Bayes Optimal Multilabel Classification via Probabilistic Classifier Chains

Krzysztof Dembczyński*, Weiwei Cheng, and Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Lab

Department of Mathematics and Computer Science

Philipps University of Marburg, Germany

Motivation

- Insufficient theoretical analysis in multi-label classification (MLC) papers
- 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*
- It is implicitly assumed that *one* algorithm is going to be beneficial for *multiple* error measures

Main Question

The risk-minimizing model $\mathbf{h}^*$ is formally defined as:

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

where $L(\mathbf{Y}, \boldsymbol{y})$ is a loss function defined on multi-label predictions.

Do we have to *take into account* the *conditional dependence* between labels in order to *obtain* a *risk-minimizing model*?

Loss Functions and Risk Minimizers

	loss function	risk minimizer
Hamming loss:	$L_H(\boldsymbol{y}, \mathbf{h}(\boldsymbol{x})) = \sum_{i=1}^{m} [\![y_i \neq h_i(\boldsymbol{x})]\!]$	$h_i^*(\boldsymbol{x}) = \arg\max_{b \in \{0,1\}} \mathbf{P}(y_i = b \,|\, \boldsymbol{x})$
Rank loss:	$L_r(\boldsymbol{y}, \mathbf{f}(\boldsymbol{x})) = \sum_{(i,j): y_i > y_j} \left([\![f_i < f_j]\!] + \frac{1}{2} [\![f_i = f_j]\!] \right)$	$f_i^*(\boldsymbol{x}) = \mathbf{P}(y_i = 1 \,|\, \boldsymbol{x})$
Subset 0/1 loss:	$L_s(\boldsymbol{y}, \mathbf{h}(\boldsymbol{x})) = [\![\boldsymbol{y} \neq \mathbf{h}(\boldsymbol{x})]\!]$	$\mathbf{h}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbf{P}(\boldsymbol{y} \,|\, \boldsymbol{x})$

Consequences and Conjectures

- The risk-minimizing prediction for the Hamming and the rank loss can be obtained from the marginal distributions $\mathbf{P}(Y_i \,|\, \boldsymbol{x})$, $i = 1, \ldots, m$, alone
- It is not necessary to know the joint label distribution $\mathbf{P}(\mathbf{Y} \,|\, \boldsymbol{x})$ on $\mathcal{Y}$ and take the conditional dependence into account
- As opposed to this, the modeling of conditional dependence has to be taken into account in order to minimize the subset zero-one loss
- In general, a specific learning and prediction strategy has to be tailored for a given performance measure.

Estimation of the Joint Distribution

- The conditional dependence can be fully described by the joint distribution
- The joint distribution enables to compute Bayes optimal predictions of any loss function

Probabilistic Classifier Chains (PCC)

Given a query instance $\boldsymbol{x}$, the (conditional) probability of each label combination $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathcal{Y}$ can be computed using the *product rule of probability*:

$$\mathbf{P}(\boldsymbol{y} \,|\, \boldsymbol{x}) = \mathbf{P}(y_1 \,|\, \boldsymbol{x}) \cdot \prod_{i=2}^{m} \mathbf{P}(y_i \,|\, \boldsymbol{x}, y_1, \ldots, y_{i-1}).$$

Algorithm:

- Learn m functions $g_i(\cdot)$ on an augmented input space $\mathcal{X} \times \{0,1\}^{i-1}$, taking $y_1, \ldots, y_{i-1}$ as additional attributes:

$$\begin{aligned} g_i : \mathcal{X} \times \{0,1\}^{i-1} &\rightarrow [0,1] \\ (\boldsymbol{x}, y_1, \ldots, y_{i-1}) &\mapsto \mathbf{P}(y_i = 1 \,|\, \boldsymbol{x}, y_1, \ldots, y_{i-1}) \end{aligned}$$

- Assuming that the function $g_i(\cdot)$ can be interpreted as a *probabilistic* classifier whose prediction is the probability that $y_i = 1$, then:

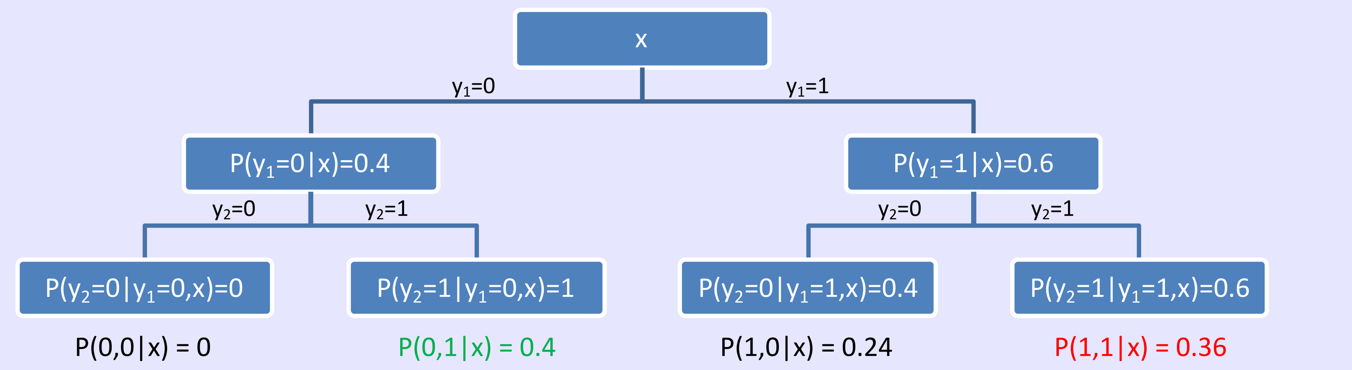
$$\hat{\mathbf{P}}(\boldsymbol{y} \,|\, \boldsymbol{x}) = g_1(\boldsymbol{x}) \cdot \prod_{i=2}^{m} g_i(\boldsymbol{x}, y_1, \ldots, y_{i-1})$$

- Given $\mathbf{P}(\boldsymbol{y} \,|\, \boldsymbol{x})$ and a loss function $L(\cdot)$ to be minimized, an optimal prediction can then be derived in an explicit way:

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

PCC vs. CC

- Original Classifier Chains (CC) (Read et al., 2009) follow a similar learning scheme
- The final prediction is computed by taking the predictions of consecutive models g_i, $i = 1, \ldots, m$
- CC can be seen as a deterministic approximation of $\mathbf{P}(\boldsymbol{y} \,|\, \boldsymbol{x})$, in the sense of using $\{0,1\}$-valued probabilities
- CC estimates the joint mode in a greedy way



Considering the chaining classifiers as searching a path in a binary tree whose leaf nodes are associated with a labeling $\boldsymbol{y} \in \mathcal{Y}$, and with 0/1 branches for y_i on level i, CC follows a single path in this tree in a greedy manner.

Simulations

- Two artificial data sets: conditionally independent and conditionally dependent data
- 10 000 training examples, 3 labels, low-dimensional problems
- Three classifiers: Binary Relevance (BR), Classifier Chains (CC), Probabilistic Classifier Chains (PCC) (all used with logistic regression as base classifier), and Bayes optimal predictions (B-O)

Results on two artificial data sets: conditionally independent (left) and conditionally dependent (right).

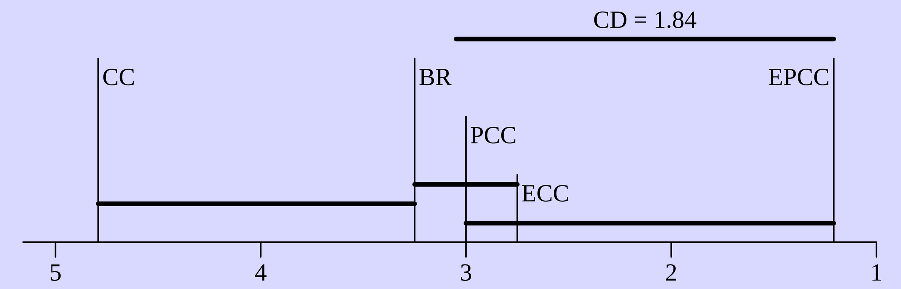
classifier	Hamming loss	rank loss	subset 0/1 loss
BR	0.4178(1.5)	0.5527(1)	0.8108(2)
CC	0.4189(3)	0.5934(3)	0.8124(3)
PCC	0.4178(1.5)	0.5528(2)	0.8088(1)
B-O	0.4179	0.5532	0.8088

classifier	Hamming loss	rank loss	subset 0/1 loss
BR	0.3921(2)	0.5675(1)	0.7374(3)
CC	0.4308(3)	0.6930(3)	0.6100(2)
PCC	0.3920(1)	0.5676(2)	0.6052(1)
B-O	0.3920	0.5671	0.6057

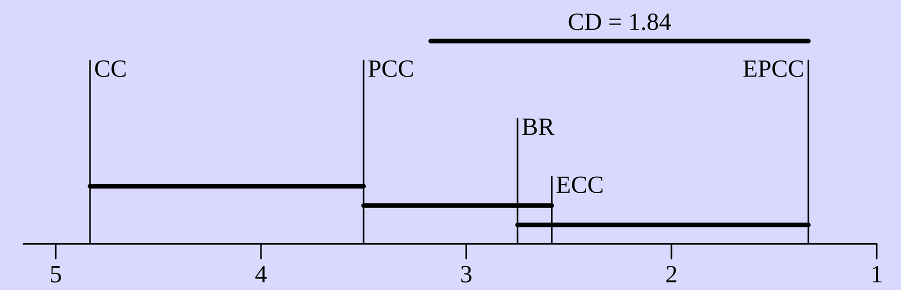
Experimental Results on Benchmark Data

- Twelve benchmark data sets taken from `http://mlkd.csd.auth.gr/multilabel.html` and `http://www.cs.waikato.ac.nz/~jmr30/#datasets`
- Five classifiers: Binary Relevance (BR), Classifier Chains (CC), Probabilistic Classifier Chains (PCC), Ensembled Classifier Chains (ECC), Ensembled Probabilistic Classifier Chains (EPCC) (all used with logistic regression as base classifier)

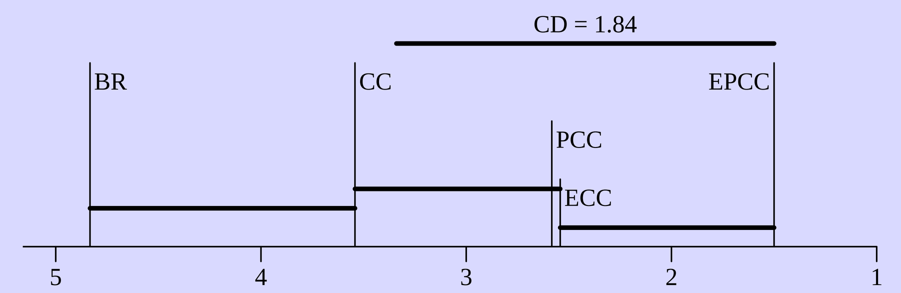
Hamming loss



Rank loss



Subset 0/1 loss



Poster presented at ICML 2010, 27th International Conference on Machine Learning, Haifa, Israel, June 2010

*Krzysztof Dembczyński is also affiliated to Intelligent Decision Support Systems Laboratory at Poznań University of Technology, Poland.