := 1 - F(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D) + f(|A(\hat{\boldsymbol{y}})|).$$

Minimizing the expectation of this loss is obviously equivalent to maximizing the following generalized F-measure in expectation:

$$F_G(\boldsymbol{y}, \hat{\boldsymbol{y}}) := \begin{cases} 1 - f(a) & \text{if } a = m \,, \\ \frac{2\sum_{i \in D(\hat{\boldsymbol{y}})} y_i \hat{y}_i}{\sum_{i \in D(\hat{\boldsymbol{y}})}(y_i + \hat{y}_i)} - f(a) & \text{otherwise,} \end{cases} \tag{28}$$

where $a := |A(\hat{\boldsymbol{y}})|$.

Remark 3. *If f in* (28) *is a strictly increasing function, then*

- *turning an incorrect prediction or an abstention on a label λ_i into a correct prediction increases the generalized F-measure, whereas*
- *turning an incorrect prediction into an abstention may decrease the measure.*

Therefore, the generalized F-measure (28) *is not monotonic.*

The F-maximizer $\hat{\boldsymbol{y}}$ of the generalized F-measure (28) is given by

$$\begin{aligned} \hat{\boldsymbol{y}} &= \arg\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}_{pa}} \mathbf{E}\left(F_G(\boldsymbol{y}, \hat{\boldsymbol{y}})\right) \\ &= \arg\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}_{pa}} \sum_{\boldsymbol{y} \in \mathcal{Y}} F(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D) \cdot \boldsymbol{p}(\boldsymbol{y} \,|\, \boldsymbol{x}) - f(a) \,. \end{aligned} \tag{29}$$

In the following, we show that the F-maximizer of the generalized F-measure (28) can be found in the time $O(m^3)$. For any $k = 0, \ldots, m$, denote by

$$\mathcal{Y}_{pa}^k := \left\{ \hat{\boldsymbol{y}} \in \mathcal{Y}_{pa} \,\middle|\, \sum_{i \in D(\hat{\boldsymbol{y}})} \hat{y}_i = k \right\} . \tag{30}$$

The optimization problem (29) is decomposed into an inner and an outer maximization as follows:

$$\hat{\boldsymbol{y}}^k := \arg\max_{\hat{\boldsymbol{y}} \in \mathcal{Y}_{pa}^k} \mathbf{E}\left(F_G(\boldsymbol{y}, \hat{\boldsymbol{y}})\right) \,, \tag{31}$$

$$\hat{\boldsymbol{y}} := \arg\max_{\hat{\boldsymbol{y}} \in \{\hat{\boldsymbol{y}}^k | k = 0, \ldots, m\}} \mathbf{E}\left(F_G(\boldsymbol{y}, \hat{\boldsymbol{y}}^k)\right) \,. \tag{32}$$

Lemma 3. *For any partial prediction $\hat{\boldsymbol{y}} \in \mathcal{Y}_{pa}$ and any index $j \in D(\hat{\boldsymbol{y}})$,*

- $\mathbf{E}\left(F_G(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D)\right)$ *is an increasing function of p_j if $\hat{y}_j = 1$;*
- $\mathbf{E}\left(F_G(\boldsymbol{y}_D, \hat{\boldsymbol{y}}_D)\right)$ *is a decreasing function of p_j if $\hat{y}_j = 0$.*

Lemma 4. *Let π be the permutation that sorts the labels in decreasing order of the marginal probability $\boldsymbol{p}_i(y_i \,|\, \boldsymbol{x})$ defined in* (19)*. Assuming (conditional) independence of label probabilities in the sense that*

$$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{m} p_i^{y_i} (1 - p_i)^{1 - y_i} \,, \tag{33}$$

the generalized F-measure (28) is maximized in expectation by an optimal k-prediction $\hat{\boldsymbol{y}}^k$ with decision set of the form

$$D(\hat{\boldsymbol{y}}^k) = \langle\!\langle k, l \rangle\!\rangle := \{1, \ldots, k\} \cup \{l, \ldots, m\} \,, \tag{34}$$

with some $l \geq k + 1$ and

$$\hat{y}_{(i)} = \begin{cases} 1 & \text{if } i \in \{1, \ldots, k\} \,, \\ 0 & \text{if } i \in \{l, \ldots, m\} \,. \end{cases} \tag{35}$$

Thanks to the previous lemma, a maximizer $\hat{\boldsymbol{y}}$ of the generalized F-measure (28) can be constructed following a procedure similar to the case of the rank loss. First, the labels are sorted according to (19). Then, we evaluate all possible partial predictions with decision sets $\langle\!\langle k, l \rangle\!\rangle$ of the form (34), and find the one with the highest expected F-measure (28) (a concrete algorithm is given in the supplementary material).

Proposition 3. *Given a query instance $\boldsymbol{x}$, assume conditional probabilities $p_i = \boldsymbol{p}(y_i = 1 \,|\, \boldsymbol{x}) = h_i(\boldsymbol{x})$ are made available by an MLC predictor* $\mathbf{h}$*. Assuming (conditional) independence of label probabilities in the sense of* (33)*, a prediction $\hat{\boldsymbol{y}}$ maximizing the generalized F-measure* (28) *in expectation is constructed in time $O(m^3)$.*

7 Related Work

In spite of extensive research on multilabel classification in the recent past, there is surprisingly little work on abstention in MLC. A notable exception is an approach by Pillai, Fumera, and Roli (2013), who focus on the F-measure as a performance metric. They tackle the problem of maximizing the F-measure on a subset of label predictions, subject to the constraint that the effort for manually providing the remaining labels (those on which the learner abstains) does not exceed a pre-defined value. The decision whether or not to abstain on a label is guided by two thresholds on the predicted degree of relevance, which are tuned in a suitable manner. Even though this is an interesting approach, it is arguably less principled than ours, in which optimal predictions are derived in a systematic way, based on decision-theoretic ideas and the notion of Bayes-optimality. Besides, Pillai, Fumera, and Roli (2013) offer a solution for a specific setting but not a general framework for MLC with partial abstention.

More indirectly related is the work by Park and Simoff (2015), who investigate the uncertainty in multilabel classification. They propose a modification of the entropy measure to quantify the uncertainty of an MLC prediction. Moreover, they show that this measure correlates with the accuracy of the prediction, and conclude that it could be used as a measure of acceptance (and hence rejection) of a prediction. While Park and Simoff (2015) focus on the uncertainty of a complete labeling $\boldsymbol{y}$, Destercke (2015) and Antonucci and Corani (2017) quantify the uncertainty in individual predictions y_i using imprecise probabilities and so-called credal classifiers, respectively. Again, corresponding estimates can be used for the purpose of producing more informed decisions, including partial predictions.

8 Experiments

In this section, we present an empirical analysis that is meant to show the effectiveness of our approach to prediction with abstention. To this end, we perform experiments on a set of standard benchmark data sets from the MULAN repository[2] (cf. Table 1), following a 10-fold cross-validation procedure.

[2]http://mulan.sourceforge.net/datasets.html

Table 1: Data sets used in the experiments

#	name	# inst.	# feat.	# lab.
1	cal500	502	68	174
2	emotions	593	72	6
3	scene	2407	294	6
4	yeast	2417	103	14
5	mediamill	43907	120	101
6	nus-wide	269648	128	81

8.1 Experimental Setting

For training an MLC classifier, we use binary relevance (BR) learning with logistic regression (LR) as base learner (in its default setting in sklearn, i.e., with regularisation parameter set to 1)[3]. Of course, more sophisticated techniques could be applied, and results using classifier chains are given in the supplementary material. However, since we are dealing with decomposable losses, BR is well justified. Besides, we are first of all interested in analyzing the effectiveness of abstention, and less in maximizing overall performance. All competitors essentially only differ in how the conditional probabilities provided by LR are turned into a (partial) MLC prediction.

We first compare the performance of reliable classifiers to the conventional BR classifier that makes full predictions (MLC) as well as the cost of full abstention (ABS)—these two serve as baselines that MLC with abstention should be able to improve on. A classifier is obtained as a risk-minimizer of the extension (10) of Hamming loss (2), instantiated by the penalty function f and the constant c. We conduct a first series of experiments (SEP) with linear penalty $f_1(a) = a \cdot c$, where $c \in [0.05, 0.5]$, and a second series (PAR) with concave penalty $f_2(a) = (a \cdot m \cdot c)/(m + a)$, varying $c \in [0.1, 1]$. The performance of a classifier is evaluated in terms of the average loss. Besides, we also compute the average abstention size $|A(\hat{\boldsymbol{y}})|/m$.

The same type of experiment is conducted for the rank loss (with MLC and ABS denoting full prediction and full abstention, respectively). A predicted ranking is a risk-minimizer of the extension (10) instantiated by the penalty function f and the constant c. We conduct a first series of experiments (SEP) with f_1 as above and $c \in [0.1, 1]$, and a second series (PAR) with f_2 as above and $c \in [0.2, 2]$.

8.2 Results

The results (illustrated in Figures 1 and 2 for three data sets–results for the other data sets are similar and can be found in the supplementary material) clearly confirm our expectations. The Hamming loss (cf. Figure 1) under partial abstention is often much lower than the loss under full prediction and full abstention, showing the effectiveness of the approach. When the cost c increases, the loss increases while the abstention size decreases, with a convergence of the performance of SEP and PAR to the one of MLC at $c = 0.5$

[3]For an implementation in Python, see http://scikit.ml/api/skmultilearn.html

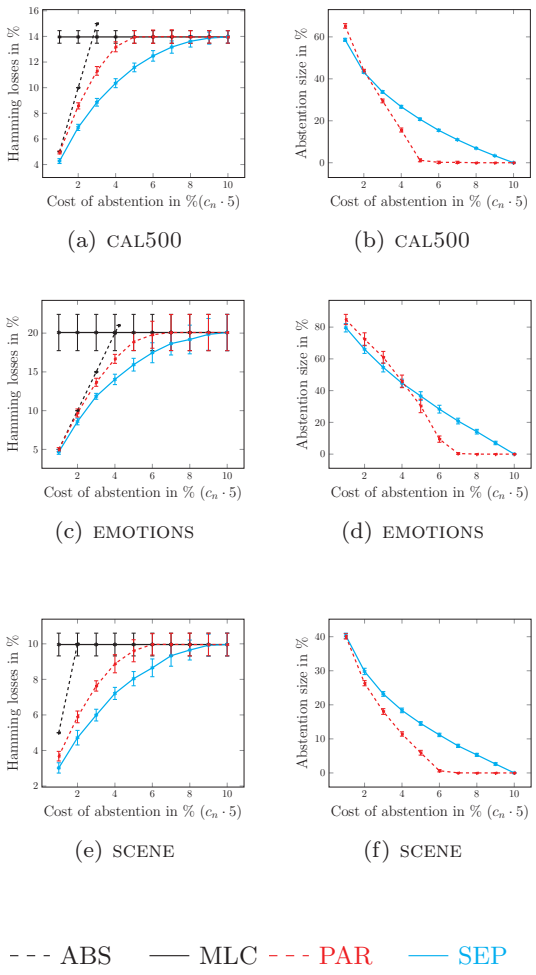

Figure 1: Binary relevance with logistic regression: Experimental results in terms of expected Hamming loss $(L_H \cdot 100)/m$ and abstention size (in percent) for $f_1(a) = a \cdot c$ (SEP) and $f_2(a) = (a \cdot m \cdot c)/(m+a)$ (PAR), as a function of the cost of abstention.

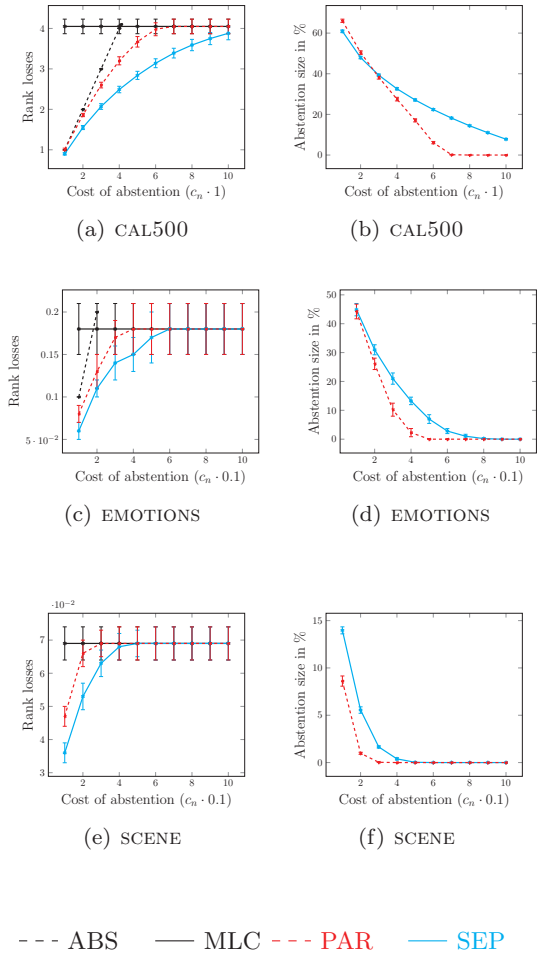

Figure 2: Binary relevance with logistic regression: Experimental results in terms of expected rank loss L_R/m and abstention size (in percent) for $f_1(a) = a \cdot c$ (SEP) and $f_2(a) = (a \cdot m \cdot c)/(m+a)$ (PAR), as a function of the cost of abstention.

and $c = 1$, respectively.

Similar results are obtained in the case of rank loss (cf. Figure 2), except that convergence to the performance of MLC is slower (i.e., requires lager cost values c, especially on the data set cal500). This is plausible, because the cost of a wrong prediction on a single label can be as high as $m-1$, compared to only 1 in the case of Hamming.

Due to space restrictions, we transferred experimental results for the generalized F-measure to the supplementary material. These results are very similar to those for the rank loss. In light of the observation that the respective risk-minimizers have the same structure, this is not very surprising.

The supplementary material also contains results for other MLC algorithms, including BR with support vector machines (using Platt-scaling (Lin, Lin, and Weng 2007; Platt 1999) to turn scores into probabilities) as base learners and classifier chains (Read et al. 2009) with LR and SVMs as base learners. Again, the results are very similar to those presented above.

9 Conclusion

This paper presents a formal framework of MLC with partial abstention, which builds on two main building blocks: First, the extension of an underlying MLC loss function so as to accommodate abstention in a proper way, and second the problem of optimal prediction, that is, minimizing this loss in expectation.

We instantiated our framework for the Hamming loss,

the rank loss, and the F-measure, which are three important and commonly used loss functions in multi-label classification. We elaborated on properties of risk-minimizers, showed them to have a specific structure, and devised efficient methods to produce optimal predictions. Experimentally, we showed these methods to be effective in the sense of reducing loss when being allowed to abstain.

In future work, we will further elaborate on our formal framework. As a concrete next step, we plan to investigate instantiations for other loss functions commonly used in MLC and the cases of label dependence (Dembczyński et al. 2012; Waegeman et al. 2014).

Acknowledgments

We thank Willem Waegeman for interesting discussions and useful suggestions. This work was supported by the Germany Research Foundation (DFG) (grant numbers HU 1284/19).

References

Antonucci, A., and Corani, G. 2017. The multilabel naive credal classifier. *International Journal of Approximate Reasoning* 83:320–336.

Bartlett, P. L., and Wegkamp, M. H. 2008. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* 9(Aug):1823–1840.

Chai, K. M. A. 2005. Expectation of f-measures: Tractable exact computation and some empirical observations of its properties. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR)*, 593–594. ACM.

Cheng, W.; Rademaker, M.; De Baets, B.; and Hüllermeier, E. 2010. Predicting partial orders: ranking with abstention. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part I (ECML/PKDD)*, 215–230. Springer-Verlag.

Cheng, W.; Hüllermeier, E.; Waegeman, W.; and Welker, V. 2012. Label ranking with partial abstention based on thresholded probabilistic models. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2501–2509.

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

Cortes, C.; DeSalvo, G.; and Mohri, M. 2016. Learning with rejection. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT)*, 67–82. Springer Verlag.

Decubber, S.; Mortier, T.; Dembczyński, K.; and Waegeman, W. 2018. Deep f-measure maximization in multi-label classification: A comparative study. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, 290–305. Springer.

Dembczyński, K.; Waegeman, W.; Cheng, W.; and Hüllermeier, E. 2011. An exact algorithm for f-measure maximization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS)*, 1404–1412. Curran Associates Inc.

Dembczyński, K.; Waegeman, W.; Cheng, W.; and Hüllermeier, E. 2012. On label dependence and loss minimization in multi-label classification. *Machine Learning* 88(1-2):5–45.

Dembczyński, K.; Cheng, W.; and Hüllermeier, E. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML)*, 279–286. Omnipress.

Destercke, S. 2015. Multilabel predictions with sets of probabilities: The hamming and ranking loss cases. *Pattern Recognition* 48(11):3757–3765.

Franc, V., and Prusa, D. 2019. On discriminative learning of prediction uncertainty. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 1963–1971.

Grandvalet, Y.; Rakotomamonjy, A.; Keshet, J.; and Canu, S. 2008. Support vector machines with a reject option. In *Proceedings of the 21st International Conference on Neural Information Processing Systems (NIPS)*, 537–544. Curran Associates Inc.

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

Jansche, M. 2007. A maximum expected utility framework for binary sequence labeling. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, 736–743.

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)*, 246–254. ACM.

Lin, H.-T.; Lin, C.-J.; and Weng, R. C. 2007. A note on platt's probabilistic outputs for support vector machines. *Machine learning* 68(3):267–276.

Park, L. A., and Simoff, S. 2015. Using entropy as a measure of acceptance for multi-label classification. In *Proceedings of the 14th International Symposium on Intelligent Data Analysis (IDA)*, 217–228. Springer.

Pillai, I.; Fumera, G.; and Roli, F. 2013. Multi-label classification with a reject option. *Pattern Recognition* 46(8):2256–2266.

Platt, J. C. 1999. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. *Advances in Large Margin Classiers*.

Quevedo, J. R.; Luaces, O.; and Bahamonde, A. 2012. Multilabel classifiers with a probabilistic thresholding strategy. *Pattern Recognition* 45(2):876–883.

Read, J.; Pfahringer, B.; Holmes, G.; and Frank, E. 2009. Classifier chains for multi-label classification. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part II (ECML/PKDD)*, 254–269. Springer-Verlag.

Tsoumakas, G.; Katakis, I.; and Vlahavas, I. 2009. Mining multilabel data. In *Data mining and knowledge discovery handbook*. Springer. 667–685.

Waegeman, W.; Dembczyński, K.; Jachnik, A.; Cheng, W.; and Hüllermeier, E. 2014. On the bayes-optimality of f-measure maximizers. *The Journal of Machine Learning Research* 15(1):3333–3388.

Ye, N.; Chai, K. M. A.; Lee, W. S.; and Chieu, H. L. 2012. Optimizing f-measures: a tale of two approaches. In *Proceedings of the 29th International Coference on International Conference on Machine Learning (ICML)*, 1555–1562. Omnipress.

Zhang, M.-L., and Zhou, Z.-H. 2014. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26(8):1819–1837.