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: $x_i \in X \quad i = 1, 2, 3...N$

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*.

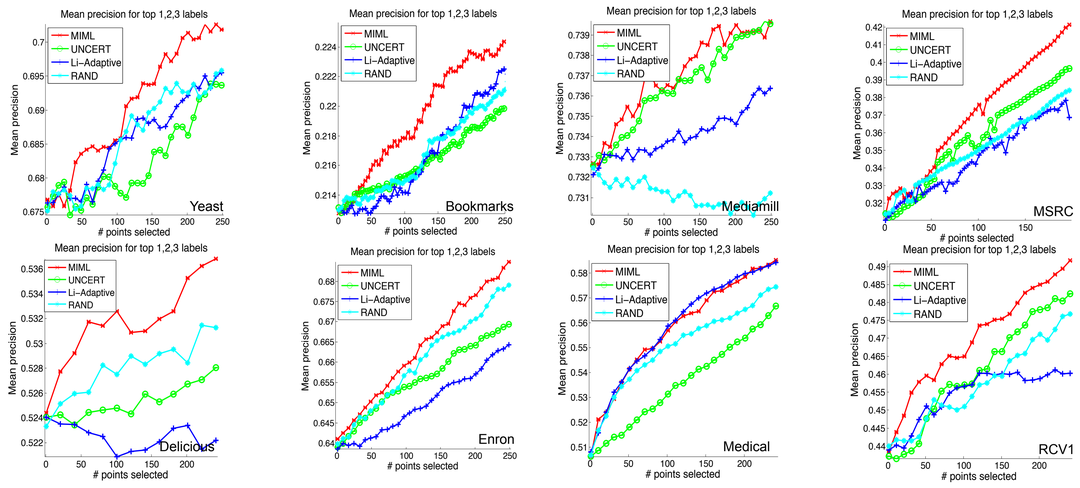
Results

Datasets

Dataset	Type	Instances	Features	Labels
Yeast	Biology	2417	103	14
MSRC	Image	591	1024	23
Medical	Text	978	1449	45
Enron	Text	1702	1001	53
Mediamill	Video	43907	129	101
RCV1	Text	6000	47236	101
Bookmarks	Text	87856	2150	208
Delicious	Text	16105	500	983

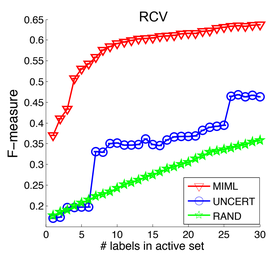
Traditional Active Learning

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



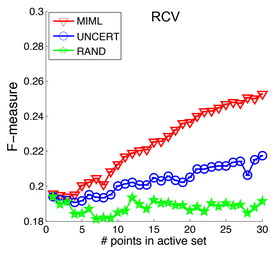
Active Diagnosis

Given a datapoints, select labels to annotate



Generalized AL

Choose both datapoint-label pairs to annotate



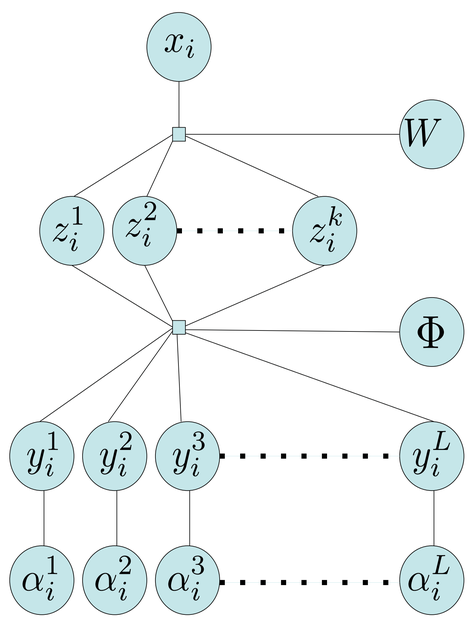
Computational Cost

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

Dataset	MIML	Li-Adaptive
Yeast	3m 25s	1m 54s
Mediamill	41m 29s	54m 35s
RCV1	30m 45s	37m 35s
Bookmarks	48m 58s	3h 57 m
Delicious	1h 11m	20h 15m

References:
[1] A Kapoor, R Viswanathan, and P. Jain. Multilabel Classification Using Bayesian Compressed Sensing, NIPS 2012
[2] A Krause, and C. Guesterin. Near-optimal Nonmyopic Value of Information in Graphical Models, UAI, 2005

Classification Model[1]



$$\mathbf{W} \sim \prod_{i=1}^{k} N(\mathbf{w}_j; 0, \mathbf{I})$$

$$f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i) = e^{\frac{-||\mathbf{W}^T\mathbf{x}_i - \mathbf{z}_i||^2}{2\sigma^2}}$$

$$g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i) = e^{\frac{-||\mathbf{\Phi}\mathbf{y}_i - \mathbf{z}_i||^2}{2\chi^2}}$$

$$y_i^j \sim N(0, \frac{1}{\alpha_i^j})$$

$$\alpha_i^j \sim \Gamma(\alpha_i^j; a_0, b_0)$$

$$P(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, [\alpha_i]_{i=1}^N | \mathbf{X}, \mathbf{\Phi}) = \frac{1}{Z} p(\mathbf{W}) \prod_{i=1}^{N} f_{\mathbf{x}_i}(\mathbf{W}, \mathbf{z}_i) g_{\mathbf{\Phi}}(\mathbf{y}_i, \mathbf{z}_i) h_{\alpha_i}(\mathbf{y}_i) p(\alpha_i)$$

Inference

$$p(\mathbf{Y}, [\alpha_i]_{i=1}^N | \mathbf{X}, \mathbf{\Phi}) = \int_{\mathbf{Z},\mathbf{W}} p(\mathbf{Y}, \mathbf{Z}, \mathbf{W}, [\alpha_i]_{i=1}^N | \mathbf{X}, \mathbf{\Phi})$$
$$= \frac{1}{Z} e^{\frac{-\mathbf{Y}^T \mathbf{\Sigma}_\mathbf{Y}^{-1} \mathbf{Y}}{2}} \prod_{i=1}^{N} h_{\alpha_i}(\mathbf{y}_i) p(\alpha_i)$$

Exact inference is not tractable => Use Variational Bayes approximation.

Approximate posterior over **Y** as a Gaussian and posterior over α to be Gamma distribution. Iterate across the two as follows (q^t denoted distribution at iteration t):

$$q^{t+1}(\mathbf{Y}) : \mathbf{\Sigma_Y}^{t+1} = [diag(\mathbb{E}(\alpha^t)) + \mathbf{\Sigma_Y}^{-1}]^{-1}$$
$$q^{t+1}(\alpha) : a_i^{t+1} = a_i^0 + 0.5; b_i^{t+1} = b_i^0 + b_i^0 + 0.5[\mathbf{\Sigma}^{t+1}(i,i)]$$

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 $\mathcal{A}^* = \arg_{\mathcal{A} \subset \mathcal{U}} \max_{|\mathcal{A}|=n} H(\mathbf{Y}_{\mathcal{U}\setminus\mathcal{A}}) - H(\mathbf{Y}_{\mathcal{U}\setminus\mathcal{A}} | \mathbf{Y}_\mathcal{A})$

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\A)-H(U\A|A), where A $\subseteq$ U, then F is submodular and non-decreasing on U and F(ϕ)=0

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

Theorem 1: Let $\hat{MI}$ denote the mutual information computed using the probability distribution $p(\mathbf{Y}) \propto e^{-\frac{\mathbf{Y}^T \mathbf{\Sigma_Y}^{-1} \mathbf{Y}}{2}}$, then for any $x \in \mathcal{U}$

$$\lim_{a_0 \to 0, b_0 \to 0} MI(\mathcal{A} \cup x) - MI(\mathcal{A}) = \hat{MI}(\mathcal{A} \cup x) - \hat{MI}(\mathcal{A})$$

Greedy Approximation: **Iteratively choose**

$$x^* = \arg\max_x (\hat{MI}(A \cup x) - \hat{MI}(A))$$