A Unified View of Multi-Label Performance Measures

Xi-Zhu Wu   Zhi-Hua Zhou

[1]LAMDA Group, National Key Lab for Novel Software Technology, Nanjing University, China

{wuxz, zhouzh}@lamda.nju.edu.cn

LAMDA Learning And Mining from DatA

http://lamda.nju.edu.cn

Background

- Multi-label classification deals with the problem where each instance is associated with multiple relevant labels.

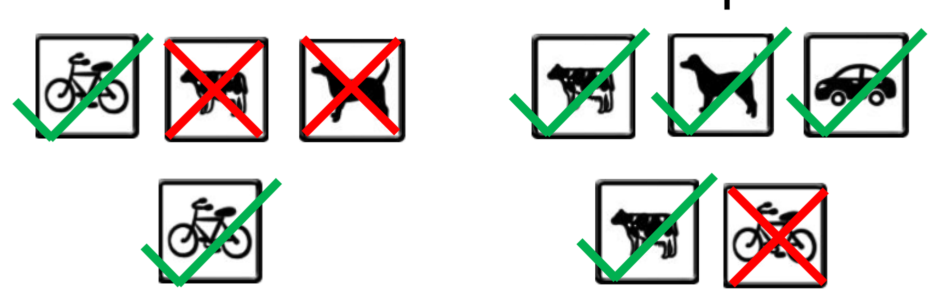
Multi-class:

Multi-label:

- Evaluation in multi-label classification is complicated.

Prediction A:

Prediction B:



  - A has more correct predictions.
  - B has less wrong predictions.

- Many performance measures are proposed to evaluate the MLC prediction. To mention a few:
  - **Hamming loss**: the fraction of misclassified labels.
  - **ranking loss**: the average fraction of reversely ordered label pairs of each instance.
  - **one-error**: the fraction of instances whose most confident label is irrelevant.
  - **coverage**: the number of more labels on average should include to cover all relevant labels.
  - **average precision**: the average fraction of relevant labels ranked higher than one other relevant label.
  - **macro-F1 / macro-AUC**: F-measure/AUC averaging on each label.
  - **instance-F1 / instance-AUC**: F-measure / AUC averaging on each instance.
  - **micro-F1 / micro-AUC**: F-measure / AUC averaging on the prediction matrix.

Contribution

- There are so many measures. We try to disclose some shared properties among different measures and established a unified margin view for multi-label performance evaluation.
- We propose two new concepts called label-wise margin and instance-wise margin to revisit eleven measures. Our theoretical results show that by maximizing each/both margin, according measures are to be optimized.
- Inspired by the theoretical findings, we design the LIMO (Label-wise and Instance-wise Margin Optimization) approach, and conduct experiments to validate our findings.

Label-wise & instance-wise margin

- Multi-label real-value predictor $F: \mathbb{R}^d \to \mathbb{R}^l, F = \{f_1, \dots, f_l\}$.
- Training set $(\boldsymbol{X}, \boldsymbol{Y})$
- The set of all the (relevant, irrelevant) label index pairs of instance i: $Y_{i\cdot}^+ \times Y_{i\cdot}^-$
- The set of all the (positive, negative) instance index pairs of label j: $Y_{\cdot j}^+ \times Y_{\cdot j}^-$
- Label-wise margin:

$$\gamma_i^{label} = \min_{u,v}\{f_u(\boldsymbol{x}_i) - f_v(\boldsymbol{x}_i) \mid (u,v) \in Y_{i\cdot}^+ \times Y_{i\cdot}^-\}.$$

- Instance-wise margin:

$$\gamma_j^{inst} = \min_{a,b}\{f_j(\boldsymbol{x}_a) - f_j(\boldsymbol{x}_b) \mid (a,b) \in Y_{\cdot j}^+ \times Y_{\cdot j}^-\}.$$

LIMO approach

- The objective function, if we use linear predictor $F = W^\top X$

$$\arg\min_{\boldsymbol{W},\xi} \sum_{i=1}^{l} \|\boldsymbol{w}_i\|^2 + \lambda_1 \sum_{i=1}^{m} \sum_{(u,v)} \xi_i^{uv} + \lambda_2 \sum_{j=1}^{l} \sum_{(a,b)} \xi_{ab}^j$$

$$\text{s.t. } \boldsymbol{w}_u^\top \boldsymbol{x}_i - \boldsymbol{w}_v^\top \boldsymbol{x}_i > 1 - \xi_i^{uv},\ \ \xi_i^{uv} \geq 0,$$

$$\text{for } i = 1, \cdots, m \text{ and } (u,v) \in Y_{i\cdot}^+ \times Y_{i\cdot}^-,$$

$$\boldsymbol{w}_j^\top \boldsymbol{x}_a - \boldsymbol{w}_j^\top \boldsymbol{x}_b > 1 - \xi_{ab}^j,\ \ \xi_{ab}^j \geq 0,$$

$$\text{for } j = 1, \cdots, l \text{ and } (a,b) \in Y_{\cdot j}^+ \times Y_{\cdot j}^-.$$

- An SGD-style algorithm is designed for optimization.

Main results

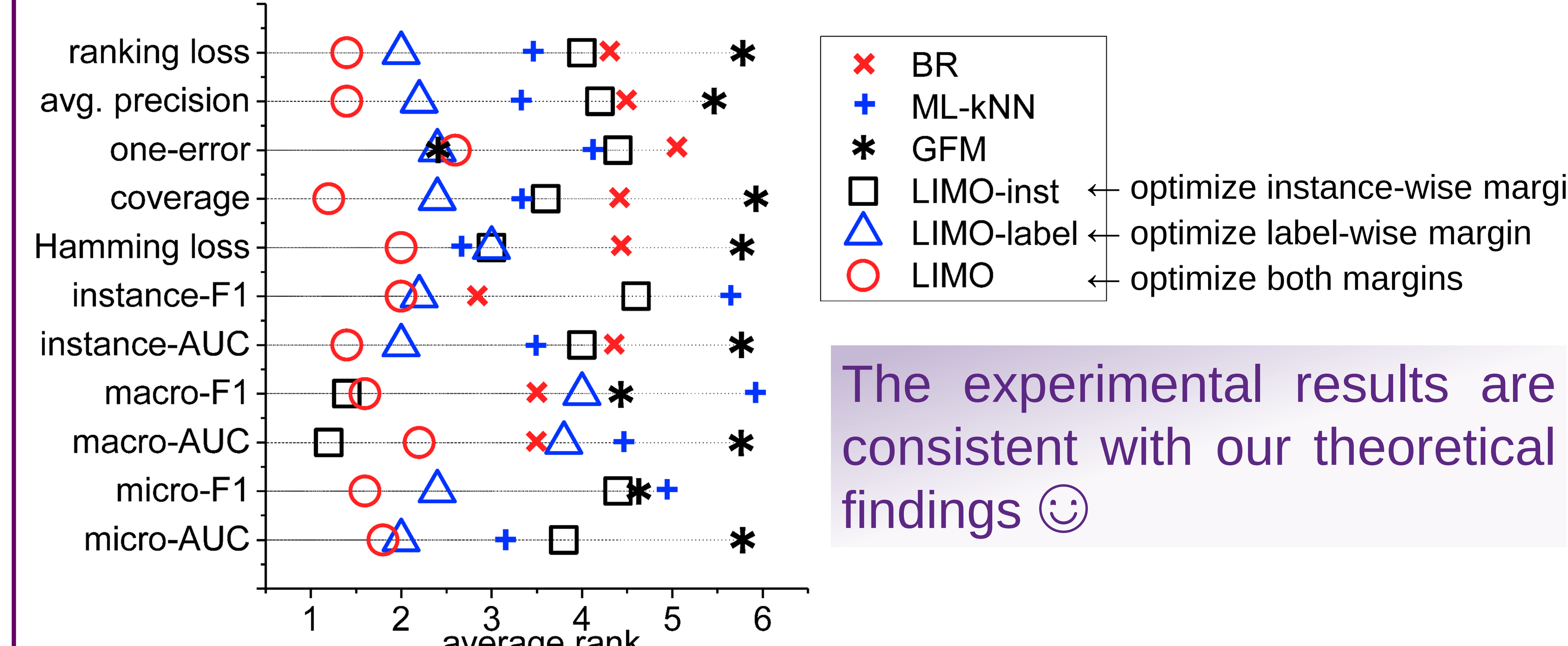
- Here is the summary table of our theoretical findings.
  - 'x-effective' means all the x margins of F on the dataset are positive. Double-effective means both the label-wise and instance-wise margins are positive;
  - '✓' means F in this cell is proved to optimize this measure;
  - '✗' means F in this cell does not necessarily optimize the measure;
  - '●'/'○' means the calculation is with/without thresholding.

Measure	x-effective F			Threshold
	label-wise	inst-wise	double	
ranking loss	✓	✗	✓	○
avg. precision	✓	✗	✓	○
one-error	✓	✗	✓	○
coverage	✓	✗	✓	○
instance-AUC	✓	✗	✓	○
macro-AUC	✗	✓	✓	○
micro-AUC	✗	✗	✓	○
macro-F1	✗	✓	✓	●
instance-F1	✓	✗	✓	●
micro-F1	✓	✗	✓	●
Hamming loss	✓	✓	✓	●

Performance measures with same combination of ✓/✗ are similar, and can be optimized by according margin(s)

Experiments

- Experiments on both synthetic data and benchmark data are conducted (results on synthetic data are omitted here).
- Benchmark datasets: CAL500, enron, medical, corel5k, bibtex.
- The smaller the average rank, the better the algorithm does.



The experimental results are consistent with our theoretical findings ☺