Fast Direct Search in an Optimally Compressed Continuous Target Space for Efficient Multi-Label Active Learning

Weishi Shi and Qi Yu

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology

Jun 2019

R·I·T
B. Thomas Golisano
College of Computing and
Information Sciences

Multi-Label Active Learning

- Multi-label classification (ML-C) aims to learn a model that automatically assigns *a set* of relevant labels to a data instance.
- Multi-label problems naturally arise in many applications, including various image classification and video/audio recognition tasks.
- Data labeling for model training becomes more labor intensive as it is necessary to check each label in a potentially large label space, making active learning more important.

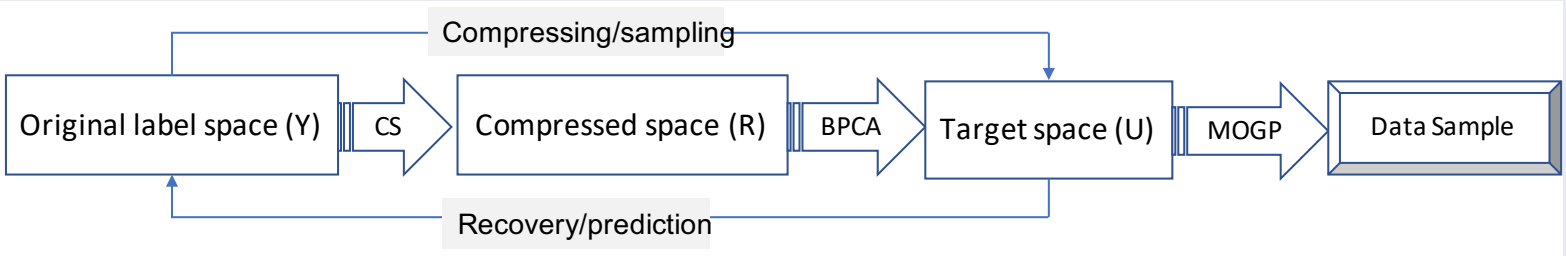
Key challenges for multi-label AL

- Sampling measure is hard to design due to label correlations.
- Rare labels are much harder to detect.
- Computational cost increases fast with the number of labels.

CS-BPCA Label Transformation

We have proposed a principled two-level label transformation (Compressed Sensing (CS) + Bayesian Principal Component Analysis (BPCA)) strategy that enables multi-label active learning to be performed in an optimally compressed target space.

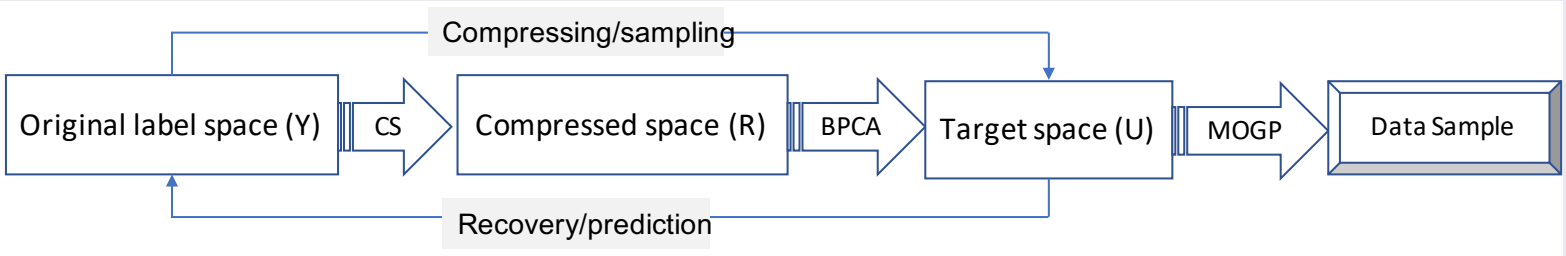
CS-BPCA: Two-level Label Transformation



CS-BPCA Label Transformation

We have proposed a principled two-level label transformation (Compressed Sensing (CS) + Bayesian Principal Component Analysis (BPCA)) strategy that enables multi-label active learning to be performed in an optimally compressed target space.

CS-BPCA: Two-level Label Transformation



Key Properties of the Transformed Label Space

- Optimally compressed: The optimal compressing rate is automatically determined.
- Orthogonal: Label correlation is fully decoupled.

Multi-output GP (MOGP) based Data Sampling

Two key benefits

- Output the predictive entropy that provides an informative measure for uncertainty based data sampling.
- Use a flexible covariance function to precisely capture the covariance structure of the input data.

A flexible kernel function

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \theta_0 \exp\{-\frac{\theta_1}{2} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\} + \theta_2 \boldsymbol{x}_i^T \boldsymbol{x}_j + \theta_3$$

Apply to the optimally compressed target space

- Continuous: Consistent with the MOGP assumption;
- Compact: Efficient computation;
- Weighted: Precise sampling;
- Orthogonality: Decoupling label correlation.

Gradient-free Hyper-parameter Optimization

High computational cost of gradient based methods

- Compute the gradient of the likelihood over each hyperparameter until convergence (via p iterations): $O|\boldsymbol{\theta}|pm^3$ [Need to run multiple times due to a non-convex likelihood].
- Construct the covariance matrix of input data: $O(m^2 n)$.

The overall complexity: $O(|\boldsymbol{\theta}|(pm^3 + m^2 n))$

Fast kernel re-estimation for covariance matrix construction

We separate two blocks of computation that are invariant to $\boldsymbol{\theta}$ and only partially update the kernel matrix for fast covariance matrix construction.

$$O(m^2 n) \longrightarrow O(m^2)$$

Gradient-free Hyper-parameter Optimization

Bayesian Optimization (B-OPT)

- Use expected improvement as a cheap surrogate of the likelihood to choose a candidate θ from the grid search space.
- Need to define a grid search space.

Simplex Optimization (S-OPT)

- Explore the search space by evolving (i.e., expanding, reflecting, and contracting) a simplex.
- Automatically explore the search space.

Overall Complexity Reduction

$O(|\theta|(pm^3 + m^2 n)) \longrightarrow O(qm^3 + m^2)$ where $q \ll p$

Benchmark Datasets and Compared Models

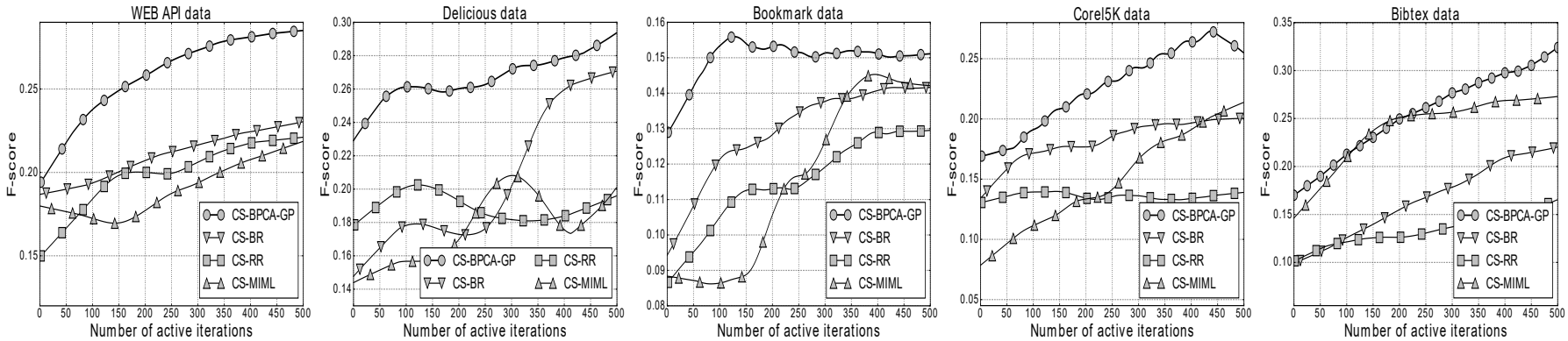
Summary of Datasets

Dataset	Domain	Instances	Features	Labels	Label Card	Label Sparsity
Delicious	web	8172	500	157	5.56	0.03
BookMark	publication	38548	2150	136	3.45	0.02
WebAPI	software	9166	5659	90	2.50	0.02
Corel5K	images	5000	499	132	3.25	0.02
Bibtex	text	7013	1836	127	2.4	0.02

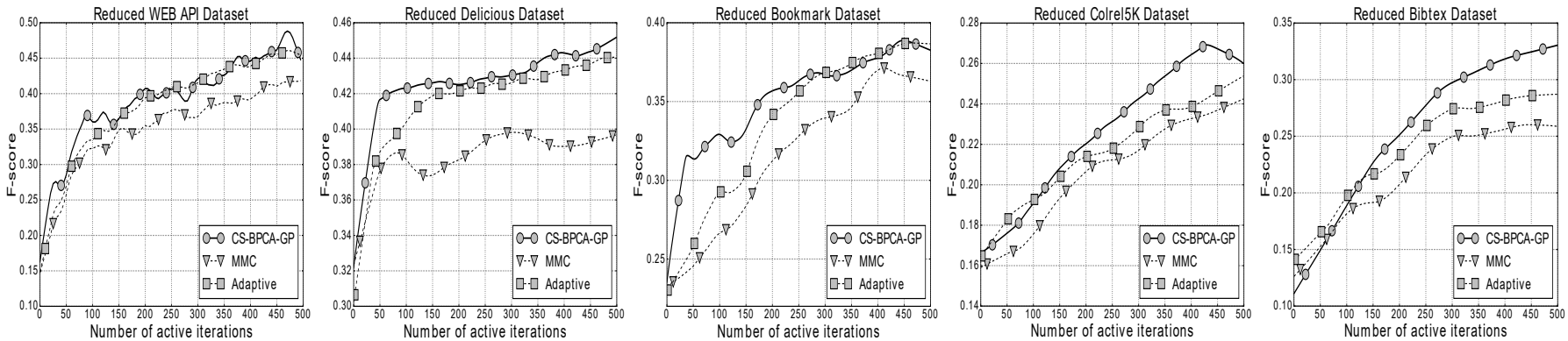
Competitive Active Learning Models for Multi-label Classification

- **Type I models**: Perform active learning in a compressed label space (CS-MIML, CS-BR, CS-RR).
- **Type II models**: Perform active learning in the original label space (MMC, Adaptive).

Comparison Results

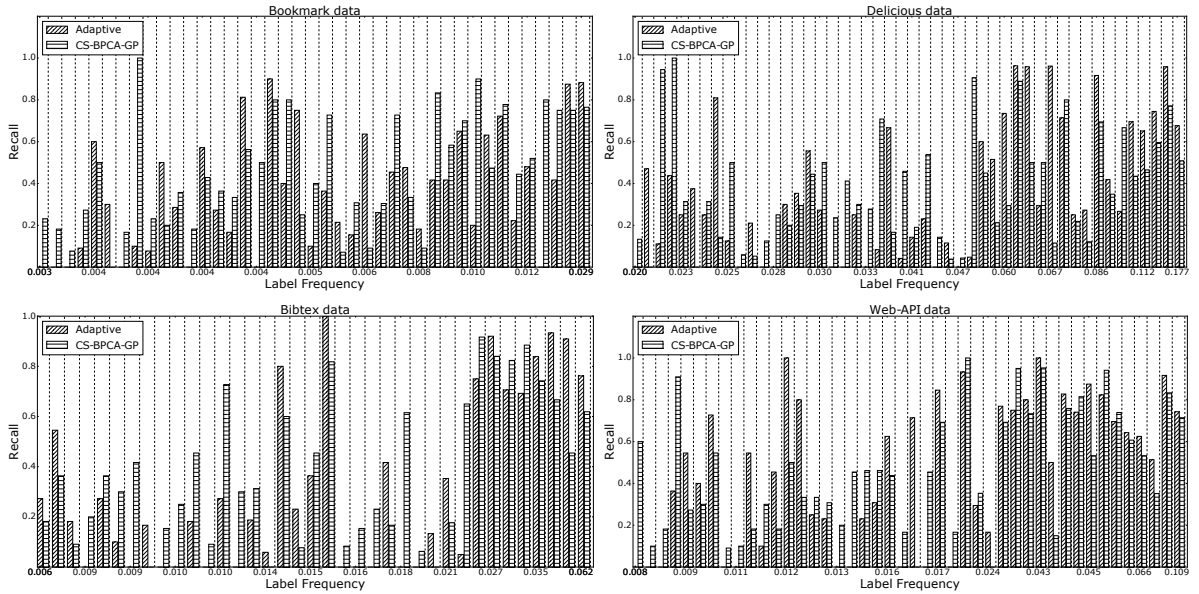


Comparison Result I



Comparison Result II

Rare Label Prediction Comparison



Missed Labels

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

Rare Label Prediction Comparison

The proposed model is effective at detecting rare labels by leveraging label correlation.

CPU Time of Hyper-parameter Optimization

Dataset	GA	B-OPT	S-OPT
Delicious	1.83	0.17	0.20
BookMark	15.0	0.80	0.79
WebAPI	10.10	0.54	0.55
Corel5K	0.58	0.08	0.08
Bibtex	8.71	0.48	0.51

The proposed direct search methods learn the kernel parameters $10 \sim 15$ times faster than the gradient based methods.

Conclusions

- Propose a two-level CS-BPCA process to generate an optimally compressed, weighted, orthogonal, and continuous target space to support multi-label data sampling.
- Propose an MOGP based sampling function that accurately captures the covariance structure of the input data.
- Propose gradient-free hyper-parameter optimization to enable fast online active learning.
- Apply to real-world multi-label datasets from diverse domains to evaluate the effectiveness of the proposed model.

Poster

Poster ID: 261