Regret Analysis for Performance Metrics in Multi-Label Classification

The Case of Hamming and Subset Zero-One Loss

Krzysztof Dembczyński[1], Willem Waegeman[2], Weiwei Cheng[1], and Eyke Hüllermeier[1]

[1]Knowledge Engineering and Bioinformatics (KEBI) Lab
Philipps-Universität Marburg

[2]Research Unit Knowledge-based Systems (KERMIT)
Universiteit Gent

Multi-Label Classification

- Given a vector $\boldsymbol{x} \in \mathcal{X}$ of features, the goal is to learn a function $\mathbf{h}(\boldsymbol{x})$ that predicts accurately a binary vector $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathcal{Y}$ of labels.
- Example: Given a news report, the goal is to learn a machine that tags the news report with relevant categories.
- The simple solution: solve the problem for each of the label independently.

Since the prediction is made for all labels **simultaneously**, two interesting issues appear:

- A multitude of loss functions defined over multiple labels,
- Dependence/correlation between labels.

Pitfalls in MLC Studies

In recent years, a plenty of algorithms has been introduced . . .

- A **large** number of **loss functions** is commonly applied as performance metrics, but a concrete **connection** between a **multi-label classifier** and a **loss function** is **rarely** established.

- A **large** number of **loss functions** is commonly applied as performance metrics, but a concrete **connection** between a **multi-label classifier** and a **loss function** is **rarely** established.
- This gives implicitly the misleading impression that the **same** method can be **optimal** for **different** loss functions.

Pitfalls in MLC Studies

- A **large** number of **loss functions** is commonly applied as performance metrics, but a concrete **connection** between a **multi-label classifier** and a **loss function** is **rarely** established.
- This gives implicitly the misleading impression that the **same** method can be **optimal** for **different** loss functions.
- It is assumed that **performance** can be improved by taking the **label dependence** into account, but this term is used in an **intuitive** manner, without any precise **formal** definition.

Pitfalls in MLC Studies

- The empirical results are given **on average** without investigation under which conditions a given algorithm benefits.

- The empirical results are given **on average** without investigation under which conditions a given algorithm benefits.
- The reasons for improvements are **not distinguished**.

Loss Functions

Let us discuss multi-label loss functions ...

Hamming and Subset 0/1 Loss

- **Hamming loss** measures the fraction of labels whose relevance is incorrectly predicted:

$$L_H(\boldsymbol{y}, \mathbf{h}(\boldsymbol{x})) = \frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h_i(\boldsymbol{x})]\!],$$

- while **subset 0/1 loss** measures whether the prediction totally agrees with the true labeling:

$$L_s(\boldsymbol{y}, \mathbf{h}(\boldsymbol{x})) = [\![\boldsymbol{y} \neq \mathbf{h}(\boldsymbol{x})]\!].$$

Analysis of Hamming and Subset 0/1 Loss

Analysis contains:

- The form of risk minimizers,
- Whether the risk minimizers coincide in some circumstances,
- Bound analysis,
- Regret analysis.

- The analysis is simplified by assuming an unconstrained hypothesis space.

Risk Minimizers

- The risk minimizer defined by:

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

- is given for the Hamming loss by:

$$h_i^*(\boldsymbol{x}) = \arg\max_{b \in \{0,1\}} \mathbf{P}(y_i = b \,|\, \boldsymbol{x}), \quad i = 1, \ldots, m,$$

- while for the subset 0/1 loss by:

$$\mathbf{h}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbf{P}(\boldsymbol{y} \,|\, \boldsymbol{x}).$$

- The minimizer of the Hamming loss is the **marginal mode**, while for the subset 0/1 loss the **joint mode**.

Coincidence of Risk Minimizers

Proposition

The Hamming loss and subset 0/1 have **the same risk minimizer**,

$$\boldsymbol{h}_H^*(\boldsymbol{x}) = \boldsymbol{h}_s^*(\boldsymbol{x}),$$

if one of the following conditions holds:

(1) *Labels* $Y_1, \ldots, Y_m$ *are* **conditionally** m**-independent**,

$$\mathbf{P}(\mathbf{Y}|\boldsymbol{x}) = \prod_{i=1}^{m} \mathbf{P}(Y_i|\boldsymbol{x}).$$

(2) *The probability of the* **joint mode** *satisfies,*

$$\mathbf{P}(\boldsymbol{h}_s^*(\boldsymbol{x})|\boldsymbol{x}) \geq 0.5.$$

Proposition

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

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

The previous results may suggest that one of the loss functions can be used as a proxy of the other:

- For some situations both risk minimizers coincide,
- One can provide mutual bounds for both loss functions,
- **However, in the worst case analysis, we will show that the regret is high** ...

The **regret** of a **classifier** $\boldsymbol{h}$ with respect to a **loss function** L_z is defined as:

$$r_{L_z}(\boldsymbol{h}) = \mathbb{E}_{\mathbf{XY}} L_z(\mathbf{Y}, \mathbf{h}(\mathbf{X})) - \mathbb{E}_{\mathbf{XY}} L_z(\mathbf{Y}, \mathbf{h}^*_z(\mathbf{X})),$$

where expectation is taken over the joint distribution $\mathbf{P}(\mathbf{X}, \mathbf{Y})$, and $\boldsymbol{h}^*_z$ is the Bayes-optimal classifier with respect to the loss function L_z.

Since both loss functions are decomposable with respect to individual instances, we analyze the expectation of $\mathbf{Y}$ for a given $\boldsymbol{x}$.

Proposition (Regret for subset 0/1 loss)

The following **upper bound** *holds:*

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

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

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

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

Proposition (Regret for Hamming loss)

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

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

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

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

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

Loss Functions

Summary:

- The risk minimizers of Hamming and subset 0/1 loss have a different structure: marginal mode vs. joint mode.
- Under specific conditions, these two types of loss minimizers are provably equivalent.
- These loss functions are mutually upper-bounded.
- Minimization of the subset 0/1 loss may cause a high regret for the Hamming loss and vice versa.

Let us empirically confirm theoretical results . . .

Binary Relevance (BR)

- The simplest classifier in which a separate binary classifier $h_i(\cdot)$ is trained for each label λ_i:

$$\begin{aligned} h_i : \mathcal{X} &\rightarrow [0,1] \\ \boldsymbol{x} &\mapsto y_i \in \{0,1\} \end{aligned}$$

- It is often criticized for treating labels independently.
- However, it is still an unbiased approach for the Hamming loss (and other losses for which the marginal distribution is sufficient for obtaining the risk-minimizing model).

Label Power-set (LP)

- The method reduces the problem to multi-class classification by considering each label subset $L \in \mathcal{L}$ as a distinct meta-class:

$$
\begin{aligned}
\boldsymbol{h} : \mathcal{X} &\rightarrow [0,1]^m \\
\boldsymbol{x} &\mapsto \boldsymbol{y} \in \{0,1\}^m
\end{aligned}
$$

- It is often claimed to be a right approach to MLC, since it takes the *label dependence* into account.
- However, this approach is clearly tailored for the subset 0/1 loss.

The artificial data experiment:

- conditionally independent data,
- conditionally dependent data,
- non-linear data,
- low-dimensional problems with 2 or 3 labels,
- two classifiers: Binary Relevance (BR), Label Power-set (LP) with linear SVM as base learners.

Table: Results on two artificial data sets: conditionally independent (top) and conditionally dependent (down).

Conditional independence		
classifier	Hamming loss	subset 0/1 loss
BR	0.4208($\pm$.0014)	0.8088($\pm$.0020)
LP	0.4212($\pm$.0011)	0.8101($\pm$.0025)
B-O	0.4162	0.8016

Conditional dependence		
classifier	Hamming loss	subset 0/1 loss
BR	0.3900($\pm$.0015)	0.7374($\pm$.0021)
LP	0.4227($\pm$.0019)	0.6102($\pm$.0033)
B-O	0.3897	0.6029

B-O is the Bayes Optimal classifier.

Figure: Data set composed of two labels: the first label is obtained by a linear model, while the second label represents the XOR problem.

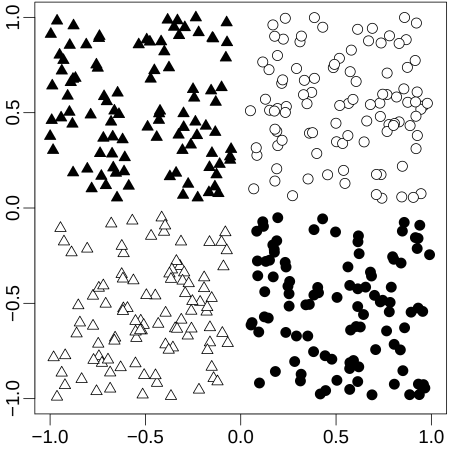


Table: Results of three classifiers on this data set.

classifier	Hamming loss	subset 0/1 loss
BR Linear SVM	0.2399($\pm$.0097)	0.4751($\pm$.0196)
LP Linear SVM	0.0143($\pm$.0020)	0.0195($\pm$.0011)
B-O	0	0

Table: Results of three classifiers on this data set.

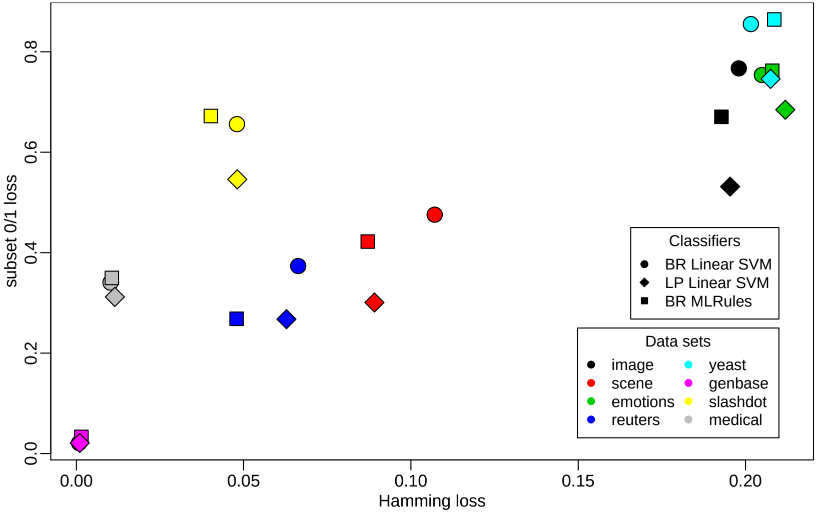
classifier	Hamming loss	subset 0/1 loss
BR Linear SVM	0.2399($\pm$.0097)	0.4751($\pm$.0196)
LP Linear SVM	0.0143($\pm$.0020)	0.0195($\pm$.0011)
BR MLRules	0.0011($\pm$.0002)	0.0020($\pm$.0003)
B-O	0	0

Experimental Settings

Summary:

- LP takes the label dependence into account, but the conditional one: it is well-tailored for the subset $0/1$ loss, but fails for the Hamming loss.
- LP may gain from the expansion of the feature or hypothesis space.
- One can easily tailor LP for solving the Hamming loss minimization problem, by marginalization of the joint probability distribution that is a by-product of this classifier.

Figure: Results of three classifiers on 8 benchmark data sets.



The experimental results on benchmark data confirm our claims.

Take-Away Message

The message to be taken home . . .

Take-Away Message

- Surprisingly, new methods are often proposed without explicitly saying what loss they intend to minimize.
- A careful distinction between loss functions seems to be even more important for MLC than for standard classification.
- One cannot expect the same MLC method to be optimal for different types of losses.
- The reasons of improvements should be carefully distinguished.