# Learning Compact Model for Large-scale Multi-label Data

Tong Wei, Yu-Feng Li

LAMDA Group
Nanjing University, China

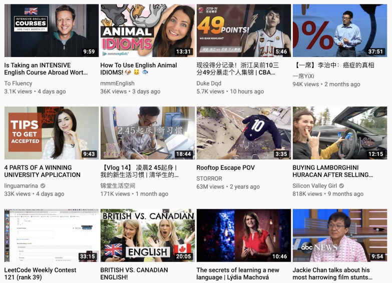
Learning And Mining from DatA
http://lamda.nju.edu.cn

February 1, 2019

# Overview

# Multi-Label Learning (MLL)

1. Multi-Label Learning aims to annotate objects with a subset of relevant labels from the entire label set.

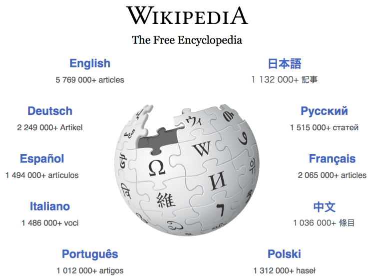


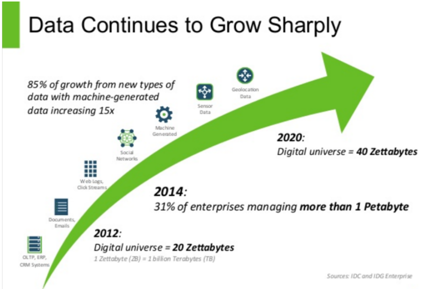

2. Multi-label objects occur in many applications, such as image tagging, recommender systems and document categorization.

# Large-scale Multi-Label Learning (LMLL)

## Goal

Learn a function $f(\mathbf{x}) : \mathbb{R}^D \to \mathbb{R}^K$ from $D$ input features to $K$ output scores that is consistent with labels $\mathbf{y} \in \{0, 1\}^K$, $K$ is large.




**Challenge:** High-dimensionality of the label space

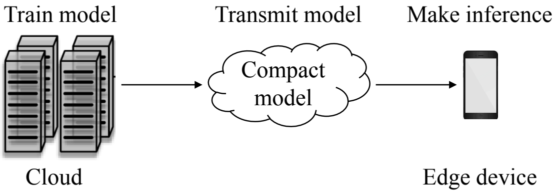
(Wikipedia Dataset: $N \approx 10^6$, $D \approx 10^6$, $K \approx 10^6$)

# Motivation

## Background

Many effective approaches *[Tsoumakas et al.,. 2009; Zhang and Zhou TKDE'14; Babbar and Schölkopf, WSDM'17; arXiv'18]* are hard to deal with LMLL data due to large storage overhead.

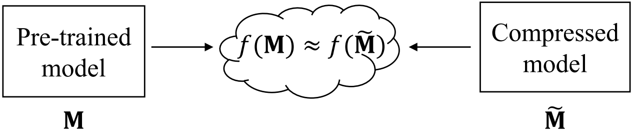
- A popular walk-around



# Motivation

## Background

Many effective approaches *[Tsoumakas et al.,. 2009; Zhang and Zhou TKDE'14; Babbar and Schölkopf, WSDM'17; arXiv'18]* are hard to deal with LMLL data due to large storage overhead.




## The task of model compression

1. compress model size as much as possible
2. maintain competitive performance

# Related Work

## Model Weights Pruning

Previous work *[Babbar and Schölkopf, WSDM'17, Niculescu-Mizil and Abbasnejad, AISTATS'17]* filter out spurious features parameters to reduce model size.

## Label Selection

Label selection methods aim to select a small subset of labels that can approximately span the original label space and subsequently model size is reduced. *[Boutsidis et al., SODA'09; Bi and Kwok, ICML'13; Weston et al., KDD'13; Niculescu-Mizil and Abbasnejad, AISTATS'17]*.

However, they either neglect label importance
or need to remove labels.

# Problem Formulation

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

$$\begin{aligned} \min_{\tilde{\mathbf{M}}} \quad & size(\tilde{\mathbf{M}}) \\ \text{s.t.} \quad & f(\tilde{\mathbf{M}}, \mathcal{D}) \geq q^* - \epsilon \end{aligned}$$

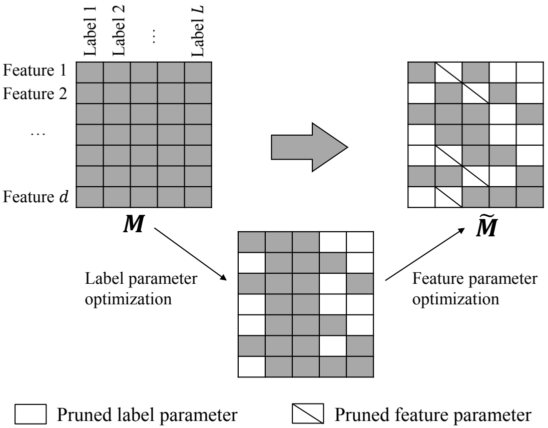
2. We consider Linear Classifier *[Babbar and Schölkopf 2017; 2018; Niculescu-Mizil and Abbasnejad 2017]*:

$$\begin{aligned} \min_{\tilde{\mathbf{M}}} \quad & ||\tilde{\mathbf{M}}||_0 \\ \text{s.t.} \quad & perf(\mathbf{X}\tilde{\mathbf{M}}, \mathbf{Y}) \geq q^* - \epsilon \end{aligned}$$

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

# Intuition

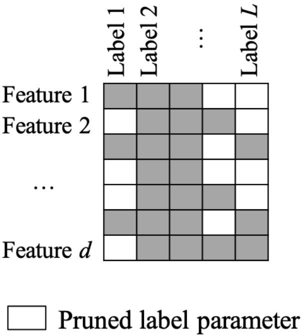
Given the pre-trained model, we propose to compress it from label and feature parameter optimization aspects jointly.



# Parameter Optimization w.r.t Label

## Naive solution

Discard part of labels may not always preferable, i.e., lose the predictive capability for some labels.




Our solution:

Step-1: identify less performance-influential labels

Step-2: preserve only a few dominant parameters (largest absolute value)

# Analysis

1. We compute the impact of labels for commonly used LMLL metrics (PSP@k and PSnDCG@k).
2. Since missing labels usually occur in LMLL, we show our results when labels are randomly missing.

**Theorem**

Suppose that relevant labels are randomly missing with probability $\pi$, the impact of the $j$-th label in terms of PSP@$k$ and PSnDCG@$k$ is upper bounded by $(1-\pi)w_j u_j$.

- $u_j$ is frequency of label $j$
- $w_j$ is the weight of label $j$.

3. In particular, when labels have equal weights, the correlation between impact of tail labels and common labels is $\frac{u_t}{u_c} \approx 0$.

# Revisit

## Main Results

1. The impact of labels on PSP@$k$ and PSnDCG@$k$ is related to label weights and label frequencies.
2. Filtering out parameters for less performance-influential labels can facilitate compact model size.

? **Challenge:** How many label parameters to trim off?

✓ **Key insight:** The performance degrades proportional to # of label parameters discarded.

# Parameter Optimization w.r.t Feature

1. We locate discriminative feature parameters and discard spurious ones.

$$\min_{\tilde{\mathbf{M}}} \ ||\tilde{\mathbf{Y}} - \mathbf{Y}^*||_F^2 + \lambda||\tilde{\mathbf{M}}||_0$$
$$\text{s.t.} \ \ \tilde{\mathbf{Y}} = \mathbf{X}\tilde{\mathbf{M}}; \mathbf{Y}^* = \mathbf{X}\mathbf{M}$$

2. Inspired by *[Zhao and Yu, JMLR'06]*, an approximate solution can be obtained by setting feature parameters that lie in range $[-\lambda, \lambda]$ to 0.

# vs. Baseline

We compare our proposed method (POP) with pure Binary Relevance (BR).

| Data set | | PSP@1 | PSP@3 | PSP@5 | PSnDCG@1 | PSnDCG@3 | PSnDCG@5 | Model size |
|---|---|---|---|---|---|---|---|---|
| bibtex | BR | 50.70 | 53.66 | 59.34 | 50.70 | 52.71 | 55.80 | 1.15 M |
| | POP | 50.71 | 53.30 | 58.86 | 50.71 | 52.39 | 55.41 | **0.59 M** |
| delicious | BR | 32.14 | 33.59 | 33.43 | 32.14 | 33.32 | 33.28 | 7.18 M |
| | POP | 32.08 | 33.59 | 33.47 | 32.08 | 33.30 | 33.29 | **1.26 M** |
| eurlex | BR | 39.93 | 45.86 | 49.74 | 39.93 | 44.24 | 46.83 | 156.38 M |
| | POP | 40.06 | 46.02 | 49.91 | 40.06 | 44.42 | 47.01 | **20.18 M** |
| wiki10 | BR | 13.57 | 13.10 | 13.96 | 13.60 | 13.82 | 13.97 | 23.50 GB |
| | POP | 13.53 | 13.10 | 13.46 | 13.53 | 13.65 | 13.67 | **67.50 M** |

- Avg. model size reduction $> 50\%$
- Avg. performance loss $< 0.5\%$

# vs. State-of-the-arts

| Data set | | FastXML | LEML | SLEEC | DiSMEC | PD-Sparse | POP |
|---|---|---|---|---|---|---|---|
| delicious | Model size | 71.29 M | 2.26 M | 7.34 M | - | **0.25 M** | **1.26 M** |
| | PSP@1 | **32.35** | 30.73 | **32.11** | - | 25.22 | 32.08 |
| | PSP@3 | **34.51** | 32.43 | 33.21 | - | 24.63 | **33.59** |
| | PSP@5 | **35.43** | 33.26 | **33.83** | - | 23.85 | 33.47 |
| | PSnDCG@1 | **32.35** | 30.73 | **32.11** | - | 25.22 | 32.08 |
| | PSnDCG@3 | **34.00** | 32.01 | 32.93 | - | 24.80 | **33.30** |
| | PSnDCG@5 | **34.73** | 32.66 | **33.41** | - | 24.25 | 33.29 |
| eurlex | Model size | 194.40 M | 34.31 M | 245.49 M | 79.86 M | 25.00 M | **20.18 M** |
| | PSP@1 | 26.62 | 24.10 | 34.25 | **41.20** | 38.28 | **40.06** |
| | PSP@3 | 34.16 | 27.20 | 39.83 | **45.40** | 42.00 | **46.02** |
| | PSP@5 | 38.96 | 29.09 | 42.76 | **49.30** | 44.89 | **49.91** |
| | PSnDCG@1 | 26.62 | 24.10 | 34.25 | **41.20** | 38.28 | **40.06** |
| | PSnDCG@3 | 32.07 | 26.37 | 38.35 | **44.30** | 40.96 | **43.55** |
| | PSnDCG@5 | 35.23 | 27.62 | 40.30 | **46.90** | 42.84 | **47.01** |
| wiki10 | Model size | **501.47 M** | 506.88 M | 924.60 M | 880.00 M | - | **67.50 M** |
| | PSP@1 | 9.80 | 9.41 | 11.14 | **13.60** | - | **13.53** |
| | PSP@3 | 10.17 | 10.07 | 11.86 | **13.10** | - | **13.10** |
| | PSP@5 | 10.54 | 10.55 | 12.40 | **13.80** | - | **13.46** |
| | PSnDCG@1 | 9.80 | 9.41 | 11.14 | **13.60** | - | **13.53** |
| | PSnDCG@3 | 10.08 | 9.90 | 11.68 | **13.20** | - | **13.65** |
| | PSnDCG@5 | 10.33 | 10.24 | 12.06 | **13.60** | - | **13.67** |

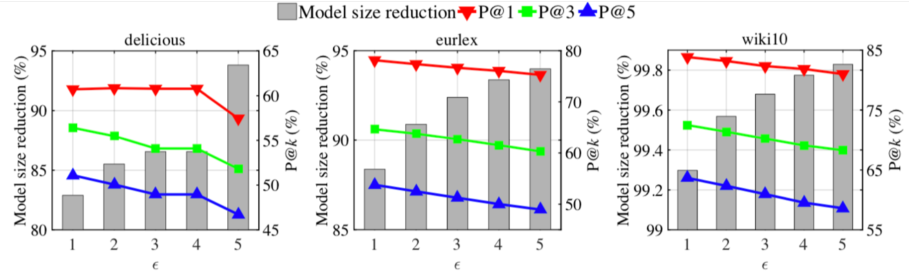
- ✓ POP achieves top 2 results in 17/21 cases.
- ✓ vs. DiSMEC: 10× smaller size on wiki10.
- ✓ vs. SLEEC: avg. 9× smaller size.
- ✓ vs. FastXML: avg: 10× smaller size.
- ✓ vs. LEML/PD-Sparse: POP consistently outperformances.

The best and the second best results are in bold.

# Parameter Sensitivity Study

## with respect to $\epsilon$

We study how different values of $\epsilon$ impact the predictive accuracy and model size.




### Observations

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

# Conclusion

- **Problem:** Learning compact model for large-scale multi-label data
- **Method:**
  - The impact of labels on PSP@$k$ and PSnDCG@$k$ is related to the label weights and label frequencies
  - We propose POP to compress the model size by jointly performing label and feature parameter optimization
- **Empirical results:**
  - Superb predictive accuracy on large-scale multi-label data
  - Much smaller model size compared with state-of-the-arts

*Thank you*

Tong Wei
weit@lamda.nju.edu.cn