Graph Transduction via Alternating Minimization

Jun Wang, Tony Jebara, and Shih-Fu Chang   *Columbia University*

Graph Transduction - A Brief Review

- Graph formulation and solution
  - ➢Samples (labeled and unlabeled) are treated as graph nodes and pair wise sample similarities are the weights of edges;
  - ➢A classification function is learned to propagate labels to unlabeled samples through minimizing a predefined loss function over graph (smoothness evaluation and fitness penalty);

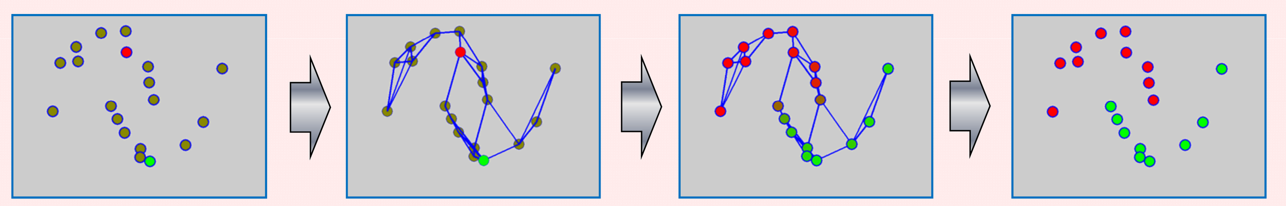
$$\mathbf{F}^* = \arg\min_{\mathbf{F}} \mathcal{Q}(\mathbf{F}) = \arg\min_{\mathbf{F}} \left\{ Q_{smooth}(\mathbf{F}) + Q_{fit}(\mathbf{F}) \right\}$$

  - ➢Related work:
    - Gaussian fields and Harmonic functions *GFHF* (Zhu, Ghahramani, and Lafferty ICML03)
    - Local and global consistency *LGC* (Zhou, Bousquet, Lal, Weston, and Scholkopf NIPS04)

$$\mathcal{Q}(F) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \left\| \frac{F_{i\cdot}}{\sqrt{D_{ii}}} - \frac{F_{j\cdot}}{\sqrt{D_{ii}}} \right\|^2 + \mu \sum_{i=1}^{n} \|F_{i\cdot} - Y_{i\cdot}\|^2$$

w_{ij} -- weight
F -- classification function
Y -- label matrix
$\mathcal{Q}$ -- loss function

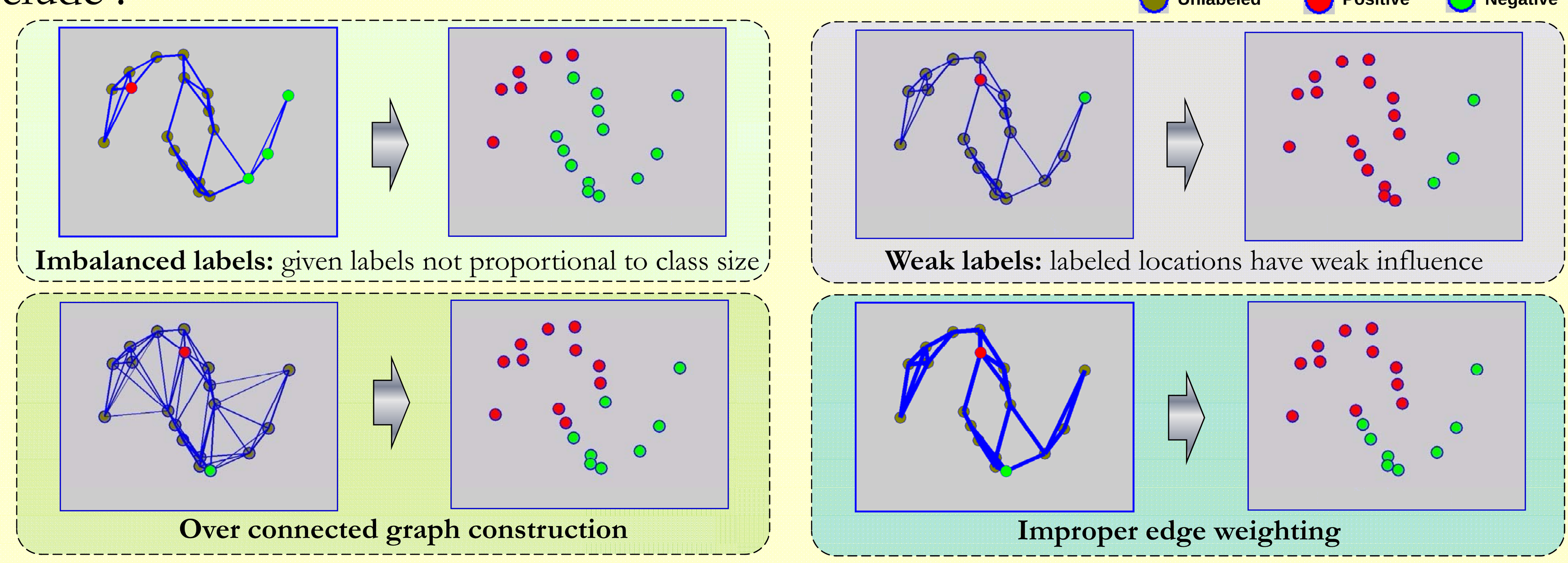
- Procedure of graph transduction learning



Given Samples | Graph Construction | Label Propagation | Classification Results

Problem and Motivation

- The derivation of the optimal classification function **F** highly relies on the label matrix **Y** and is very sensitive to graph construction **W**. Several problematic cases include :



Imbalanced labels: given labels not proportional to class size
Weak labels: labeled locations have weak influence
Over connected graph construction
Improper edge weighting

- Traditional approaches rely on initial labels which may be imbalanced, noisy, or ill positioned, and try to find optimal **F** alone. We propose a new formulation (Graph Transduction via Alternating Minimization ***GTAM***) to simultaneously refine the label matrix **Y** and find the optimal **F**.

New Loss Function

- Revisit the *LGC* formulation: prior work optimizes **F** alone

$$\mathcal{Q}(\mathbf{F}) = \frac{1}{2}\mathrm{tr}\left\{\mathbf{F}^T\mathbf{L}\mathbf{F} + \mu(\mathbf{F}-\mathbf{Y})^T(\mathbf{F}-\mathbf{Y})\right\}$$

L: normalized graph Laplacian

Smoothness   Fitness

- New formulation: jointly optimizes **F** and **Y**

$$\mathcal{Q}(\mathbf{F},\mathbf{Y}) = \frac{1}{2}\mathrm{tr}\left\{\mathbf{F}^T\mathbf{L}\mathbf{F} + \mu(\mathbf{F}-\mathbf{V}\mathbf{Y})^T(\mathbf{F}-\mathbf{V}\mathbf{Y})\right\}$$

Bivariate function   Label regularizer

$$\mathbf{v} = \sum_{j=1}^{c} \frac{\mathbf{Y}_{\cdot j} \odot \mathbf{D}\vec{1}}{\mathbf{Y}_{\cdot j}^T \mathbf{D}\vec{1}}$$

- Additionally, we propose a novel label regularizer with the following two merits.
  - ➢Normalizing labels among classes to handle imbalance issue (can incorporate class portion knowledge);
  - ➢Weighting labels based on their degree (reduce the impact of unreliable labels);

positive negative

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \frac{1}{1+3} & 0 & 0 & 0 \\ 0 & \frac{2}{2} & 0 & 0 \\ 0 & 0 & \frac{3}{1+3} & 0 \\ 0 & 0 & 0 & \frac{0}{4} \end{bmatrix}$$

Greedy Solution by Alternating Minimization

- Optimization on continues valued **F**

$$\frac{\partial \mathcal{Q}}{\partial \mathbf{F}^*} = 0 \Rightarrow \mathbf{F}^* = (\mathbf{L}/\mu + \mathbf{I})^{-1}\mathbf{V}\mathbf{Y} = \mathbf{P}\mathbf{V}\mathbf{Y}$$

$$\mathbf{P} = (\mathbf{L}/\mu + \mathbf{I})^{-1}$$

Propagation Matrix

$$\mathcal{Q}(\mathbf{F}^*, \mathbf{Y})$$

$$\mathcal{Q}(\mathbf{Y}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Y}^T\mathbf{V}^T\left[\mathbf{P}^T\mathbf{L}\mathbf{P} + \mu(\mathbf{P}^T - \mathbf{I})(\mathbf{P} - \mathbf{I})\right]\mathbf{V}\mathbf{Y}\right)$$

$$\mathbf{Z} = \mathbf{V}\mathbf{Y} \qquad \mathbf{A} = \mathbf{P}^T\mathbf{L}\mathbf{P} + (\mathbf{P}^T - \mathbf{I})(\mathbf{P} - \mathbf{I})$$

- Iterative gradient optimization on discrete valued **Y**

$$\mathcal{Q}(\mathbf{Z}) = \frac{1}{2}\mathrm{tr}\left(\mathbf{Z}^T\mathbf{A}\mathbf{Z}\right)$$

$$\frac{\partial \mathcal{Q}^{(t)}}{\partial \mathbf{Z}} = \mathbf{A}\mathbf{Z}^t = \mathbf{A}\mathbf{V}^t\mathbf{Y}^t$$

gradient matrix

$$\mathbf{Y}_{i^*j^*}^{t+1} = 1$$

set the best label that achieves the largest cost reduction

$$(i^*, j^*) = \arg\min_{\mathbf{x}_i \in \mathcal{X}_u, 1 \le j \le c} \nabla_{\mathbf{Z}_{ij}} \mathcal{Q}$$

$$\mathbf{v}^{t+1} = \sum_{j=1}^{c} \frac{\mathbf{Y}_{\cdot j}^{t+1} \odot \mathbf{D}\vec{1}}{\mathbf{Y}_{\cdot j}^{t+1\,T}\mathbf{D}\vec{1}}$$

Update label regularizer

$$\mathcal{X}_l^{t+1} \longleftarrow \mathcal{X}_l^t + \mathbf{x}_{i^*} \qquad \mathcal{X}_u^{t+1} \longleftarrow \mathcal{X}_u^t - \mathbf{x}_{i^*}$$

Update labeled and unlabeled sets

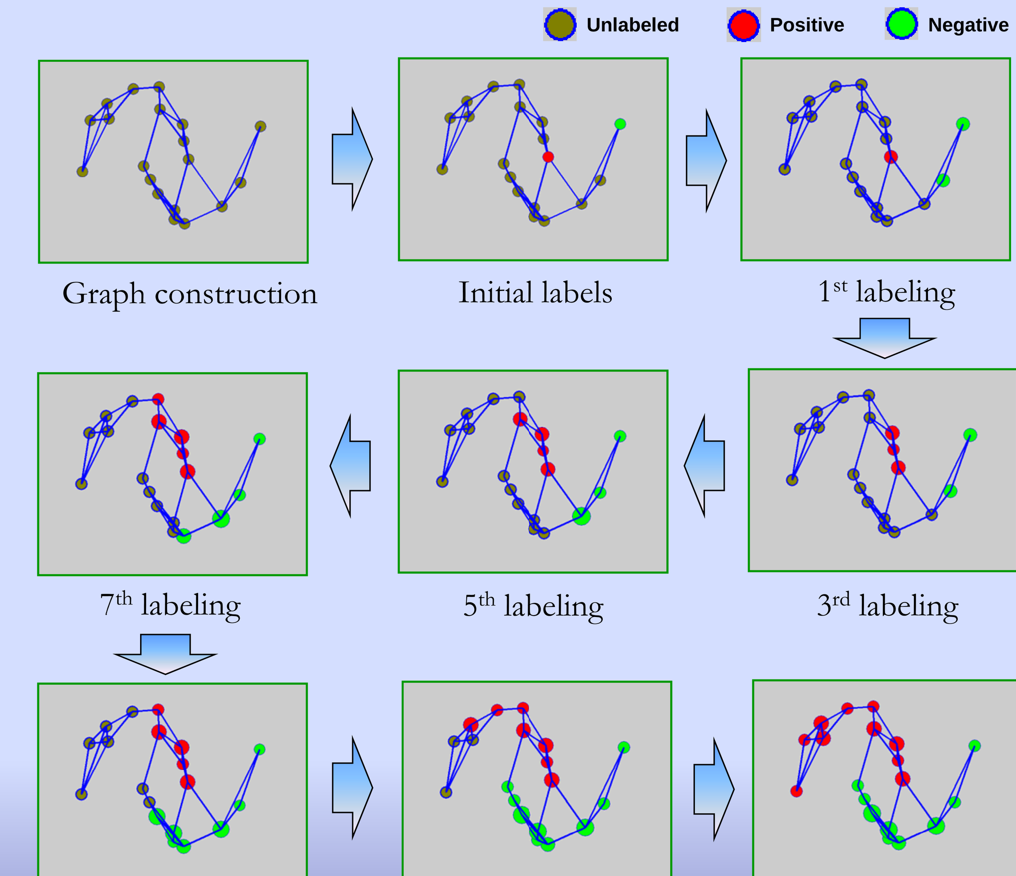
$t \leftarrow t+1$

update iteration counter

Output labels until $\mathcal{X}_u^{t+1} = \emptyset$

Intuitive Demonstration

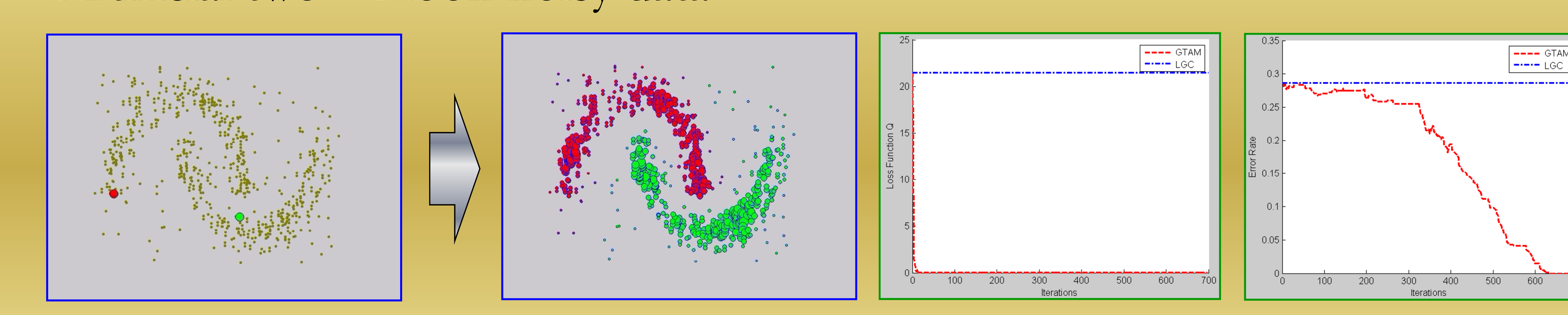
- Given a set of labeled samples, ***GTAM*** chooses the most beneficial data with largest cost reduction over the whole graph for label assignment, one sample in each iteration.



Graph construction | Initial labels | 1st labeling | 3rd labeling | 5th labeling | 7th labeling | 10th labeling | 14th labeling | 17th labeling

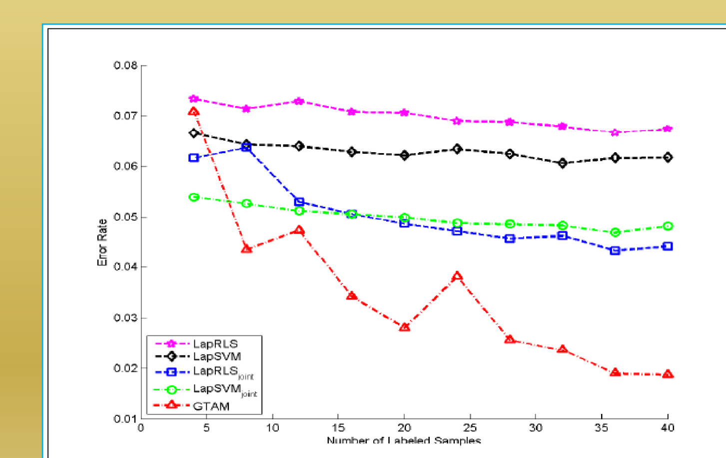
Experiments

- Artificial two – moon noisy data

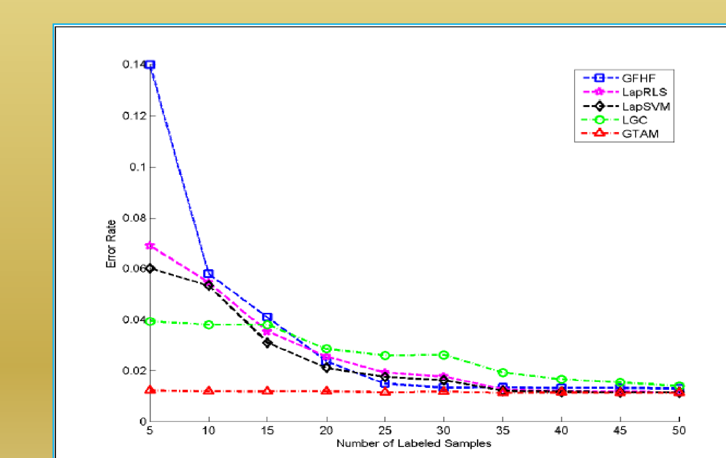


Input samples | Output labels | Loss function | Error rate

- Text Classification (WebKB data)
- Image Classification – USPS digits



Performance comparison on text classification (WebKB dataset). The horizontal axis represents the number of randomly observed labels (guaranteeing there is at least one label for each class). The vertical axis shows the average error rate over 100 random trials.



Performance comparison on image classification (USPS digits). The horizontal axis shows the total number of randomly observed labels (guarantee ing there is at least one label for each class). The vertical axis shows the average error rate over 20 random trials.