Active Sampling on Graphs via Expected Model Change

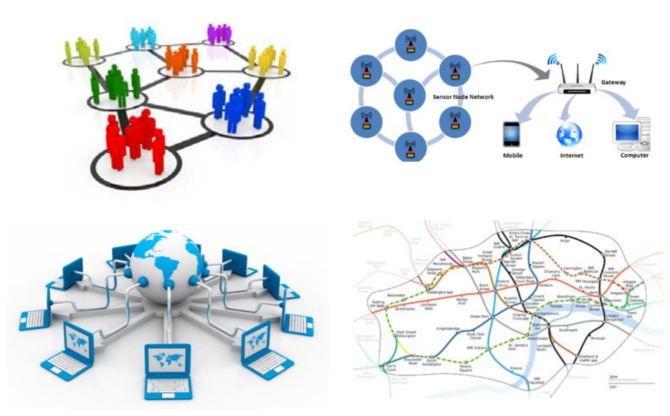
November 14, 2017

D. Berberidis and G. B. Giannakis
Dept. of ECE, University of Minnesota, 200 Union St. SE, Minneapolis, MN 55455, USA

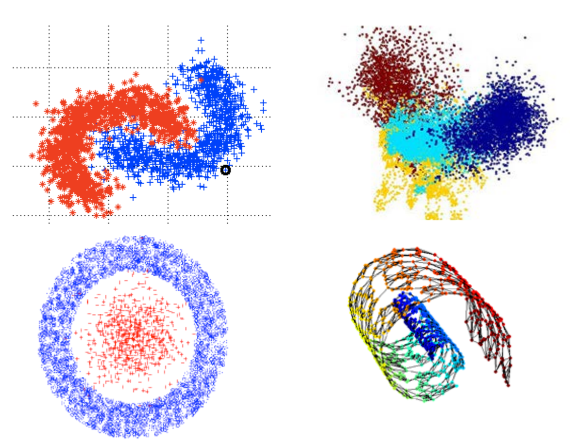
Motivation

Graph representations

Real networks



Data similarities

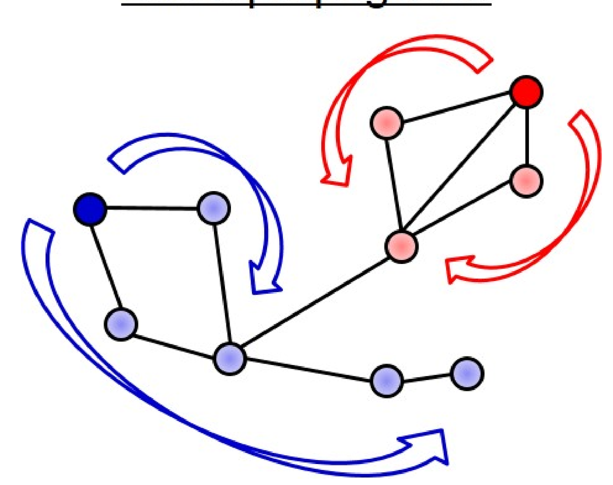


- Inference goal: Estimate values/labels defined over nodes
- Challenge: Obtaining observations often difficult/costly
  - Privacy issues, battery consumption, human annotators and other

Classification on Graphs

- Graph $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$
  - Weighted adjacency matrix $\mathbf{W}$
  - Node v_i has label $y_i \in \{-1,1\}$
- Topology (un)known
  - Given (in e.g. WSNs and social nets)
  - Identified via nodal similarities
- Semi-supervised classification (on graphs)
  - Set $\mathcal{L} \subseteq \mathcal{V}$ of labeled nodes is given
  - Labels of $\mathcal{U}=\mathcal{V}-\mathcal{L}$ are to be inferred

Label propagation



Goal: Maximize classification accuracy on $\mathcal{U}$ by actively selecting $\mathcal{L}$

Semi-supervised learning with GMRFs

- Labels modeled as a Markov Random Field (smoothness over graph)

$$p(\mathbf{Y})=\frac{1}{C(\beta)}\exp\left(-\frac{\beta}{2}\Phi(\mathbf{Y})\right) \qquad \mathbf{y}=(y_1,y_2,\ldots,y_N)$$

$$\Phi(\mathbf{y})=\sum_{i,j\in\mathcal{V}} w_{i,j}\,(y_i-y_j)^2=\mathbf{y}^T\mathbf{L}\mathbf{y} \qquad \mathbf{L}=\mathbf{D}-\mathbf{W}$$

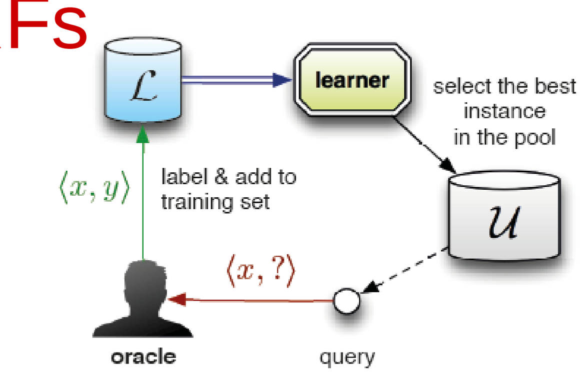
Computing marginal posteriors $p(y_i|\mathbf{y}_{\mathcal{L}})$ is NP-hard

- Unknown (discrete) labels approximated by (continuous) Gaussian field (GMRF)

$$\psi_{\mathcal{U}|\mathcal{L}}\sim\mathcal{N}(\boldsymbol{\mu}_{\mathcal{U}|\mathcal{L}},\mathbf{L}_{\mathcal{U}\mathcal{U}}^{-1}) \qquad \mathbf{L}=\begin{bmatrix}\mathbf{L}_{\mathcal{U}\mathcal{U}} & \mathbf{L}_{\mathcal{U}\mathcal{L}}\\ \mathbf{L}_{\mathcal{L}\mathcal{U}} & \mathbf{L}_{\mathcal{L}\mathcal{L}}\end{bmatrix}$$

- Conditional mean: $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{L}}=\mathbf{C}_{\mathcal{U}\mathcal{L}}\mathbf{C}_{\mathcal{L}\mathcal{L}}^{-1}\psi_{\mathcal{L}}=-\mathbf{L}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{L}_{\mathcal{U}\mathcal{L}}\boxed{\psi_{\mathcal{L}}}\rightarrow\mathbf{y}_{\mathcal{L}}$
- Predictor of unknown labels via GMRF mean: $\hat{y}_i=\begin{cases}1 & [\boldsymbol{\mu}_{\mathcal{U}|\mathcal{L}}]_i>0\\ -1 & \text{else}\end{cases}$
- Approximation of marginal posteriors: $p(y_i=1|\mathbf{y}_{\mathcal{L}}):=\frac{1}{2}\left([\boldsymbol{\mu}_{\mathcal{U}|\mathcal{L}}]_i+1\right)$

Active Sampling on GMRFs



- Greedy selection of most "informative" node

$$k_t=\arg\max_{i\in\mathcal{U}^{t-1}} U(v_i,\mathcal{L}^{t-1})$$

- GMRF mean update ($\mu_{k_t}\rightarrow y_{k_t}$)

$$\boldsymbol{\mu}^{+y_{k_t}}_{\mathcal{U}^{t-1}|\mathcal{L}^{t-1}}=\boldsymbol{\mu}_{\mathcal{U}^{t-1}|\mathcal{L}^{t-1}}+\frac{1}{g_{k_tk_t}}(y_{k_t}-[\boldsymbol{\mu}_{\mathcal{U}^{t-1}|\mathcal{L}^{t-1}}]_{k_t})\mathbf{g}_{k_t}$$

- Update Laplacian inverse $(\mathbf{G}:=\mathbf{L}_{\mathcal{U}\mathcal{U}}^{-1})$ when *k*-th node is removed
  - $\mathcal{O}(|\mathcal{U}|^2)$ complexity $\begin{bmatrix}\mathbf{G}_t^{-k_t} & \mathbf{0}\\ \mathbf{0}^T & 0\end{bmatrix}=\mathbf{G}_t-\frac{1}{g_{k_tk_t}}\mathbf{g}_{k_t}\mathbf{g}_{k_t}^T$

Key Active Sampling issue: How do we select $U(v_i,\mathcal{L})$?

Related work

- Non-adaptive approaches
  - Error upper bound minimization [Gu-Han'12]
  - GMRF variance minimization [Ji-Han'12]
  - Σ-optimal design [Ma, Garnett, Schneider'13]
- Adaptive approaches
  - Expected error (EER) minimization [Zhu et al'03]
  - EER with two-step approximation [Jun-Nowak'16]
  - Information gain maximization [Long et al'08]
  - Class boundary search [Ortega '16], [Zapella '13]

Expected model change (ECM)

- Our method: Sample node expected to inflict the largest change on label model
  - Intuition: Take larger steps to arrive faster at a "good" model
  - Various measures of change considered
- Expected number of prediction changes ("flips")

$$U_{FL}(v_i,\mathcal{L})=\mathbb{E}_{y_i|\mathbf{y}_{\mathcal{L}}}\left[F(y_i,\boldsymbol{\mu}_{\mathcal{U}|\mathcal{L}})\right] \qquad F(y_i,\boldsymbol{\mu}_{\mathcal{U}|\mathcal{L}}):=\sum_{j\in\mathcal{U}-\{i\}}\mathbf{1}_{\{\hat{y}_j^{+y_i}\neq\hat{y}_j\}}$$

- Aggregated mutual information

$$U_{KL}(v_i,\mathcal{L})=\sum_{j\in\mathcal{U}-\{i\}} I(y_j,y_i)$$

$$I(y_j,y_i)=\mathbb{E}_{y_i|\mathbf{y}_{\mathcal{L}}}\left[D_{KL}\left(y_j^{+y_i}||y_j\right)\right]$$

$$D_{KL}(y_j^{+y_i}||y_j)=H(y_j^{+y_i},y_j)-H(y_j^{+y_i})$$

Scalable with number of unlabeled nodes and classes?

EMC without retraining

- Total variation (TV) between *p(x)* and *q(x)* : $\delta(p,q)=\frac{1}{2}\sum_{x\in\mathcal{X}}|p(x)-q(x)|$
- Sum of total variations over marginal posteriors of unlabeled ($y_i\sim\mathrm{Ber}(\mu_i)$)

$$\Delta(\mathbf{y}_{\mathcal{U}}^{+y_i},\mathbf{y}_{\mathcal{U}})=\sum_{j\in\mathcal{U}}\delta(y_j^{+y_i},y_j)=\frac{|y_i-\mu_i|}{g_{ii}}\|\mathbf{g}_i\|_1$$

- Expected sum of total variations score function

$$U_{TV}(v_i,\mathcal{L})=\mathbb{E}_{y_i|\mathbf{y}_{\mathcal{L}}}\left[\Delta(\mathbf{y}_{\mathcal{U}}^{+y_i},\mathbf{y}_{\mathcal{U}})\right]=2(1-\mu_i^2)\frac{\|\mathbf{g}_i\|_1}{g_{ii}}$$

- Expected mean-square deviation (MSD) of Gaussian field yields

$$U_{MSD}(v_i,\mathcal{L})\propto(1-\mu_i^2)\frac{\|\mathbf{g}_i\|_2^2}{g_{ii}^2}$$

- TV- and MSD-based utility functions available *without model retraining*
  - Significantly faster especially for large-scale graphs

Sampling bias reduction

- Bias due to averaging over available (possibly flawed) model

$$U(v_i,\mathcal{L})=\mathbb{E}_{y_i|\mathbf{y}_{\mathcal{L}}}\left[C(y_i,\mathcal{L})\right]$$

- Possible remedies

Combine w/ random sampling: $k=\begin{cases}\arg\max_{i\in\mathcal{U}^{t-1}}U(v_i,\mathcal{L}^{t-1}), & \text{w.p.} \quad (1-p_{\mathrm{r}}^t)\\ \mathrm{Unif}\{1,\ldots,|\mathcal{L}^{t-1}|\}, & \text{w.p.} \quad p_{\mathrm{r}}^t\end{cases}$

Max-min change: $U(v_i,\mathcal{L})=\min_{y_i\in\{0,1\}} C(y_i,\mathcal{L})$

- **Our approach**: Use a convex combination between prior and model

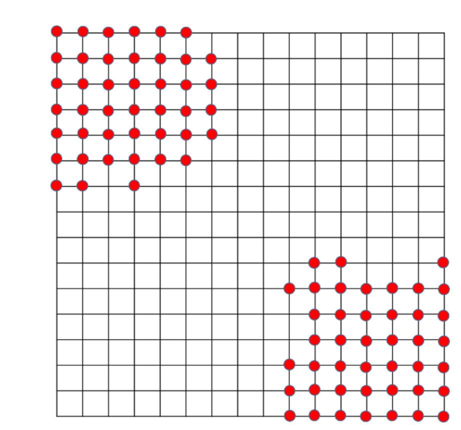
$$\check{p}\,(y_i|\mathbf{y}_{\mathcal{L}};\alpha)=\alpha\pi(y_i)+(1-\alpha)p\,(y_i|\mathbf{y}_{\mathcal{L}})\,,\quad 0\leq\alpha\leq1$$

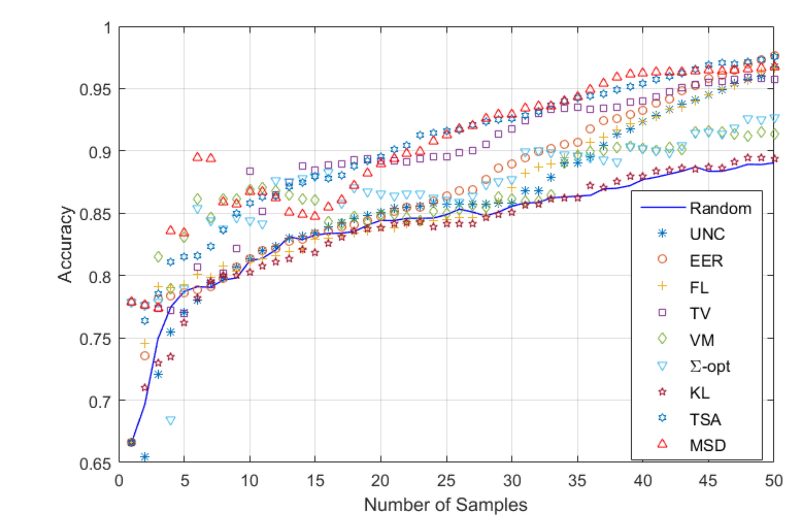
- **Example**: Total variation measure

$$U_{MSD}(v_i,\mathcal{L},a)\propto\left[0.5a+(1-a)(1-\mu_i^2)\right]\frac{\|\mathbf{g}_i\|_2^2}{g_{ii}^2}$$

- Use sequence $\{\alpha_t\}_{t=1}^T$ where $a_t\rightarrow0$ as model improves

Synthetic experiments- Rectangular grid



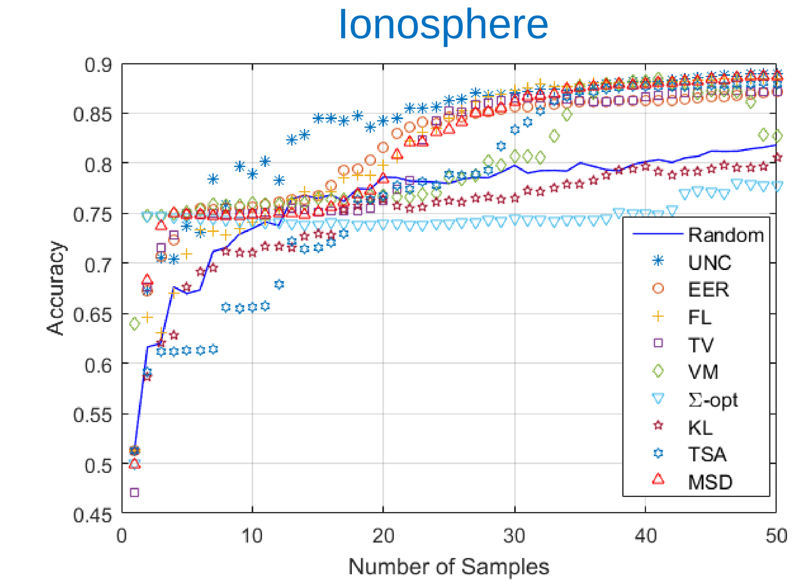


Real datasets

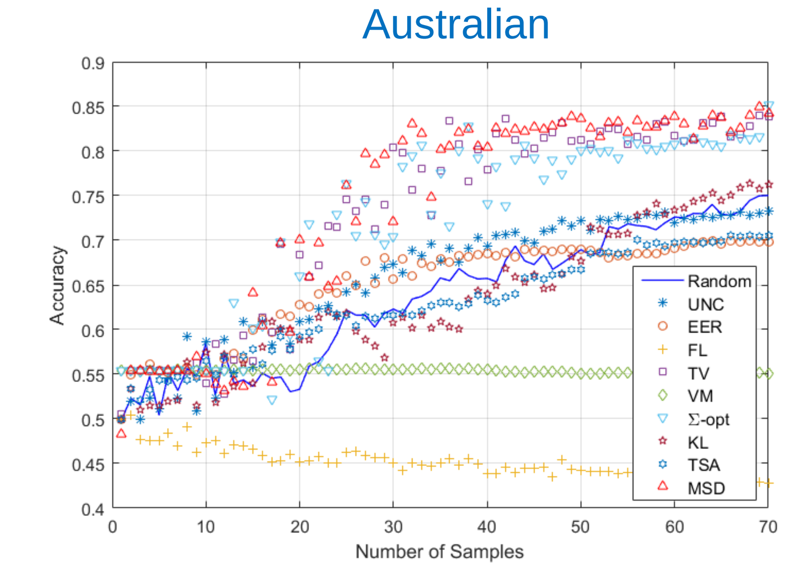
- Graph connectivity using Pearson correlation

Weighted adjacency matrix entries $w_{i,j}=\frac{\langle\mathbf{x}_i,\mathbf{x}_j\rangle}{\|\mathbf{x}_i\|_2\|\mathbf{x}_j\|_2}$

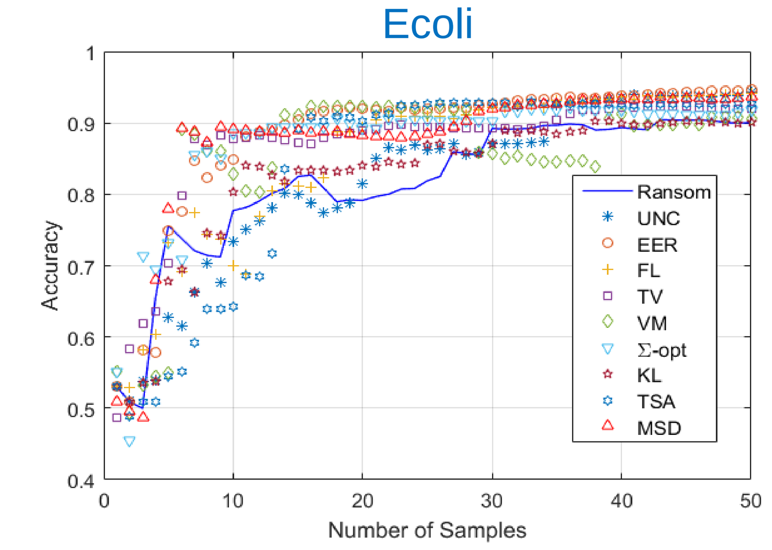
Ionosphere



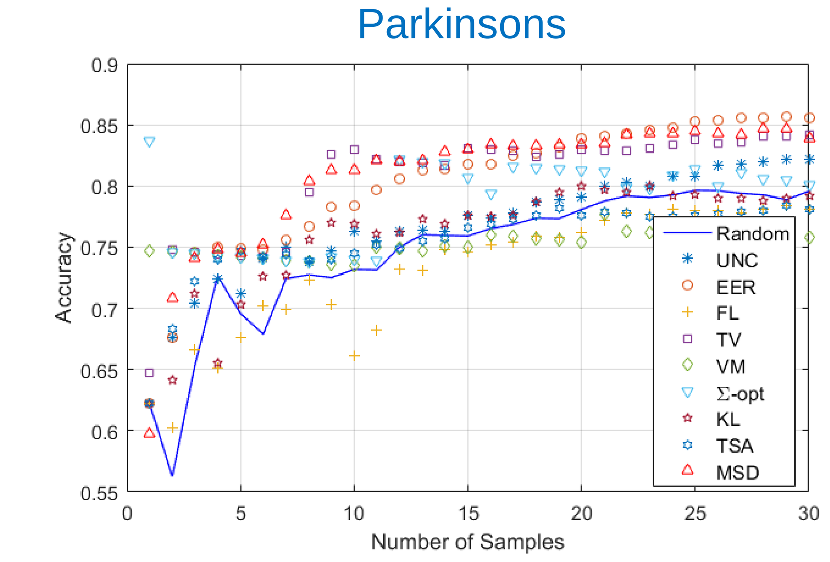
Australian



Ecoli



Parkinsons



Work in this paper was supported by NSF 1514056 and 1500713, and NIH 1R01GM104975-01.