Supplemental materials

Definition 1. (k,δ)-RIP. A matrix A $\in \mathbb{R}^{l\times d}$ has (k,δ)-restricted isometry property if

$$(1-\delta)\left\|\boldsymbol{y}\right\|_2^2 \leq \left\|A^T\boldsymbol{y}\right\|_2^2 \leq (1+\delta)\left\|\boldsymbol{y}\right\|_2^2 \qquad (20)$$

where $\boldsymbol{y}$ can be recovered with sparsity level k.

With a fixed sparsity level, A can be efficiently generated from appropriate distributions as a uni-variant Gaussian (Rudelson & Vershynin, 2006):

Theorem 1. *A Matrix $A \in \mathbb{R}^{l\times d}$ is (k,δ)-RIP if $A_{i,j} \sim \mathcal{N}(0,l^{-1})$, where $i \in [1,l], j \in [1,d]$, and $d = O(k\log(\frac{l}{k}))$.*

The big-O notation indicates the upper bound of the compressing rate. Intuitively, if d is set to be smaller than the bound, then it is impossible to fully recover the compressed signal $\boldsymbol{y} \in \mathbb{R}^l$ from k-sparse signal $\boldsymbol{r} \in \mathbb{R}^d$.

Pseudo Code of Multi-label Active Learning

Algorithm 1 Data sampling from unlabeled pool X_u

```
Input: $X, Y, X_u$ {training data and unlabeled pool}
Output: $z^*$ {optimal data sample}
  Conduct compressed sensing on the training label matrix $Y$
  $R \leftarrow CS(Y)$
  Conduct BPCA to generate target space $U$
  for all $x \in X$ do
    $\langle u_x \rangle = M^{-1}W_{ML}^T(r - \bar{r})$ {using eq. (3)}
  end for
  Optimize MOGP hyper-parameters using B-OPT/S-OPT
  for all $g \in 1..p$ do
    Compute the covariance matrix $C(z)_{g,g}$ for the $g$-th target
    using eq. (4)
  end for
  Select data sample $z^*$ using eq. (13)
```

Searching $\boldsymbol{\theta}_{new}$ in Simplex Optimization

The $\boldsymbol{\theta}_{new}$ is assumed to be located on the straight line passing through $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_{min}$ and the update rules of $\boldsymbol{\theta}_{new}$ can be obtained by examining the reflection point, $\boldsymbol{\theta}_{ref} = \hat{\boldsymbol{\theta}} + \beta(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{min})$ and the expansion point, $\boldsymbol{\theta}_{exp} = \boldsymbol{\theta}_{ref} + \gamma(\boldsymbol{\theta}_{ref} - \hat{\boldsymbol{\theta}})$ according to three cases:

1. (Expansion search) If $\mathcal{L}(\boldsymbol{\theta}_{max}) < \mathcal{L}(\boldsymbol{\theta}_{ref})$, then

$$\boldsymbol{\theta}_{new} = \begin{cases} \boldsymbol{\theta}_{exp} & \text{if } \mathcal{L}(\boldsymbol{\theta}_{exp}) > \mathcal{L}(\boldsymbol{\theta}_{ref}) \\ \boldsymbol{\theta}_{ref} & \text{otherwise} \end{cases}$$

2. (Reflection search) If $\min\{\mathcal{L}(\boldsymbol{\theta}_i)|\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_{min}\} < \mathcal{L}(\boldsymbol{\theta}_{ref}) < \mathcal{L}(\boldsymbol{\theta}_{max})$, then

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{ref}$$

3. (Contraction search) If $\mathcal{L}(\boldsymbol{\theta}_{ref}) < \min\{\mathcal{L}(\boldsymbol{\theta}_i)|\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_{min}\}$, then

$$\boldsymbol{\theta}_{new} = \begin{cases} \lambda\boldsymbol{\theta}_{min} + (1-\lambda)\hat{\boldsymbol{\theta}} & \text{if } \mathcal{L}(\boldsymbol{\theta}_{ref}) < \mathcal{L}(\boldsymbol{\theta}_{min}) \\ \lambda\boldsymbol{\theta}_{ref} + (1-\lambda)\hat{\boldsymbol{\theta}} & \text{otherwise} \end{cases}$$

Parameters $\beta > 0$, $\gamma > 0$, and $0 < \lambda < 1$ are used to control the rate of reflection, expansion, and contraction, respectively. The simplex method requires no parameter candidates for the search, which makes the searching range almost unbounded as the iteration goes. Each iteration in the parameter searching process is analogous to a binary search over one dimension of the searching space when $\beta = 1$, $\gamma = 1$, and $\lambda = 0.5$.

Additional Results

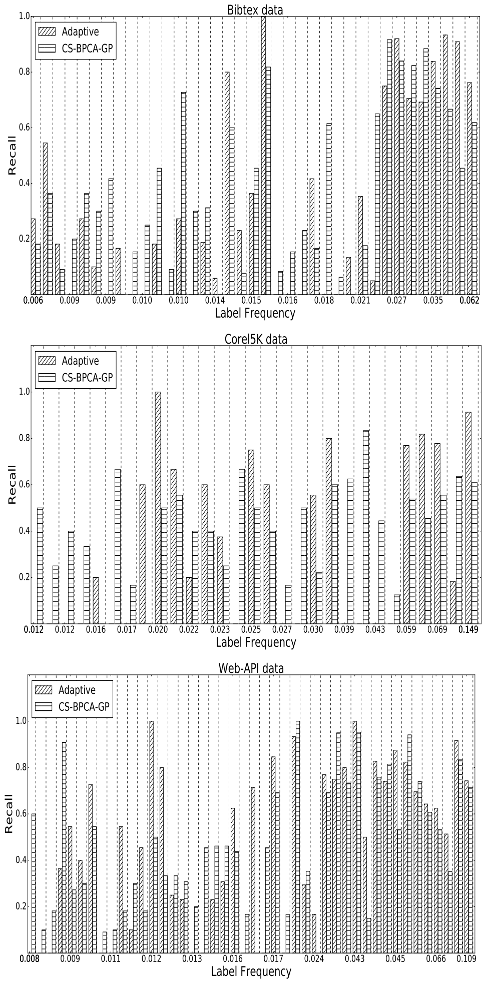

Figure 8. Rare Label Prediction Comparison

Figure 8 shows the recall performance on the three remaining datasets. In Table 3, we count the number of labels that are totally missed by each model. A label is missed by the model if the model predicts no true positives for that label. In general, rare labels can be missed by the model easily

Table 3. Missed Labels

Dataset	Adaptive	CS-BPCA-GP
Delicious	9	2
BookMark	7	1
WebAPI	13	5
Corel5K	28	17
Bibtex	18	10

as the information learned from the feature space is insufficient for model training. However, our proposed model can reduce the missing label by utilizing the label correlation as additional information for training. The result in Table 3 show that the proposed model has less number of missing labels thus is more robust in terms of multi-label prediction compared with the adaptive method.