Optimizing the F-Measure in Multi-Label Classification: Plug-in Rule Approach versus Structured Loss Minimization

Krzysztof Dembczyński[1], Arkadiusz Jachnik[1], Wojciech Kotłowski[1], Willem Waegeman[2], and Eyke Hüllermeier[3]

[1] Intelligent Decision Support Systems Laboratory, Poznań University of Technology, Poland
[2] NGDATA-Europe, Belgium [3] Mathematics and Computer Science, Marburg University, Germany

Multi-Label Classification (MLC)

- For a feature vector $\boldsymbol{x}$ predict a binary vector of responses $\boldsymbol{y}$ using a prediction function $\boldsymbol{h}(\boldsymbol{x})$:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_p) \xrightarrow{\boldsymbol{h}(\boldsymbol{x})} \boldsymbol{y} = (y_1, y_2, \ldots, y_m)$$

- **Main challenges in multi-label classification:**
  - ∘ Appropriate modeling of label dependencies between labels

$$y_1, y_2, \ldots, y_m$$

  - ∘ A multitude of multivariate loss functions defined over the binary vectors

$$\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))$$

F_β-measure

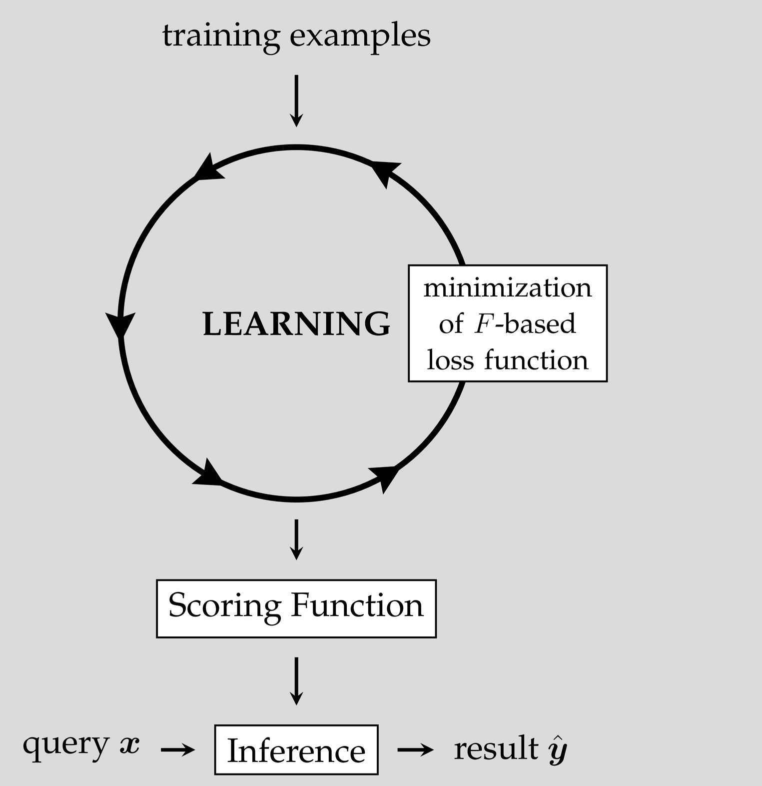
- We focus on the F_β-measure-based loss function (F_β-loss):

$$\begin{aligned}\ell_{F_\beta}(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) &= 1 - F_\beta(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) \\ &= 1 - \frac{(1+\beta^2)\sum_{i=1}^m y_i h_i(\boldsymbol{x})}{\beta^2 \sum_{i=1}^m y_i + \sum_{i=1}^m h_i(\boldsymbol{x})} \in [0,1]\,.\end{aligned}$$

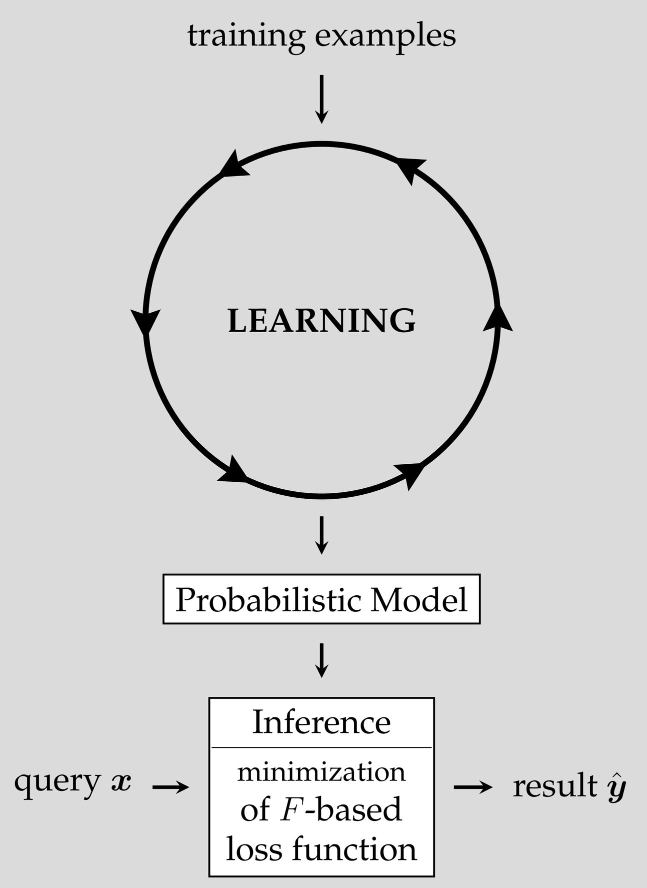
- Provides a **better balance** between relevant and irrelevant labels.
- However, it **is not easy** to optimize.

Two Approaches

Structured loss minimization



Plug-in rule approach



Structured Loss Minimization with SSVM

- Use a scoring function $f(\boldsymbol{y}, \boldsymbol{x})$
- Minimize the **structured hinge loss** [5]:

$$\tilde{\ell}_h(\boldsymbol{y}, \boldsymbol{x}, f) = \max_{\boldsymbol{y}' \in \mathcal{Y}} \{\ell_F(\boldsymbol{y}, \boldsymbol{y}') + f(\boldsymbol{y}', \boldsymbol{x})\} - f(\boldsymbol{y}, \boldsymbol{x})\,,$$

- With $\ell_F(\boldsymbol{y}, \boldsymbol{y}')$ used for margin rescaling.
- Predict according to:

$$\boldsymbol{h}(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{y}, \boldsymbol{x})\,.$$

- Requires solving the arg max and constraint generation problem.
- Two algorithms:

RML [3]
No label interactions:

$$f(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^m f_i(y_i, \boldsymbol{x})$$

Quadratic learning and linear prediction

SML [4]
Submodular interactions:

$$f(\boldsymbol{y}, \boldsymbol{x}) = \sum_{i=1}^m f_i(y_i, \boldsymbol{x}) + \sum_{y_k, y_l} f_{k,l}(y_k, y_l)$$

More complex (graph-cut and approximate algorithms)

Plug-in Rule Approaches with LR

- Plug estimates of required parameters into the **Bayes classifier**.
- The brute-force algorithm is **intractable**:

$$\boldsymbol{h}^* = \arg\min_{\boldsymbol{h} \in \mathcal{Y}} \mathbb{E}\left[\ell_{F_\beta}(\boldsymbol{Y}, \boldsymbol{h})\right] = \arg\max_{\boldsymbol{h} \in \mathcal{Y}} \sum_{\boldsymbol{y} \in \mathcal{Y}} \Pr(\boldsymbol{y}) \frac{(\beta+1)\sum_{i=1}^m y_i h_i}{\beta^2 \sum_{i=1}^m y_i + \sum_{i=1}^m h_i}$$

- Approximation needed? Not really. The exact solution is tractable!

LFP
- ∘ Assumes label independence
- ∘ Linear number of parameters: $\Pr(y_i = 1)$
- ∘ Inference based on dynamic programming [6]
- ∘ Reduction to LR for each label

EFP
- ∘ No assumptions
- ∘ Quadratic number of parameters: $\Pr(y_i = 1, s = \sum_i y_i)$
- ∘ Inference based on matrix multiplication and top k selection [1]
- ∘ Reduction to multinomial LR for each label

Theoretical Analysis

- **Computational complexity** (with respect to the number of labels):

	RML	SML	LFP	EFP
learning	$\mathcal{O}(m^2)$	$\mathcal{O}(m^4)$	$\mathcal{O}(m)$	$\mathcal{O}(m^2)$
prediction	$\mathcal{O}(m)$	$\mathcal{O}(m^3)$	$\mathcal{O}(m^2)$	$\mathcal{O}(m^3)$

- **Statistical consistency** of multi-label classifiers [2]:
  - ∘ RML and SML are **not consistent**
  - ∘ EFP is **consistent**.

Empirical Evaluation On Benchmark Datasets

	HL[%]↓	F[%]↑	t_{cv}	t_{train}	t_{inf}	HL[%]↓	F[%]↑	t_{cv}	t_{train}	t_{inf}
	IMAGE					SCENE				
BR	**19.90**	43.63	**9**	**0.392**	0.087	10.51	55.73	**29**	**0.733**	0.241
LFP	27.55	58.86	**9**	**0.392**	0.119	12.18	74.38	**29**	**0.733**	0.270
EFP	26.07	**59.77**	24	0.606	0.183	12.22	**74.44**	72	0.995	0.399
RML	25.07	57.49	94	1.104	**0.051**	**9.70**	73.92	73	1.001	**0.118**
SML	28.82	56.99	156	7.116	0.052	15.65	68.50	52	1.129	0.123
	YEAST					MEDICAL				
BR	**20.03**	60.59	**12**	**0.429**	0.128	**1.17**	70.19	**9**	**1**	0.952
LFP	22.24	65.02	**12**	**0.429**	0.146	1.18	**81.27**	**9**	**1**	1.513
EFP	22.82	**65.47**	101	2.004	0.367	1.23	80.39	16	1	1.883
RML	22.82	64.78	206	5.194	**0.056**	1.20	80.63	1253	30	**0.144**
SML	24.52	63.96	319	4.385	0.070	2.50	67.90	715	23	0.773
	ENRON					MEDIAMILL				
BR	**4.54**	55.49	**52**	**4**	1.016	**3.19**	51.21	**3238**	**118**	13
LFP	6.09	56.86	**52**	**4**	1.519	3.67	55.15	**3238**	**118**	20
EFP	5.34	**61.04**	214	6	2.628	3.63	**55.16**	24620	440	30
RML	6.35	57.69	3897	41	**0.143**	4.12	49.35	–	1125	**7**
SML	7.82	54.61	18780	62	0.887	4.18	50.02	–	10365	131

Experimental results for Hamming loss (HL), F_1, and running times (in seconds) of cross-validation (t_{cv}), training (t_{train}) (for the best set of parameters) and inference (t_{inf}). The best results are marked by a '*'. BR denotes Binary Relevance which is a baseline in the comparison.

References

[1] K. Dembczyński, W. Waegeman, W. Cheng, and E. Hüllermeier. An exact algorithm for F-measure maximization. In *NIPS*, volume 25, 2011.

[2] W. Gao and Z.-H. Zhou. On the consistency of multi-label learning. *Journal of Machine Learning Research - Proceedings Track*, 19:341–358, 2011.

[3] J. Petterson and T. S. Caetano. Reverse multi-label learning. In *Advances in Neural Information Processing Systems 24*, pages 1912–1920, 2010.

[4] J. Petterson and T. S. Caetano. Submodular multi-label learning. In *Advances in Neural Information Processing Systems 24*, pages 1512–1520, 2011.

[5] Y. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005.

[6] N. Ye, K. Chai, W. Lee, and H. Chieu. Optimizing F-measures: a tale of two approaches. In *ICML*, 2012.

This project is partially supported by the Foundation of Polish Science under the Homing Plus programme, co-financed by the European Regional Development Fund.

INNOVATIVE ECONOMY NATIONAL COHESION STRATEGY
FNP Foundation for Polish Science
EUROPEAN UNION EUROPEAN REGIONAL DEVELOPMENT FUND