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
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.
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

Original label space (Y) $\xrightarrow{\text{CS}}$ Compressed space (R) $\xrightarrow{\text{BPCA}}$ Target space (U) $\xrightarrow{\text{MOGP}}$ Data Sample

Compressing/sampling

Recovery/prediction
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.

**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

\[ k(x_i, x_j) = \theta_0 \exp\left\{-\frac{\theta_1}{2} \|x_i - x_j\|^2\right\} + \theta_2 x_i^T x_j + \theta_3 \]

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): $O(|\theta| pm^3$ [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: $O(|\theta| (pm^3 + m^2 n))$

**Fast kernel re-estimation for covariance matrix construction**

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

$$O(m^2 n) \rightarrow O(m^2)$$
**Gradient-free Hyper-parameter Optimization**

### Bayesian Optimization (B-OPT)
- Use expected improvement as a cheap surrogate of the likelihood to choose a candidate $\theta$ 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

$$O(|\theta|(pm^3 + m^2n)) \rightarrow O(qm^3 + m^2) \text{ where } q \ll p$$
Benchmark Datasets and Compared Models

### Summary of Datasets

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

#### 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

The proposed model is effective at detecting rare labels by leveraging label correlation.
# CPU Time of Hyper-parameter Optimization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GA</th>
<th>B-OPT</th>
<th>S-OPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delicious</td>
<td>1.83</td>
<td>0.17</td>
<td>0.20</td>
</tr>
<tr>
<td>BookMark</td>
<td>15.0</td>
<td>0.80</td>
<td>0.79</td>
</tr>
<tr>
<td>WebAPI</td>
<td>10.10</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>Corel5K</td>
<td>0.58</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>Bibtex</td>
<td>8.71</td>
<td>0.48</td>
<td>0.51</td>
</tr>
</tbody>
</table>

The proposed direct search methods learn the kernel parameters 10 ~ 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.