

Regret Analysis for Performance Metrics in Multi-Label Classification: The Case of Hamming and Subset Zero-One Loss

Krzysztof Dembczyński^{1,3}, Willem Waegeman², Weiwei Cheng¹, and Eyke Hüllermeier¹

¹ Department of Mathematics and Computer Science, Marburg University, Hans-Meerwein-Str., 35039 Marburg, Germany

{cheng, dembczynski, eyke}@informatik.uni-marburg.de

² Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

willem.waegeman@ugent.be

³ Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland

Abstract. In multi-label classification (MLC), each instance is associated with a subset of labels instead of a single class, as in conventional classification, and this generalization enables the definition of a multitude of loss functions. Indeed, a large number of losses has already been proposed and is commonly applied as performance metrics in experimental studies. However, even though these loss functions are of a quite different nature, a concrete connection between the type of multi-label classifier used and the loss to be minimized is rarely established, implicitly giving the misleading impression that the same method can be optimal for different loss functions. In this paper, we elaborate on risk minimization and the connection between loss functions in MLC, both theoretically and empirically. In particular, we compare two important loss functions, namely the Hamming loss and the subset 0/1 loss. We perform a regret analysis, showing how poor a classifier intended to minimize the subset 0/1 loss can become in terms of Hamming loss and vice versa. The theoretical results are corroborated by experimental studies, and their implications for MLC methods are discussed in a broader context.

1 Introduction

The setting of *multi-label classification* (MLC) which, in contrast to conventional (single-label) classification, allows an instance to belong to several classes simultaneously, has received increasing attention in machine learning in recent years [1–6]. In particular, several approaches aiming at the exploitation of dependencies between class labels have been proposed. Even though the goal itself is clearly worthwhile, and empirically, many approaches have indeed been shown to improve predictive performance, a thorough theoretical analysis of the MLC setting is still missing.

Indeed, the notion of “label dependence” is often used in a purely intuitive manner. In this paper, we will argue that a careful distinction should be made between two different but related forms of statistical dependence in MLC, namely conditional and unconditional dependence. Moreover, we will establish a close connection between conditional label dependence and loss minimization. In MLC, a multitude of loss functions can be considered, and indeed, a large number of losses has already been proposed and is commonly applied as performance metrics in experimental studies. However, even though these loss functions are of a quite different nature, a concrete connection between the type of multi-label classifier used and the loss to be minimized is rarely established, implicitly giving the misleading impression that the same method can be optimal for different loss functions.

More specifically, this paper extends our previous work [7], in which we analyzed the connection between conditional label dependence and risk minimization for three loss functions commonly used in MLC problems: Hamming, rank and subset 0/1 loss. According to our results, the first two losses can in principle be minimized without taking conditional label dependence into account, which is not the case for the subset 0/1 loss.

In this paper, we further elaborate on the relationship between the Hamming and subset 0/1 loss. Our main theoretical result states that, even though we can establish mutual bounds for these loss functions, the bounds are not very tight. On the contrary, we can show that the minimization of subset 0/1 loss may come along with a very high regret in terms of Hamming loss and vice versa. As will be discussed in more detail later on, these results have important implications and suggest that previous experimental studies have often been interpreted in an incorrect way.

Let us also remark that the analysis performed in this paper is simplified by assuming an unconstrained hypothesis space. This allows for an analysis with respect to the joint conditional distribution alone. Regarding related work, we mention that generalization bounds have already been considered for problems with structured outputs. Some of these results apply directly to MLC as a special case [8, 9]. Moreover, it is worth mentioning that a similar problems can be found in information theory, namely bitwise and codeword decoding [10]. One can easily notice that the bitwise and codeword decoding correspond to Hamming loss and subset 0/1 loss minimization, respectively.

The structure of the paper is the following. Section 2 introduces the MLC problem in a formal way. Section 3 contains the main theoretical results concerning the bound and regret analysis. In Section 4, we present some experimental results confirming our theoretical claims. The last section concludes the paper.

2 Multi-Label 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.

2.1 Problem Statement

Let \mathcal{X} denote an instance space, and let $\mathcal{L} = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ be a finite set of class labels. We assume that an instance $\mathbf{x} \in \mathcal{X}$ is (non-deterministically) associated with a subset of labels $L \in 2^{\mathcal{L}}$; this subset is often called the set of relevant labels, while the complement $\mathcal{L} \setminus L$ is considered as irrelevant for \mathbf{x} . We identify a set L of relevant labels with a binary vector $\mathbf{y} = (y_1, y_2, \dots, y_m)$, in which $y_i = 1 \Leftrightarrow \lambda_i \in L$. By $\mathcal{Y} = \{0, 1\}^m$ we denote the set of possible labelings.

We assume observations to be generated independently and randomly according to a probability distribution $\mathbf{p}(\mathbf{X}, \mathbf{Y})$ on $\mathcal{X} \times \mathcal{Y}$, i.e., an observation $\mathbf{y} = (y_1, \dots, y_m)$ is the realization of a corresponding random vector $\mathbf{Y} = (Y_1, Y_2, \dots, Y_m)$. We denote by $\mathbf{p}_{\mathbf{x}}(\mathbf{Y}) = \mathbf{p}(\mathbf{Y} | \mathbf{x})$ the conditional distribution of \mathbf{Y} given $\mathbf{X} = \mathbf{x}$, and by $\mathbf{p}_{\mathbf{x}}^{(i)}(Y_i) = \mathbf{p}^{(i)}(Y_i | \mathbf{x})$ the corresponding marginal distribution of Y_i :

$$\mathbf{p}_{\mathbf{x}}^{(i)}(b) = \sum_{\mathbf{y} \in \mathcal{Y}: y_i = b} \mathbf{p}_{\mathbf{x}}(\mathbf{y})$$

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

$$\mathbf{h}(\mathbf{x}) = (h_1(\mathbf{x}), h_2(\mathbf{x}), \dots, h_m(\mathbf{x})).$$

The problem of MLC can be stated as follows: Given training data in the form of a finite set of observations $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, drawn independently from $\mathbf{p}(\mathbf{X}, \mathbf{Y})$, the goal is to learn a classifier $\mathbf{h} : \mathcal{X} \rightarrow \mathcal{Y}$ that generalizes well beyond these observations in the sense of minimizing the risk with respect to a specific loss function.

2.2 Label Dependence

As already announced in the introduction, we propose to distinguish two types of dependence. We call the labels \mathbf{Y} *unconditionally independent* if and only if

$$\mathbf{p}(\mathbf{Y}) = \prod_{i=1}^m \mathbf{p}^{(i)}(Y_i). \quad (1)$$

On the other hand, the labels are *conditionally independent* if the joint posterior distribution is the product of the marginals:

$$\mathbf{p}_{\mathbf{x}}(\mathbf{Y}) = \prod_{i=1}^m \mathbf{p}_{\mathbf{x}}^{(i)}(Y_i)$$

Obviously, both types of dependence are related to each other, since

$$\mathbf{p}(\mathbf{Y}) = \int_{\mathcal{X}} \mathbf{p}_{\mathbf{x}}(\mathbf{Y}) d\mathbf{P}(\mathbf{x}).$$

Nevertheless, unconditional dependence does *not* imply *nor* is implied by conditional dependence.

It has been widely established in statistics that exploiting unconditional label dependence can improve the generalization performance, because unconditional label dependence mainly originates from a similar structure of the different models [11]. The same arguments have played a key role in the development of related areas like multi-task learning and transfer learning, where task i and task j as well their models are assumed to be related [12]. As we will show the conditional dependence is rather connected with loss functions and their minimizers.

Let us remind that the joint distribution of a random vector $\mathbf{Y} = (Y_1, \dots, Y_m)$ can be expressed by the product rule of probability:

$$\mathbf{p}(\mathbf{Y}) = \mathbf{p}(Y_1) \prod_{i=2}^m \mathbf{p}(Y_i | Y_1, \dots, Y_{i-1})$$

If Y_1, \dots, Y_m are independent, then the product rule simplifies to (1).

2.3 Loss Functions

The performance in MLC is perhaps most frequently reported in terms of the Hamming loss, which is defined as the fraction of labels whose relevance is incorrectly predicted.⁴

$$L_H(\mathbf{y}, \mathbf{h}(\mathbf{x})) = \frac{1}{m} \sum_{i=1}^m \llbracket y_i \neq h_i(\mathbf{x}) \rrbracket. \quad (2)$$

Another natural loss function in the MLC setting is generalization of the well-known 0/1 loss from the conventional to the multi-label setting:

$$L_s(\mathbf{y}, \mathbf{h}(\mathbf{x})) = \llbracket \mathbf{y} \neq \mathbf{h}(\mathbf{x}) \rrbracket. \quad (3)$$

This loss function is referred to as *subset 0/1* loss. Admittedly, it may appear overly stringent, especially in the case of many labels. Moreover, since making a mistake on a single label is punished as hardly as a mistake on all labels, it does not discriminate well between “almost correct” and completely wrong predictions. However, mainly because of the fact that it is so extreme, it is especially relevant for our discussion about label dependence. Besides, as will be seen in more detail later on, it is a strong complement to the Hamming loss.

3 Analysis of Hamming and Subset 0/1 Loss

In this section, we analyze the Hamming and the subset 0/1 loss. The analysis is performed by assuming an unconstrained hypothesis space. This allows us to simplify the analysis by considering the conditional distribution for a given \mathbf{x} . First, we recall the risk minimizers of the two loss functions, already presented

⁴ For a predicate P , the expression $\llbracket P \rrbracket$ evaluates to 1 if P is true and to 0 if P is false.

in [7], and then show that, despite being different in general, they may coincide under specific conditions. Further, we derive mutual bounds for the two loss functions. Finally, we will show how poorly a classifier intended to minimize the subset 0/1 loss can perform in terms of Hamming loss and vice versa.

3.1 Risk Minimization

The risk of a classifier \mathbf{h} is defined as the expected loss over the joint distribution $\mathbf{p}(\mathbf{X}, \mathbf{Y})$:

$$R_L(\mathbf{h}) = \mathbb{E}_{\mathbf{X}\mathbf{Y}}L(\mathbf{Y}, \mathbf{h}(\mathbf{X})), \quad (4)$$

where $L(\cdot)$ is a loss function on multi-label predictions. A risk-minimizing model \mathbf{h}^* is given by

$$\mathbf{h}^* = \arg \min_{\mathbf{h}} \mathbb{E}_{\mathbf{X}\mathbf{Y}}L(\mathbf{Y}, \mathbf{h}(\mathbf{X})) = \arg \min_{\mathbf{h}} \mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\mathbf{Y}|\mathbf{X}}L(\mathbf{Y}, \mathbf{h}(\mathbf{X}))] \quad (5)$$

and determined in a pointwise way by the *Bayes optimal decisions*

$$\mathbf{h}^*(\mathbf{x}) = \arg \min_{\mathbf{y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}}L(\mathbf{Y}, \mathbf{y}). \quad (6)$$

For the Hamming loss (2), it is easy to see that the risk minimizer (6) is obtained by

$$\mathbf{h}_H^*(\mathbf{x}) = (h_{H_1}(\mathbf{x}), \dots, h_{H_m}(\mathbf{x})),$$

where

$$h_{H_i}(\mathbf{x}) = \arg \max_{b \in \{0,1\}} \mathbf{p}_{\mathbf{x}}^{(i)}(b) \quad (i = 1, \dots, m). \quad (7)$$

The Bayes prediction for (3) is also straight-forward. As for any other 0/1 loss, it simply consists of predicting the mode of the distribution:

$$\mathbf{h}_s^*(\mathbf{x}) = \arg \max_{\mathbf{y} \in \mathcal{Y}} \mathbf{p}_{\mathbf{x}}(\mathbf{y}) \quad (8)$$

As one of the most important consequences of the above results we note that, according to (7), a risk-minimizing prediction for the Hamming loss can be obtained from the marginal distributions $\mathbf{p}_{\mathbf{x}}^{(i)}(Y_i)$ ($i = 1, \dots, m$) alone. In other words, it is not necessary to know the joint label distribution $\mathbf{p}_{\mathbf{x}}(\mathbf{Y})$ on \mathcal{Y} . For finding the minimizer (8), on the other hand, the joint distribution must obviously be known. These results suggest that taking conditional label dependence into account is less important for Hamming loss than for subset 0/1 loss.

Despite the differences noted above, we can show that the two risk minimizers coincide under specific conditions. More specifically, we can show the following proposition.

Proposition 1. *The Hamming loss and subset 0/1 have the same risk minimizer, i.e., $\mathbf{h}_H^*(\mathbf{x}) = \mathbf{h}_s^*(\mathbf{x})$, if one of the following conditions holds:*

- (1) *Labels Y_1, \dots, Y_m are conditionally independent, i.e., $\mathbf{p}_{\mathbf{x}}(\mathbf{Y}) = \prod_{i=1}^m \mathbf{p}_{\mathbf{x}}(Y_i)$.*

- (2) *The probability of the mode of the joint probability is greater or equal than 0.5, i.e., $\mathbf{p}_x(\mathbf{h}_s^*(\mathbf{x})) \geq 0.5$.*

Proof. (1) Since the joint probability of any combination of \mathbf{y} is given by the product of marginal probabilities, the highest value of this product is given by the highest values of the marginal probabilities. Thus, the joint mode is composed of the marginal modes.

(2) If $\mathbf{p}_x(\mathbf{h}_s^*(\mathbf{x})) \geq 0.5$, then $\mathbf{p}_x(h_{s_i}^*(\mathbf{x})) \geq 0.5$, $i = 1, \dots, m$, and from this it follows that $h_{s_i}^*(\mathbf{x}) = h_{H_i}^*(\mathbf{x})$. \square

As a simple corollary of this proposition, we have the following.

Corollary 1. *In the separable case (i.e., the joint conditional distribution is deterministic, $\mathbf{p}_x(\mathbf{Y}) = \llbracket \mathbf{Y} = \mathbf{y} \rrbracket$, where \mathbf{y} is a binary vector of size m), the risk minimizers of the Hamming loss and subset 0/1 coincide.*

Proof. If $\mathbf{p}_x(\mathbf{Y}) = \llbracket \mathbf{Y} = \mathbf{y} \rrbracket$, then $\mathbf{p}_x(\mathbf{Y}) = \prod_{i=1}^m \mathbf{p}_x(Y_i)$. In this case, we also have $\mathbf{p}_x(\mathbf{h}_s^*(\mathbf{x})) \geq 0.5$. Thus, the result follows from both (1) and (2) in Proposition 1. \square

3.2 Bound Analysis

So far, we have looked at the minimizers of Hamming and subset 0/1 loss and we have seen that these minimizers may coincide under special conditions. In general, however, they are different and, therefore, will call for different classifiers. Another natural question one may ask is the following: If we fix a classifier and we know, say, its subset 0/1 loss, can we say anything about its Hamming loss? This question is answered by the following proposition.

Proposition 2. *For all distributions of \mathbf{Y} given \mathbf{x} , and for all models \mathbf{h} , the expectation of the subset 0/1 loss can be bounded in terms of the expectation of the Hamming loss as follows:*

$$\frac{1}{m} \mathbb{E}_{\mathbf{Y}}[L_s(\mathbf{Y}, \mathbf{h}(\mathbf{x}))] \leq \mathbb{E}_{\mathbf{Y}}[L_H(\mathbf{Y}, \mathbf{h}(\mathbf{x}))] \leq \mathbb{E}_{\mathbf{Y}}[L_s(\mathbf{Y}, \mathbf{h}(\mathbf{x}))] \quad (9)$$

Proof. For a fixed $\mathbf{x} \in \mathcal{X}$, we can express the expected loss as follows:

$$\mathbb{E}_{\mathbf{Y}}[L(\mathbf{Y}, \mathbf{h}(\mathbf{x}))] = \sum_{\mathbf{y} \in \mathcal{Y}} \mathbf{p}(\mathbf{y}) L(\mathbf{y}, \mathbf{h}(\mathbf{x}))$$

Suppose we can express an MLC loss in terms of an aggregation $G : \{0, 1\}^m \rightarrow [0, 1]$ of the standard zero-one losses $L_{0/1}$ on individual labels (as used in conventional classification):

$$L(\mathbf{y}, \mathbf{h}(\mathbf{x})) = G(L_{0/1}(y_1, h_1(\mathbf{x})), \dots, L_{0/1}(y_m, h_m(\mathbf{x}))). \quad (10)$$

Indeed, the subset 0/1 loss and the Hamming loss can be written, respectively, as

$$\begin{aligned} G_{\max}(\mathbf{a}) &= G_{\max}(a_1, \dots, a_m) = \max\{a_1, \dots, a_m\} \\ G_{\text{mean}}(\mathbf{a}) &= G_{\text{mean}}(a_1, \dots, a_m) = \frac{1}{m}(a_1 + \dots + a_m). \end{aligned}$$

This immediately leads to the above lower and upper bound for the Hamming loss. The proposition then immediately follows from the fact that $\frac{1}{m}G_{\max}(\mathbf{a}) \leq G_{\text{mean}}(\mathbf{a}) \leq G_{\max}(\mathbf{a})$ for all $\mathbf{a} \in [0, 1]^m$. \square

Interestingly, it turns out that the location of the Hamming loss between the bounds in (9) is in direct correspondence with the conditional dependence between the labels, and when the dependence structure of the conditional distribution of \mathbf{Y} is given, the difference between Hamming loss and subset 0/1 loss can be determined in a more precise way. Roughly speaking, the less (more) dependent the labels are, the more the Hamming loss moves toward the lower (upper) bound. Without going into detail, we note that more precise estimations of the difference between subset 0/1 loss and Hamming loss can be derived with the help of copulas [13].

Nevertheless, we need to emphasize that a complete analysis of the relationship between bound (9) and conditional label dependence has to take additional factors into account. One of these factors is the hypothesis space one considers; the lower bound will become more tight when restricting to certain hypothesis spaces. Another important factor is the interplay between conditional and unconditional label dependence. For example, in case of full positive dependence between all labels, estimating the Hamming loss might still be more simple than estimating the subset 0/1 loss, implying that the upper bound will not behave as an equality. The analysis of this section mainly provides insights on the behavior of the different loss functions for the underlying distribution. Yet, we realize that the picture might look different in terms of estimated performance after training on a finite set of examples.

3.3 Regret Analysis

Some of the previous results may suggest that, for learning a risk minimizing classifier, either loss functions can be used as a proxy of the other one. For example, the bounds in (9) may suggest that a low subset 0/1 loss will also imply a low Hamming loss. On the other hand, one may argue that the bounds themselves are rather weak, and reckon that the concrete difference in terms of Hamming and subset 0/1 loss may become quite high. In this section, we present a regret analysis showing that minimization of Hamming loss does not guarantee good performance in terms of subset 0/1 loss and vice versa.

The regret of a classifier \mathbf{h} with respect to a loss function L_z is defined as follows:

$$r_{L_z}(\mathbf{h}) = R_{L_z}(\mathbf{h}) - R_{L_z}(\mathbf{h}_z^*), \quad (11)$$

where R is the risk given by (4), and \mathbf{h}_z^* is the Bayes-optimal classifier with respect to the loss function L_z .

In the following, we consider the regret with respect to the Hamming loss, given by

$$r_H(\mathbf{h}) = \mathbb{E}_{\mathbf{X}\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}(\mathbf{X})) - \mathbb{E}_{\mathbf{X}\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{X})),$$

and the subset 0/1 loss, given by

$$r_H(\mathbf{h}) = \mathbb{E}_{\mathbf{X}\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}(\mathbf{X})) - \mathbb{E}_{\mathbf{X}\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{X})).$$

Since both loss functions are decomposable with respect to individual instances, we analyze the expectation over \mathbf{Y} for a given \mathbf{x} . The first result concerns the highest value of the regret in terms of the subset 0/1 loss for $\mathbf{h}_H^*(\mathbf{X})$, the optimal strategy for the Hamming loss.

Proposition 3. *The following upper bound holds:*

$$\mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x})) - \mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) < 0.5.$$

Moreover, this bound is tight, i.e.,

$$\sup_{\mathbf{p}} (\mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x})) - \mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x}))) = 0.5,$$

where the supremum is taken over all probability distributions on \mathcal{Y} .

Proof. Since the risk of any classifier \mathbf{h} is within the range $[0, 1]$, the maximal value of the regret is 1. However, according to the second part of Proposition 1, both risk minimizers coincide if $\mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) \leq 0.5$. Consequently, the regret must be (strictly) smaller than 0.5. To prove the tightness of the bound, we show that, for any $\delta \in (0, \frac{1}{6})$, there is a probability distribution \mathbf{p} that yields the regret $0.5 - \delta$. Define \mathbf{p} as follows:

$$\mathbf{p}(\mathbf{y}) = \begin{cases} \frac{1}{2} - \delta, & \text{if } \mathbf{y} = (a_1, \dots, a_{k-1}, \bar{a}_{k+1}, \dots, \bar{a}_m) \\ \frac{1}{2} - \delta, & \text{if } \mathbf{y} = (\bar{a}_1, \dots, \bar{a}_{k-1}, a_{k+1}, \dots, a_m) \\ 2\delta, & \text{if } \mathbf{y} = (a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_m) \end{cases}$$

where $a_i \in \{0, 1\}$ and $\bar{a}_i = 1 - a_i$. Such a distribution can be constructed for all $m > 1$. Obviously,

$$\begin{aligned} \mathbf{h}_s^*(\mathbf{x}) &= (a_1, \dots, a_{k-1}, \bar{a}_{k+1}, \dots, \bar{a}_m) \quad \text{or} \\ \mathbf{h}_s^*(\mathbf{x}) &= (\bar{a}_1, \dots, \bar{a}_{k-1}, a_{k+1}, \dots, a_m) \end{aligned}$$

and

$$\mathbf{h}_H^*(\mathbf{x}) = (a_1, \dots, a_{k-1}, a_{k+1}, \dots, a_m)$$

Finally, we thus obtain

$$\mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x})) = 1 - 2\delta$$

and

$$\mathbb{E}_{\mathbf{Y}} L_s(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) = 0.5 - \delta,$$

which immediately proves the proposition. \square

The second result concerns the highest value of the regret in terms of the Hamming loss for $\mathbf{h}_s^*(\mathbf{X})$, the optimal strategy for the subset 0/1 loss.

Proposition 4. *The following upper bound holds for $m > 3$:*

$$\mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) - \mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x})) < \frac{m-2}{m+2}.$$

Moreover, this bound is tight, i.e.

$$\sup_{\mathbf{p}} (\mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) - \mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x}))) = \frac{m-2}{m+2},$$

where the supremum is taken over all probability distributions on \mathcal{Y} .

Proof. Because of space limitations, we only show how to construct the distribution for which the regret is close to the given bound.

Let $a_i \in \{0, 1\}$ and $\bar{a}_i = 1 - a_i$. If $\mathbf{a}_m = (a_1, a_2, \dots, a_m)$ is a $\{0, 1\}$ -vector of length m , then $\bar{\mathbf{a}}_m$ denotes the vector $(\bar{a}_1, \bar{a}_2, \dots, \bar{a}_m)$. Furthermore, let $d_H(\mathbf{a}, \mathbf{b})$ denote the Hamming distance, given by

$$d_H(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^m |a_i - b_i|$$

for all $\mathbf{a}, \mathbf{b} \in \{0, 1\}^m$. Now, consider a joint probability distribution defined as follows:

$$\mathbf{p}(\mathbf{y}) = \begin{cases} \frac{1}{m+2} + \delta & \text{if } \mathbf{y} = \mathbf{a}_m \\ \frac{1}{m+2} - \frac{\delta}{m+1} & \text{if } d_H(\mathbf{y}, \bar{\mathbf{a}}_m) \leq 1 \\ 0 & \text{otherwise} \end{cases},$$

where $\delta > 0$. Hence, we obtain:

$$\begin{aligned} \mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) &= \frac{1}{m+2} - \frac{\delta}{m+1} + m \left(\frac{1}{m+2} - \frac{\delta}{m+1} \right) \frac{m-1}{m}, \\ \mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x})) &= \frac{1}{m+2} + \delta + m \left(\frac{1}{m+2} - \frac{\delta}{m+1} \right) \frac{1}{m}. \end{aligned}$$

The difference is then given by

$$\mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_s^*(\mathbf{x})) - \mathbb{E}_{\mathbf{Y}} L_H(\mathbf{Y}, \mathbf{h}_H^*(\mathbf{x})) = \frac{m-2}{m+2} - \delta \left(\frac{m-1}{m+1} + 1 \right).$$

Since this holds for any $\delta > 0$, the regret is close to the bound. \square

As we can see, the regret is quite high in both cases, suggesting that a single classifier will not be able to perform equally well in terms of both loss functions. Instead, a classifier specifically tailored for the Hamming (subset 0/1) loss will indeed perform much better for this loss than a classifier trained to minimize the subset 0/1 (Hamming) loss.

3.4 Summary and Implications of Theoretical Results

Our theoretical results so far can be summarized as follows:

- The risk minimizers of Hamming and subset 0/1 loss have a different structure: In the latter case, the minimizer is the mode of a joint distribution on the label space, whereas in the former, it is a combination of the modes of (one-dimensional) marginal distributions.
- Under specific conditions, these two types of loss minimizers are provably equivalent, though in general, they will produce different predictions.
- The Hamming loss is upper-bounded by the subset 0/1 loss, which in turn is bounded by the Hamming loss multiplied by the number of labels m .
- Minimization of the subset 0/1 loss may cause a high regret for the Hamming loss and vice versa.

These results have a number of implications, not only from a theoretical but also from a methodological and empirical point of view:

- The idea to exploit label dependencies, one of the main research topics in MLC, should be reconsidered in a careful way, distinguishing the two types of dependence mentioned above. The conditional dependence, for example, is arguably more related to non-decomposable (with respect to labels) losses like subset 0/1 loss than to decomposable ones like Hamming. This distinction is largely ignored in papers on that topic.
- A careful distinction between loss functions seems to be even more important for MLC than for standard classification, and one cannot expect the same MLC method to be optimal for different types of losses. Surprisingly, new methods are often proposed without explicitly saying what loss they intend to minimize. Instead, they are typically shown to perform well across a wide spectrum of different loss functions, casting some doubts on the reliability of such studies.

4 Experimental Studies

To corroborate our theoretical results by means of empirical evidence, this section presents a number of experimental studies, using both synthetic and benchmark data. As MLC methods, two meta-techniques will be employed, namely the Binary Relevance (BR) and the Label Power-set (LP) classifier. These methods are commonly used as baselines in experimental studies and are of a quite complementary nature [4]. Besides, we will also propose a simple modification of LP that allows for adapting this approach to any loss function. We present results on three artificial data sets pointing to some important pitfalls often encountered in experimental studies of MLC. Finally, we present some results on benchmark data sets and discuss them in the light of these pitfalls.

In the experimental study, we used the WEKA [14] and Mulan [6] packages.

4.1 Binary Relevance and Label Power-set Classifier

BR is arguably the simplest approach to MLC. It trains a separate binary classifier $h_i(\cdot)$ for each label λ_i . Learning is performed independently for each label, ignoring all other labels. At prediction time, a query instance \mathbf{x} is submitted to all binary classifiers, and their outputs are combined into an MLC prediction.

Obviously, BR is tailored for Hamming loss minimization or, more generally, every loss whose risk minimizer can be expressed solely in terms of marginal distributions; as shown in [7], this also includes the rank loss. However, BR does not take label dependence into account, neither conditional nor unconditional, and this is what it is most often criticized for. Indeed, as suggested by our theoretical results, BR will in general not be able to yield risk minimizing predictions for losses like subset 0/1. Moreover, one may suspect that, even though it can minimize Hamming loss theoretically, exploiting label dependencies may still be beneficial practically.

LP reduces the MLC problem to multi-class classification, considering each label subset $L \in \mathcal{L}$ as a distinct meta-class. The number of these meta-classes may become as large as $|\mathcal{L}| = 2^m$, although it is often reduced considerably by ignoring label combinations that never occur in the training data. Nevertheless, the large number of classes produced by this reduction is generally seen as the most important drawback of LP.

Since prediction of the most probable meta-class is equivalent to prediction of the mode of the joint label distribution, LP is tailored for the subset 0/1 loss. Interestingly, however, it can easily be extended to any other loss function, given that the underlying multi-class classifier $\mathbf{f}(\cdot)$ does not only provide a class prediction but a reasonable estimate of the probability of all meta-classes (label combinations), i.e., $\mathbf{f}(\mathbf{x}) \approx \mathbf{p}_{\mathbf{x}}(\mathbf{Y})$. Given a loss function $L(\cdot)$ to be minimized, an optimal prediction can then be derived in an explicit way:

$$\mathbf{h}^*(\mathbf{x}) = \arg \min_{\mathbf{y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}} L(\mathbf{Y}, \mathbf{y})$$

In particular, LP can be improved for the Hamming loss, simply by computing the marginal distributions and combining the marginal modes into a single MLC prediction. In this regard, we note that computing margins is not harder than searching the mode of $\mathbf{p}_{\mathbf{x}}(\mathbf{Y})$. We refer to this modification of LP as LP+.

Practically, we improve LP+ by regularizing the (joint) probability estimation. To this end, we make use of shrinking. In Bayesian inference, it is well-known that the estimated parameters are shrunk toward the prior distribution. We mimic such kind of shrinkage by means of a regularized probability estimate:

$$\tilde{p}_{\mathbf{x}}(\mathbf{y}) = \alpha \hat{p}_{\mathbf{x}}(\mathbf{y}) + (1 - \alpha) \hat{p}(\mathbf{y}),$$

where $\hat{p}_{\mathbf{x}}(\mathbf{y})$ is given by LP, $\hat{p}(\mathbf{y})$ is a prior estimated from the training data, and α is the shrinkage parameter. This parameter can be determined empirically so as to maximize performance on the test set: For given α , the accuracy of the classifier is estimated on a validation set, and an optimal α is found through line-search. In the following, we use $\alpha = 0.95$.

Table 1. Results on two artificial data sets: conditionally independent (left) and conditionally dependent (right). Standard errors are given in parentheses.

classifier	Conditional independence		Conditional dependence	
	Hamming loss	subset 0/1 loss	Hamming loss	subset 0/1 loss
BR	0.4208(.0014)	0.8088(.0020)	0.3900(.0015)	0.7374(.0021)
LP	0.4212(.0011)	0.8101(.0025)	0.4227(.0019)	0.6102(.0033)
LP+	0.4181(.0013)	0.8093(.0021)	0.3961(.0033)	0.6135(.0034)
B-O	0.4162	0.8016	0.3897	0.6029

4.2 Artificial Data

We consider three artificial data sets, each one reflecting a typical situation for MLC. In each case, we generated 30 training and testing folds, each containing 1000 instances.

The first data set represents the case of conditional independence. Data are drawn uniformly from the square $\mathbf{x} \in [-0.5, 0.5]^2$. The label distribution is given by the product of the marginal distributions defined by $\mathbf{p}_{\mathbf{x}}(y_i) = 1/(1 + \exp(-f_i(\mathbf{x})))$, where the f_i are linear functions: $f_1(\mathbf{x}) = x_1 + x_2$, $f_2(\mathbf{x}) = -x_1 + x_2$, $f_3(\mathbf{x}) = x_1 - x_2$. The cardinality of labels (the average number of relevant labels for an instance) is 1.503.

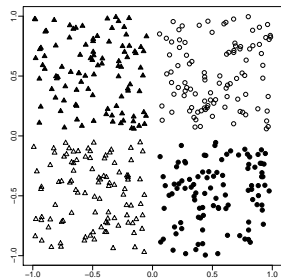
In the second data set, the labels are dependent. Data are drawn from the univariate uniform distribution $\mathbf{x} \in [-0.5, 0.5]$. The joint distribution is obtained by applying the product rule of probability:

$$\mathbf{p}_{\mathbf{x}}(\mathbf{Y}) = \mathbf{p}_{\mathbf{x}}(Y_1) \prod_{i=2}^3 \mathbf{p}_{\mathbf{x}}(Y_i | Y_1, \dots, Y_{i-1}),$$

where the probabilities are modeled by linear functions in a similar way as before: $f_1(x) = x$, $f_2(y_1, x) = -x - 2y_1 + 1$, $f_3(y_2, y_1, x) = x + 12y_1 - 2y_2 - 11$. The cardinality of labels for this data set is 1.314.

The results of the experiment are reported for both data sets in Table 4.2. All approaches are used with linear support vector machine as base learner. In the case of LP+, we used the approach of [15] to turn SVM scores into probabilities, thus obtaining an estimation of the joint distribution and its marginals. Since the true data generating process is known, we also report the loss of the Bayes-optimal classifier. In the case of the independent data, we observe that both approaches perform equally well in terms of both loss functions. For the dependent data, however, we see that BR and LP are tailored toward different loss functions. As expected, the former performs well in terms of Hamming loss, whereas the latter is superior in terms of subset 0/1 loss. As expected, LP+ is able to adapt to both loss functions. Overall, the results are in complete agreement with our theoretical findings.

In the literature, LP is often shown to outperform BR even in terms of Hamming loss. Given our results so far, this is somewhat surprising and calls



classifier	Hamming loss	subset 0/1 loss
BR Linear SVM	0.2399(.0097)	0.4751(.0196)
BR MLRules	0.0011(.0002)	0.0020(.0003)
LP Linear SVM	0.0143(.0020)	0.0195(.0011)
B-O	0	0

Fig. 1. Plot of the data set composed of two labels: the first label is obtained by a linear model, while the second label represents the exclusive disjunction. The table contains results of three classifiers on this data set.

for an explanation. We argue that results of that kind should be considered with caution, mainly because a meta learning technique (such as BR and LP) must always be considered in conjunction with the underlying base learner. In fact, differences in performance should not only be attributed to the meta but also to the base learner. In particular, since BR uses binary and LP multi-class classification, they are typically applied with different base learners, and hence are not directly comparable.

We illustrate this by means of an example. For simplicity, suppose that data is generated without noise (whence the risk of the Bayes optimal classifier for both Hamming and subset 0/1 loss is 0), and consider a problem with two-dimensional instances $\mathbf{x} = (x_1, x_2) \in \mathcal{X} = [-1, 1]^2$ and two labels: $y_1 = \llbracket x_1 < 0 \rrbracket$ and $y_2 = \llbracket x_1 > 0 \rrbracket \odot \llbracket x_2 > 0 \rrbracket$, where \odot is the exclusive (logical) disjunction. Obviously, using a linear base learner, BR is not able to solve this problem properly, whereas LP, using a multi-class extension of the linear support vector machine (based on a one-vs-one decomposition) yields almost perfect predictions. However, this multi-class extension is no longer a truly linear classifier. Instead, several linear classifiers are wrapped in a decomposition and an aggregation procedure, yielding a more complex classifier that can produce non-linear decision boundaries. And indeed, giving BR access to a more complex base learner, like a rule ensemble [16], it is able to solve the problem equally well; see results and the scatter plot of data in Fig. 1.

4.3 Benchmark Data

The second part of the experiments was performed on a collection of 8 MLC data sets.⁵ In the case of the *Reuters* data, we used the preprocessed version as in [5]. A summary of the data sets and their properties are given in Table 2.

We used BR and LP with linear support vector machines as base learner, and additionally BR with MLRules and LP+ based on the probabilistic SVM.

⁵ Data sets are taken from <http://mlkd.csd.auth.gr/multilabel.html> and <http://www.cs.waikato.ac.nz/~jmr30/#datasets>

Results of a 3-fold cross-validation are given for Hamming loss in Table 3 and for subset 0/1 loss in Table 4. Overall, the results are again in agreement with our expectations. In particular, LP achieves better results for the subset 0/1 loss, while BR is on average superior in terms of Hamming loss.

Let us have a closer look at the results for the *scene* data set. As reported in [17], LP outperforms BR on this data set in terms of Hamming loss; both methods were used with linear SVM as base learner. Although our results here give the same picture, note that BR with MLRules outperforms both approaches. As pointed out above, comparing LP and BR with the same base learner is questionable and may lead to unwarranted conclusions.

Let us underline that LP+ outperforms LP using probabilistic SVM in terms of the Hamming loss on all datasets. This confirms our theoretical claims and justifies the modification of LP. Also shrinking used in LP+ improves the results for the subset 0/1 loss. However, let us notice that probabilistic SVM performs worse than classical SVM in the case of classification. We can observe that, in terms of the subset 0/1 loss, the latter is much better than the former. Moreover, the latter receives good results for Hamming loss minimization on some data sets. We suppose that, for these data sets, one of the conditions that imply equivalence of the risk minimizers will hold (cf. Proposition 1).

5 Conclusions

In this paper, we have addressed a number of issues related to loss minimization in multi-label classification. In our opinion, this topic has not received enough attention so far, despite the increasing interest in MLC in general. However, as we have argued in this paper, empirical studies of MLC methods are often meaningless or even misleading without a careful interpretation, which in turn requires a thorough understanding of underlying theoretical conceptions.

In particular, by looking at the current literature, we noticed that papers proposing new methods for MLC, and for exploiting label dependencies, rarely distinguish between the type of loss function to be minimized. Instead, a new method is often shown to be better than existing ones “on average”, evaluating on a number of different loss functions. Our technical results in this paper, already summarized in Section 3.4 and therefore not repeated here, indicate that studies of that kind might be less illuminative than they could be. First, we have shown that the type of loss function has a strong influence on whether or not, and perhaps to what extent, an exploitation of conditional label dependencies can be expected to yield a true benefit. Consequently, some loss functions will be more suitable than others for showing the benefit of label dependencies. Second, using the example of Hamming and subset 0/1 loss, we have shown that loss functions in MLC cover a broad spectrum, and that minimizing different losses will normally require different estimators. Consequently, one cannot expect an MLC method to perform equally well for various losses of different type.

Our focus on Hamming and subset 0/1 loss can be justified by their complementarity, and by noting that these losses can be considered representative of

decomposable and non-decomposable loss functions, respectively. Besides, they are among the most well-known and frequently used performance measures in MLC. Nevertheless, looking at other loss functions is of course worthwhile and holds the promise to gain further insight into the nature of MLC. Expanding our studies in this direction is therefore on our agenda for future work.

References

1. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning multi-label scene classification. *Pattern Recognition* **37**(9) (2004) 1757–1771
2. Ghamrawi, N., McCallum, A.: Collective multi-label classification. In: *CIKM 2005*. (2005) 195–200
3. Amit, Y., Dekel, O., Singer, Y.: A boosting algorithm for label covering in multilabel problems. In: *JMLR W&P. Volume 2*. (2007) 27–34
4. Tsoumakas, G., Katakis, I.: Multi label classification: An overview. *Int J Data Warehousing and Mining* **3**(3) (2007) 1–13
5. Cheng, W., Hüllermeier, E.: Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning* **76**(2-3) (2009) 211–225
6. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In Maimon, O., Rokach, L., eds.: *Data Mining and Knowledge Discovery Handbook*, Springer (2010)
7. Dembczyński, K., Cheng, W., Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: *ICML 2010*. (2010)
8. Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. In: *NIPS 16*, MIT Press (2004)
9. McAllester, D.: Generalization bounds and consistency for structured labeling. In: *Predicting Structured Data*. MIT Press (2007)
10. MacKay, D.J.C.: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press (2003)
11. Breiman, L., Friedman, J.: Predicting multivariate responses in multiple linear regression. *J R Stat. Soc. Ser. B* **69** (1997) 3–54
12. Caruana, R.: Multitask learning: A knowledge-based source of inductive bias. *Machine Learning* **28** (1997) 41–75
13. Nelsen, R.: *An Introduction to Copulas*, Second Edition. Springer (2006)
14. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*, 2nd Edition. Morgan Kaufmann, San Francisco (2005)
15. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in Large Margin Classifiers*, MIT Press (1999) 61–74
16. Dembczyński, K., Kotłowski, W., Słowiński, R.: Maximum likelihood rule ensembles. In: *ICML 2008*. (2008) 224–231
17. Tsoumakas, G., Vlahavas, I.: Random k-labelsets: An ensemble method for multi-label classification. In: *ECML 2007*. (2007) 406–417

Table 2. Data sets used in the experiment.

data set	# inst.	# attr.	# labels	card.	data set	# inst.	# attr.	# labels	card.
image	2000	135	5	1.236	yeast	2417	103	14	4.237
scene	2407	294	6	1.074	genbase	662	1186	27	1.252
emotions	593	72	6	1.868	slashdot	3782	1079	22	1.181
reuters	7119	243	7	1.241	medical	978	1449	45	1.245

Table 3. Results for Hamming loss. Ranks of classifiers are given in parentheses.

	BR SVM	BR MLRules	LP+ pSVM	LP pSVM	LP SVM
image	0.1980 (4)	0.1928(2)	0.1888(1)	0.2021(5)	0.1954(3)
scene	0.1071 (5)	0.0871(1)	0.0919(3)	0.0950(4)	0.0891(2)
emotions	0.2049 (1)	0.2080(2)	0.2091(3)	0.2232(5)	0.2119(4)
reuters	0.0663 (5)	0.0479(1)	0.0565(2)	0.0596(3)	0.0628(4)
yeast	0.2016 (1)	0.2086(3)	0.2156(4)	0.2523(5)	0.2075(2)
genbase	0.0008 (1)	0.0015(5)	0.0011(3)	0.0012(4)	0.0010(2)
slashdot	0.0480 (2)	0.0402(1)	0.0534(4)	0.0631(5)	0.0481(3)
medical	0.0102 (1)	0.0106(2)	0.0132(4)	0.0135(5)	0.0115(3)
Avg. Rank	2.7	2.1	2.75	4.25	3.2

Table 4. Results for subset 0/1 loss. Ranks of classifiers are given in parentheses.

	BR SVM	BR MLRules	LP+ pSVM	LP pSVM	LP SVM
image	0.7670 (5)	0.6705(4)	0.5595(2)	0.5600(3)	0.5315(1)
scene	0.4757 (5)	0.4221(4)	0.3299(2)	0.3303(3)	0.3008(1)
emotions	0.7538 (4)	0.7622(5)	0.7353(2)	0.7386(3)	0.6846(1)
reuters	0.3735 (5)	0.2684(4)	0.2391(1)	0.2406(2)	0.2676(3)
yeast	0.8552 (4)	0.8643(5)	0.8155(2)	0.8159(3)	0.7460(1)
genbase	0.0211 (1.5)	0.0332(5)	0.0257(3.5)	0.0257(3.5)	0.0211(1.5)
slashdot	0.6560 (2)	0.6721(3)	0.6819(4)	0.6835(5)	0.5460(1)
medical	0.3405 (2)	0.3497(3)	0.3630(4.5)	0.3630(4.5)	0.3119(1)
Avg. Rank	3.56	4.13	2.63	3.38	1.31