Multiclass Active Learning for Graph-Based Semi-Supervised Learning

Kevin Miller and Andrea L. Bertozzi
Department of Mathematics, University of California, Los Angeles

Abstract

Active learning in semi-supervised classification seeks to select points from the pool of unlabeled data to label that will in turn help to improve the underlying classifier's accuracy. Such potential for classifier improvement is quantified in a real-valued *acquisition function*. We present a "model-change" acquisition function for batch active learning in graph-based semi-supervised learning (SSL) for multiclass classification.

Graph-Based SSL Models

Given similarity graph of N nodes corresponding to input data $X = \{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$, with normalized graph Laplacian matrix $L = \mathrm{I}_N - D^{-1/2}WD^{-1/2}$ and *shifted graph Laplacian* $L_\tau = L + \tau^2 \mathrm{I}_N$. We consider the semi-supevised learning objective

$$\mathcal{J}(U;Y) := \frac{1}{2}\langle U, L_\tau U\rangle_F + \sum_{j\in\mathcal{L}} \ell(\mathbf{u}^j, \mathbf{y}^j),$$

where $\mathbf{y}^j$ is one-hot encoding of class label for $j \in \mathcal{L} \subset \{1,2,\dots,N\}$ and $\langle\cdot,\cdot\rangle_F$ is the Frobenius inner product. Loss functions that are considered:

- Multiclass Gaussian Regression (MGR): $\ell(\mathbf{s},\mathbf{t}) = \frac{1}{2\gamma^2}\|\mathbf{s}-\mathbf{t}\|_2^2$
- Cross-Entropy (CE): $\ell(\mathbf{s},\mathbf{t}) = \sum_{c=1}^{n_c} s_c \ln t_c$, where $\mathbf{s},\mathbf{t}$ represent discrete probability distributions

Instead of using the full graph Laplacian matrix, use only $M < N$ smallest eigenvalues $0 = \lambda_1 < \lambda_2 \le \dots \le \lambda_M$ of L with corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M$. Define

- $V = [\mathbf{v}_1 \ \ \mathbf{v}_2 \ \ \dots \ \ \mathbf{v}_M] \in \mathbb{R}^{N\times M}$,
- $\Lambda_\tau = \mathrm{diag}\left(\lambda_1+\tau^2, \lambda_2+\tau^2, \dots, \lambda_M+\tau^2\right) \in \mathbb{R}^{M\times M}$,
- $A \in \mathbb{R}^{M\times n_c}$ with classifier $U = VA$.

Spectral truncation objective

$$\mathcal{J}(A;Y) := \frac{1}{2}\langle A, \Lambda_\tau A\rangle_F + \sum_{j\in\mathcal{L}} \ell(\underbrace{\mathbf{e}_j^T VA}_{j^{th}\text{ row}}, \mathbf{y}^j) \tag{1}$$

Consider the variable $\mathbf{a} \in \mathbb{R}^{Mn_c}$ that is the columns of $A \in \mathbb{R}^{M\times n_c}$ successively stacked. Objective now becomes

$$\mathcal{J}(\mathbf{a};Y) := \frac{1}{2}\langle \mathbf{a}, \Lambda_\tau^{\otimes}\mathbf{a}\rangle + \sum_{j\in\mathcal{L}} \ell(P_j\mathcal{V}\mathbf{a}, \mathbf{y}^j), \tag{2}$$

- $\Lambda_\tau^{\otimes} := \mathrm{diag}(\Lambda_\tau, \Lambda_\tau, \dots, \Lambda_\tau) \in \mathbb{R}^{Mn_c\times Mn_c}$,
- $\mathcal{V} = \mathrm{diag}(V, V, \dots, V) \in \mathbb{R}^{Nn_c\times Mn_c}$,
- $P_j \in \mathbb{R}^{n_c\times Nn_c}$, projection onto the entries of j^{th} row of U.

Acknowledgements

Supported by the DOD NDSEG Fellowship, DARPA award FA8750-18-2-0066, and NSF grant DMS-1952339.

Model Change Acquisition Function

For each unlabeled index $k \notin \mathcal{L}$, suppose we had hypothetical one-hot encoded label $\mathbf{y}^k$:

$$\mathcal{J}^{+k,\mathbf{y}^k}(\mathbf{a};Y,\mathbf{y}^k) = \frac{1}{2}\langle \mathbf{a}, \Lambda_\tau^{\otimes}\mathbf{a}\rangle + \sum_{j\in\mathcal{L}\cup\{k\}} \ell(P_j\mathcal{V}\mathbf{a}, \mathbf{y}^j) = \mathcal{J}(\mathbf{a};Y) + \ell(P_k\mathcal{V}\mathbf{a}, \mathbf{y}^k).$$

Approximate how much classifier ("model") $\hat{\mathbf{a}}$ would change if were to add k to the labeled set $\mathcal{L}$:

$$\hat{\mathbf{a}}^{+k,\mathbf{y}^k} := \arg\min_{\mathbf{a}} \ \mathcal{J}^{+k,\mathbf{y}^k}(\mathbf{a};Y,\mathbf{y}^k), \qquad \text{measure the distance } \|\hat{\mathbf{a}}^{+k,\mathbf{y}^k} - \hat{\mathbf{a}}\|_2.$$

Approximate with one-step of Newton's method starting at previous minimizer $\hat{\mathbf{a}}$:

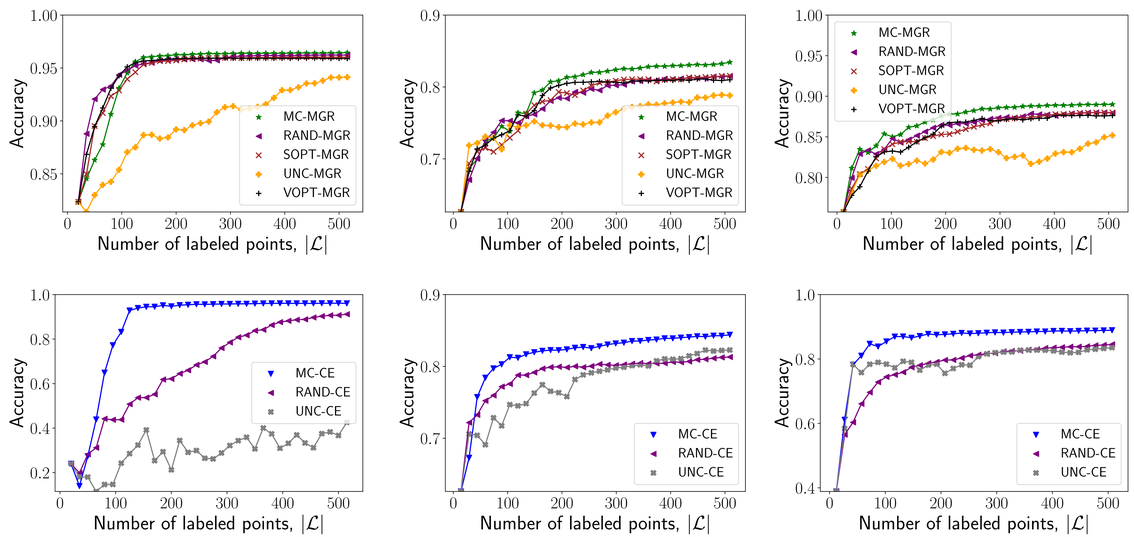
$$\begin{aligned}\hat{\mathbf{a}}^{+k,\mathbf{y}^k} &\approx \hat{\mathbf{a}} - \left(\nabla^2\mathcal{J}^{+k,\mathbf{y}^k}(\hat{\mathbf{a}};Y,\mathbf{y}^k)\right)^{-1}\left(\nabla\mathcal{J}^{+k,\mathbf{y}^k}(\hat{\mathbf{a}};Y,\mathbf{y}^k)\right)\\ &= \hat{\mathbf{a}} - \left(\nabla^2\mathcal{J}(\hat{\mathbf{a}};Y) + \nabla^2\ell(P_k\mathcal{V}\hat{\mathbf{a}},\mathbf{y}^k)\right)^{-1}\left(\underbrace{\nabla\mathcal{J}(\hat{\mathbf{a}};Y)}_{0} + \nabla\ell(P_k\mathcal{V}\hat{\mathbf{a}},\mathbf{y}^k)\right)\end{aligned}$$

$$\implies \|\hat{\mathbf{a}}^{+k,\mathbf{y}^k} - \hat{\mathbf{a}}\|_2 \approx \left\| \underbrace{\left(\nabla^2\mathcal{J}(\hat{\mathbf{a}};Y) + \nabla^2\ell(P_k\mathcal{V}\hat{\mathbf{a}},\mathbf{y}^k)\right)^{-1}}_{\text{low-rankupdate via Woodbury identity}} \nabla\ell(P_k\mathcal{V}\hat{\mathbf{a}},\mathbf{y}^k) \right\|_2. \tag{3}$$

Results

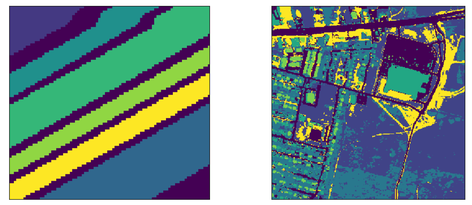
Accuracy plots for the acquisition functions for the MGR (top row) and CE (bottom row) losses for the MNIST, Salinas A, and Urban datasets. For each dataset,

- 2 points from each class are initially labeled
- 100 batches of size 5 that *maximize* acquisition functions on 10% randomly sampled subset of unlabeled points



(d) MNIST (e) Salinas A (f) Urban

HSI Datasets



(g) Salinas A (h) Urban

HSI Similarity Graph Construction:

- 15 k-nearest neighbors, using *cosine similarity* $\kappa(\mathbf{x}_i,\mathbf{x}_j) = \langle \mathbf{x}_i,\mathbf{x}_j\rangle / \|\mathbf{x}_i\|_2\|\mathbf{x}_j\|_2$
- 50 eigenvalues calculated from sparse matrix, L
- $\tau = 0.005, \gamma_{MGR} = 0.1, \gamma_{CE} = 0.5$

Acquisition Functions

- MC - Model Change (current work)
- UNC - Uncertainty Sampling[5]
- RAND - Randomly selected points
- VOPT - [1], adapted to spectral truncation
- SOPT - [2], adapted to spectral truncation

Discussion of Results

In both the MGR and CE models, the Model Change (MC) criterion outperforms the other acquisition functions:

1. Achieve higher overall accuracy
2. Faster initial increase in accuracy (e.g. Figures 1d, 1e, 1f)

References

[1] M. Ji and J. Han, *A Variance Minimization Criterion to Active Learning on Graphs*, in Artificial Intelligence and Statistics, Mar. 2012, pp. 556 -564, ISSN: 1938-7228 Section: Machine Learning.

[2] Y. Ma, R. Garnett, and J. Schneider, *Σ-Optimality for Active Learning on Gaussian Random Fields*, in Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013, pp. 2751-2759.

[3] K. Miller, H. Li, and A.L. Bertozzi, *Efficient Graph-Based Active Learning with Probit Likelihood via Gaussian Approximations*, ICML Workshop on Real-World Experiment Design and Active Learning, 2020.

[4] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning, MIT Press, Cambridge, Mass, 2006.

[5] B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning, 6, 2012.