Learning Compact Model for Large-scale Multi-label Data
Tong Wei, Yu-Feng Li {weit, liyf}@lamda.nju.edu.cn
Department of Computer Science & Technology, Nanjing University, China

1. Multi-Label Learning

2. Motivation

In real-world applications, data sets with tens of thousands of labels and features are already becoming benchmark in many domains.

Consequently, many effective approaches [Tsoumakas et al., DMKD 2009; Zhang and Zhou, TKDE 2014] are hard to deal with LMLL (Large-scale Multi-Label Learning) data due to large storage overhead.

A popular walk around

Cloud
Edge device
Compact
model
Train model
Transmit model
Make inference

Challenges (retain competitive performance while doing compression)

3. Methodology

Problem Formulation
Given a pre-trained model $M$, the goal is to find a compact model $\tilde{M}$ with comparable performance. Such objective can be formulated as:

$$\min_M \text{ size}(M) \quad \text{s.t. perf}(M, D) \geq q^* - \epsilon$$

In this work, we focus on linear models, and the optimization problem can be rewritten as:

$$\min_{\tilde{M}} \text{ size}(\tilde{M}) \quad \text{s.t. perf}(XM, Y) \geq q^* - \epsilon$$

Joint Label and Feature Parameter Optimization (POP)

Since the resultant optimization problem is difficult, we propose to solve it from label and feature parameter optimizations jointly.

Parameter Optimization w.r.t Label

We compute the impact of labels for commonly used LMLL metrics (PSP@k and PSnDCG@k). Since missing labels commonly occur in LMLL [Bi and Kwok 2013; Xu, et. al., 2016], we capture the impact of labels under this scenario.

Parameter Optimization w.r.t Feature

We further locate the most important feature parameters for the influential labels.

4. Empirical Studies

Comparison with baseline (BR)
- Avg. model size reduction > 50%.
- Avg. performance loss < 0.5%.

Comparison with State-of-the-arts
- POP achieves top 2 results in 23/28 cases.
- Vs. DiSMEC: 10x size reduction on wiki10 and 4x reduction on eurlex.
- Vs. SLEEC: avg. 9x size reduction.
- Vs. FastXML: avg: 10x size reduction.
- Vs. LEML/CoH/PD-Sparse: POP consistently outperforms.

Parameter study
- POP filters out more than 80% model parameters when $\epsilon = 1$.
- Predictive accuracy goes down very slowly as $\epsilon$ becomes bigger.

5. Take-home Messages

Main contribution
- The impact of labels on PSP@k and P$n$DCG@k is related to the label weights and label frequencies.
- We propose the POP method to compress the model size by jointly performing label and feature parameter optimization.

Feel free to check out the code

The best and the second best results are in bold.