Active Learning for Sparse Bayesian Multilabel Classification

Deepak Vasisht
MIT

Andreas Domianou
University of Sheffield, UK

Manik Varma
Microsoft Research, India

Ashish Kapoor
Microsoft Research, Redmond

Multilabel Classification

Given a set of datapoints, the goal is to choose the subset of labels that best describe the data point.

Datapoints: \( x_i \in X, \quad i = 1, 2, 3 \ldots N \)

Labels: \( y_i \in \{0, 1\}^L \)

Train Sky Building Sea Mountain Person

Obtaining training data for multilabel classification is hard:
- There can be thousands of labels
- Each label is expensive. Example: biological data.

Active Learning

The goal of active learning is to select a set of \( n \) points, \( A \), to label from a set of unlabeled pool, \( U \), such that the resulting classifier is the most accurate.

Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Type</th>
<th>Instances</th>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yeast</td>
<td>Biology</td>
<td>2417</td>
<td>103</td>
<td>14</td>
</tr>
<tr>
<td>MSRC</td>
<td>Image</td>
<td>591</td>
<td>1024</td>
<td>23</td>
</tr>
<tr>
<td>Medical</td>
<td>Text</td>
<td>978</td>
<td>1449</td>
<td>45</td>
</tr>
<tr>
<td>Enron</td>
<td>Text</td>
<td>1702</td>
<td>1091</td>
<td>53</td>
</tr>
<tr>
<td>Mediawiki</td>
<td>Video</td>
<td>43907</td>
<td>129</td>
<td>101</td>
</tr>
<tr>
<td>RCV1</td>
<td>Text</td>
<td>6000</td>
<td>47236</td>
<td>101</td>
</tr>
<tr>
<td>Bookmarks</td>
<td>Text</td>
<td>87856</td>
<td>2150</td>
<td>208</td>
</tr>
<tr>
<td>Delicious</td>
<td>Text</td>
<td>16105</td>
<td>500</td>
<td>983</td>
</tr>
</tbody>
</table>

Datasets

Traditional Active Learning

The goal is to choose the best subset of datapoints to annotate.

Active Diagnosis

Given a datapoint, select labels to annotate

Generalized AL

Choose both datapoint-label pairs to annotate

Computational Cost

4 core, Intel-i7, 3.4 GHz Processor 16 GB RAM

Classification Model \([1]\)

\[
W \sim \prod_{i=1}^k N(w_j; 0, I)
\]

\[
f_{x_i}(W, z_i) = e^{-\frac{|W^T x_i - z_i|^2}{2\sigma^2}}
\]

\[
g_{x_i}(y_i, z_i) = e^{-\frac{|\Phi y_i - z_i|^2}{2\sigma^2}}
\]

\[
P(Y, Z, W, [\alpha_i]_{i=1}^N | X, \Phi) = \frac{1}{Z}p(W) \prod_{i=1}^N f_{x_i}(W, z_i)g_{x_i}(y_i, z_i)p(h_{\alpha_i}(y_i)p(\alpha_i)
\]

Inference

\[
p(Y, [\alpha_i]_{i=1}^N | X, \Phi) = \int_{Z, W} p(Y, Z, W, [\alpha_i]_{i=1}^N | X, \Phi)
\]

\[
= \frac{1}{Z} e^{-\frac{\gamma^T \Sigma^{-1} \gamma}{2} - \sum_{i=1}^N \ln h_{\alpha_i}(y_i)p(\alpha_i)}
\]

Exact inference is not tractable \( \Rightarrow \) Use Variational Bayes approximation. Approximate posterior over \( Y \) as a Gaussian and posterior over \( \alpha \) to be Gamma distribution. Iterate across the two as follows (\( q \) denoted distribution at iteration \( t \)):

\[
q^{t+1}(Y) : \Sigma_y^{t+1} = [\text{diag}(E(\alpha_i^t)) + \Sigma_y^{-1}]^{-1}
\]

\[
q^{t+1}(\alpha) : a_i^{t+1} = a_i^0 + 0.5; b_i^{t+1} = b_i^0 + 0.5[\Sigma^{t+1}(i, i)]
\]

Mutual Information

Entropy (H): For a random variable, \( X \), \( H(X) = \sum_{i=1}^n -p(x_i) \log p(x_i) \)

Mutual Information (MI): For \( A, B \subseteq X, MI(A, B) = H(A) - H(A | B) \)

Mutual information measures reduction in uncertainty.

Goal: Select \( A^* \) such that \( A^* = \arg_{A \subseteq U} \max_{|A|=\ell} MI(Y_{\bar{A}}, A) \)

Proposition 1: The subset selection problem defined above is NP complete.

Proposition 2: \([2]\): Let \( S, U \) be disjoint sets of random variables such that the variables in \( S \) are independent given \( U \), then information gain, \( I \), defined as \( I(A) = H(U \setminus A) - H(U \setminus A | A) \), where \( A \subseteq U \), then \( F \) is submodular and non-decreasing on \( U \) and \( F(\Phi)=0 \)

Corollary 1: Let \( S = [\alpha_i]_{i=1}^N \) and \( U = Y_U, \), then the mutual information objective is sub-modular and non-decreasing.

Theorem 1: Let \( \tilde{M}I \) denote the mutual information computed using the probability distribution \( p(Y) \sim e^{-\frac{\gamma^T \Sigma^{-1} \gamma}{2}} \), then for any \( x \in U \)

\[
\lim_{a_0 \to 0, b_0 \to 0} MI(A \cup x) - MI(A) = \tilde{M}I(A \cup x) - \tilde{M}I(A)
\]

Greedy Approximation: Iteratively choose

\[
x^* = \arg \max_x (\tilde{M}I(A \cup x) - \tilde{M}I(A))
\]

References:
