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 & Bioinformatics Lab, University of Marburg, Germany*

[2] *Research Unit Knowledge-Based Systems, Ghent University, Belgium*

Motivation

- 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
- The notion of "label dependence" is often used in a purely *intuitive* manner, without a precise *formal* definition
- The results are given *on average* without investigation under which conditions a given algorithm benefits
- The reasons for improvements are *not* carefully *distinguished*

Analysis of 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:

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

Can one of the loss functions be used as a proxy of the other?

- The risk minimizer of the Hamming loss is the *marginal mode*:

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

while for the subset 0/1 loss, it is the *joint mode*:

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

- In some situations both risk minimizers coincide, for example, if:
  - labels $Y_1, \ldots, Y_m$ are conditionally independent, i.e.,

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

  - probability of the joint mode is ≥ 0.5, i.e., $\mathbf{P}(\boldsymbol{h}_s^*(\boldsymbol{x}) \,|\, \boldsymbol{x}) \geq 0.5$
- One can also provide mutual bounds for both loss functions:

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

- *However*, one can show that the following upper bounds are tight:

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

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

what means that **minimization of the Hamming loss may cause a high regret for the subset 0/1 loss and vice versa**

Conclusions

- 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

Experimental Evidence of Theoretical Claims

Binary Relevance:

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

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

- It is often criticized for treating labels independently
- However, it is still an *unbiased approach for the Hamming loss*

Label Power-set:

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

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

- 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*

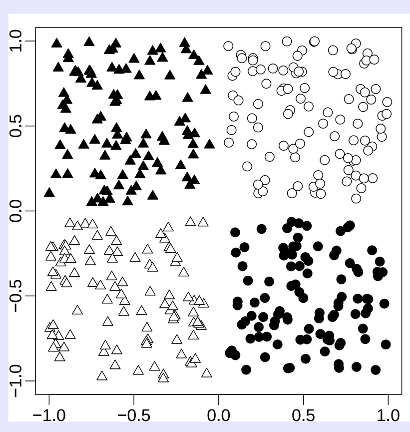
Simulations:

- Artificial data sets: conditional independence (left) and conditional dependence (right)

classifier	Hamming loss	subset 0/1 loss
BR	0.4208(±.0014)	0.8088(±.0020)
LP	0.4212(±.0011)	0.8101(±.0025)
Bayes Optimal	0.4162	0.8016

classifier	Hamming loss	subset 0/1 loss
BR	0.3900(±.0015)	0.7374(±.0021)
LP	0.4227(±.0019)	0.6102(±.0033)
Bayes Optimal	0.3897	0.6029

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



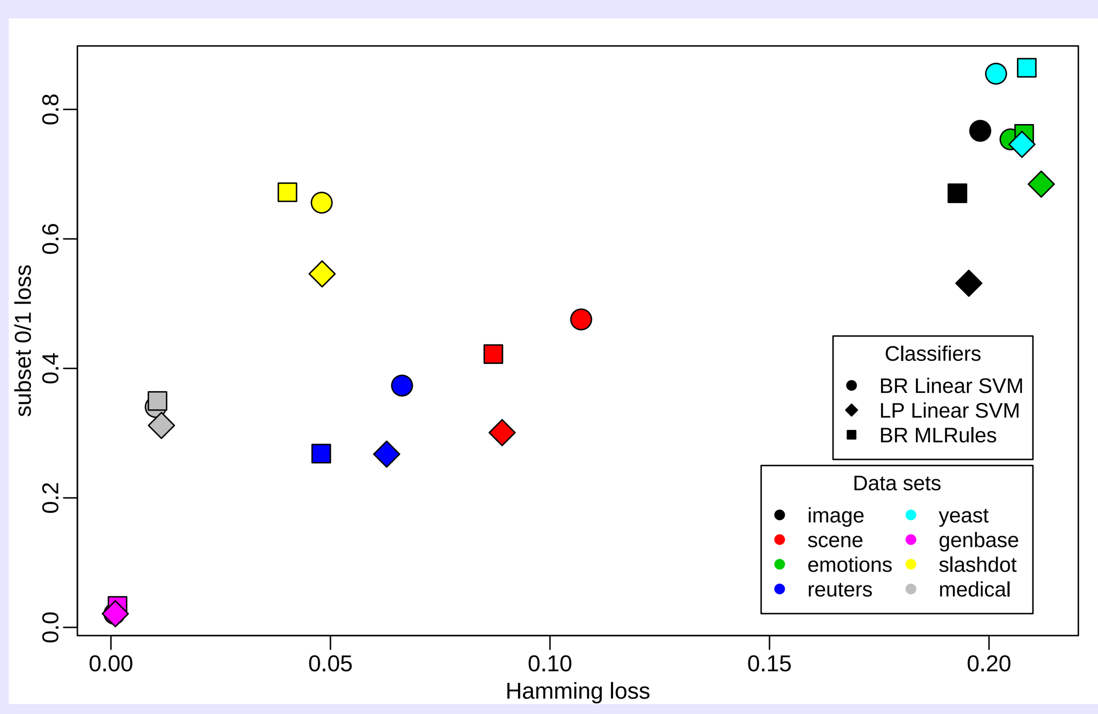
classifier	Hamming loss	subset 0/1 loss
BR Linear SVM	0.2399(±.0097)	0.4751(±.0196)
LP Linear SVM	0.0143(±.0020)	0.0195(±.0011)
BR MLRules	0.0011(±.0002)	0.0020(±.0003)
Bayes Optimal	0	0

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: the reasons of improvements should be carefully distinguished

Benchmark Data:

- The experimental results on benchmark data confirm the main claims



ECML PKDD 2010, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, Barcelona, September 2010.