Multi-Label Classification with Relevance Ordering

Miao Xu **Yu-Feng Li** **Zhi-Hua Zhou**
{xum,liyf,zhouzh}@lamda.nju.edu.cn

1 Introduction

In real-world applications, one object may be associated with multiple class labels simultaneously. Multi-label learning studies such problems and many effective approaches have been developed [4] which have been found useful in diverse applications. For any multi-label task, generally one object is only associated with a subset of labels called *relevant* ones while the remaining as *irrelevant*. The first goal of multi-label learning is usually *label prediction*. While in many applications, there is also the second goal, i.e., to get a good order of the relevant labels. For example, both the two images in Figure 1 have the relevant labels *mountain*, *cattle* and *road*; if we know the relevant label orders are {*cattle*, *mountain*, *road*} and {*mountain*, *road*, *cattle*}, respectively, then when the user wishes to get *cattle* images, we will recommend the first image in front of the second one, while if the user wishes to get *road* images, we will recommend the second image in front of the first one. It is very worth noting that in such tasks, the order of the irrelevant labels are meaningless.

Figure 1: Relevant label order. Left: {*cattle*, *mountain*, *road*}; Right: {*mountain*, *road*, *cattle*}.

Though there are many effective multi-label learning algorithms, few of them can accomplish the above task well. A crucial reason lies in the fact that most algorithms were designed to optimize some multi-label criteria, but none existing criterion expresses the above requirement exactly. For example, the `Hamming Loss` considers only the label prediction while totally neglects the label orders, the `One-error` considers only the top-predicted label while totally neglects other labels, etc. Though the `Ranking Loss` considers some label order information, it only concerns about the order of each relevant-irrelevant label pair, while regards relevant labels as equally important. It is evident that to address our concerned problem, new criteria as well as new algorithms are needed.

2 The `PRO Loss` and the PROMPT Approach

Given an instance X, we define the `PRO Loss` as:

$$\widehat{\mathcal{L}}(X, R, \prec, \mathbf{g}) = \frac{1}{|R||\overline{R}|}\left|\{(a,b) : g_a < g_b, (l_a, l_b) \in R \times \overline{R})\}\right| + \frac{1}{|R|}|\{a : l_a \in R, g_a < g_\Theta\}| \quad (1)$$
$$+\frac{1}{|\overline{R}|}\left|\{b : l_b \in \overline{R}, g_b > g_\Theta\}\right| + \frac{2}{|R|(|R|-1)}|\{(a,c) : g_a < g_c, l_a \succ l_c, (l_a, l_c) \in R \times R\}|,$$

For an instance X, R denotes the relevant label set with an order $\prec$, and $\overline{R}$ denotes the irrelevant label set on which the order is not concerned. $\mathbf{g}(X)$ assigns a score $g_t(X)$ to each label $l_t \in L$, with a threshold $g_\Theta(X)$ ordered above all irrelevant labels but below all relevant ones.

The `PRO Loss` has very natural meanings. The first part is actually the `Ranking Loss`. The second and third part reassembles the `Hamming Loss` but with normalization to relevant and irrelevant labels, respectively. The last part favors more pairs of correctly ordered relevant labels.

Table 1: Comparison results on PRO Loss. Each entry presents the PRO Loss value and the rank (in parenthesis) of the method among all compared methods; the best result on each data is bolded. The last row presents the sum of ranks; the smaller the R-total, the better the overall performance.

DATA SET	PROMPT	RSVM	BSVM
EMOTIONS	**.9741(1)**	1.256(3)	1.103(2)
ENRON	**.5919(1)**	.6027(2)	1.012(3)
MSRA	**1.019(1)**	1.191(2)	1.409(3)
SCENE	**.4155(1)**	.4793(2)	1.118(3)
SLASHDOT	**.4649(1)**	.6745(2)	.8244(3)
YEAST	**.7419(1)**	.7734(2)	1.190(3)
TOTAL RANK	6	13	17

This model is comprehensible for users from application perspectives, and provides users with the flexibility to balance the different parts by setting different weights according to domain knowledge.

Instead of optimizing the difficult non-convex PRO Loss, we consider to optimize a surrogate convex loss function followed by large margin principle:

$$\min_{\mathbf{g}} \quad \lambda \sum_{i=1}^{n} \mathcal{L}(X_i, R_i, \prec, \mathbf{g}) + \Omega(\mathbf{g}), \tag{2}$$

Without losing the generality, suppose that $\mathbf{g}$ is a linear model, then Eq.2 can be cast as the following optimization problem:

$$\min_{\bar{\mathbf{w}} \triangleq [\mathbf{w}_1; \ldots; \mathbf{w}_m; \mathbf{w}_\Theta]} F(\bar{\mathbf{w}}, D) \triangleq \frac{1}{2}\|\bar{\mathbf{w}}\|^2 + \lambda \mathbf{C}^\top \boldsymbol{\xi}, \quad \text{s.t.} \quad \mathbf{A}\bar{\mathbf{w}} \geq \mathbf{1}_p - \boldsymbol{\xi}, \quad \boldsymbol{\xi} \geq \mathbf{0}_p, \tag{3}$$

We can solve Eq.3 directly by optimization software, and get the PROMPT (PRO loss Max-margin OPTimization) algorithm.

3 Experiments

We have experimented with a broad range of six data sets covering different domains. The original experimental data sets do not contain label ordering information. For MSRA, we invited volunteers to manually provide the orders for relevant labels for each object. For other data sets, we automatically simulated the relevant label ordering by aggregating predictions of benchmark approaches.

We compared PROMPT with two state-of-the-art multi-label learning approaches: RankSVM [3] and BSVM (Binary Relevance based on SVM) [1]. RankSVM was designed for Ranking Loss. It is proved in [2] that the BR(Binary Relevance) and are tailored for Hamming Loss. Table 1 shows the performances of the compared approaches according to PRO Loss.

References

[1] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Patt. Recogn.*, 37(9):1757–1771, 2004.

[2] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier. Regret analysis for performance metrics in multi-label classification: The case of hamming and subset zero-one loss. In *ECML*, 2010.

[3] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *NIPS 14*. 2002.

[4] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2010.