Graph-based Semi-Supervised Learning as Optimization

Partha Pratim Talukdar
CMU

Machine Learning with Large Datasets (10-605)
April 3, 2012
Graph-based Semi-Supervised Learning
Graph-based Semi-Supervised Learning

Labeled (seed)
Graph-based Semi-Supervised Learning

Unlabeled

Labeled (seed)
Graph-based Semi-Supervised Learning
Graph-based Semi-Supervised Learning

Various methods:
LP (Zhu et al., 2003); QC (Bengio et al., 2007); Adsorption (Baluja et al., 2008)
Notations

Seed Scores
Label Priors
Estimated Scores
Nota%ons

\( \hat{Y}_{v,l} \) : score of estimated label \( l \) on node \( v \)
Notations

\( \hat{Y}_{v,l} \) : score of estimated label \( l \) on node \( v \)

\( Y_{v,l} \) : score of seed label \( l \) on node \( v \)
Notations

\( \hat{Y}_{v,l} \) : score of estimated label \( l \) on node \( v \)

\( Y_{v,l} \) : score of seed label \( l \) on node \( v \)

\( R_{v,l} \) : regularization target for label \( l \) on node \( v \)
Notations

$\hat{Y}_{v,l}$: score of estimated label $l$ on node $v$

$Y_{v,l}$: score of seed label $l$ on node $v$

$R_{v,l}$: regularization target for label $l$ on node $v$

$S$: seed node indicator (diagonal matrix)
Notations

$\hat{Y}_{v,l}$ : score of estimated label $l$ on node $v$

$Y_{v,l}$ : score of seed label $l$ on node $v$

$R_{v,l}$ : regularization target for label $l$ on node $v$

$S$ : seed node indicator (diagonal matrix)

$W_{uv}$ : weight of edge $(u, v)$ in the graph
LP-ZGL (Zhu et al., ICML 2003)
LP-ZGL (Zhu et al., ICML 2003)

$$\arg \min_{\hat{Y}} \sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2$$

such that \( Y_{ul} = \hat{Y}_{ul} \), \( \forall S_{uu} = 1 \)
LP-ZGL (Zhu et al., ICML 2003)

\[
\arg\min_{\hat{Y}} \sum_{l=1}^{m} W_{uv} (\hat{Y}_{ul} - \hat{Y}_{vl})^2
\]

such that \( Y_{ul} = \hat{Y}_{ul}, \forall S_{uu} = 1 \)
LP-ZGL (Zhu et al., ICML 2003)

\[
\arg \min_{\hat{Y}} \sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2
\]

such that

\[
Y_{ul} = \hat{Y}_{ul}, \ \forall S_{uu} = 1
\]

Smooth

Match Seeds (hard)
LP-ZGL (Zhu et al., ICML 2003)

\[
\text{arg min} \sum_{l=1}^{m} W_{uv} (\hat{Y}_{ul} - \hat{Y}_{vl})^2 = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l
\]

such that

\[
Y_{ul} = \hat{Y}_{ul}, \forall S_{uu} = 1
\]

Smooth

Matching Seeds (hard)

Graph Laplacian

\[L = D - W\] (PSD)
LP-ZGL (Zhu et al., ICML 2003)

**Smoothness**

- two nodes connected by an edge with high weight should be assigned similar labels
LP-ZGL (Zhu et al., ICML 2003)

\[
\arg\min_{\hat{Y}} \sum_{l=1}^{m} W_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2 = \sum_{l=1}^{m} \hat{Y}_l^T L \hat{Y}_l
\]

such that

\[
Y_{ul} = \hat{Y}_{ul}, \quad \forall S_{uu} = 1
\]

Smooth

Match Seeds (hard)

• Smoothness
  – two nodes connected by an edge with high weight should be assigned similar labels

• Solution satisfies harmonic property
Two Related Views

Random Walk

L → U
Two Related Views

Random Walk

Label Diffusion
Random Walk View
Random Walk View

what next?
Random Walk View

what next?

- Continue walk with prob. \( p_{v}^{\text{cont}} \)
- Assign \( V \)'s seed label to \( U \) with prob. \( p_{v}^{\text{inj}} \)
- Abandon random walk with prob. \( p_{v}^{\text{abnd}} \)
  - assign \( U \) a dummy label
Discounting Nodes
Discounting Nodes

• Certain nodes can be unreliable (e.g., high degree nodes)
  • do not allow propagation/walk through them
Discounting Nodes

• Certain nodes can be unreliable (e.g., high degree nodes)
  • do not allow propagation/walk through them

• Solution: increase abandon probability on such nodes:
Discounting Nodes

• Certain nodes can be unreliable (e.g., high degree nodes)
  • do not allow propagation/walk through them

• Solution: increase abandon probability on such nodes:

\[ p_v^{\text{abnd}} \propto \text{degree}(v) \]
Redefining Matrices
Redefining Matrices

$$W'_{uv} = p^\text{cont}_u \times W_{uv}$$

New Edge Weight
Redefining Matrices

\[ W'_{uv} = p_{u}^{cont} \times W_{uv} \]

New Edge Weight

\[ S_{uu} = \sqrt{p_{u}^{inj}} \]
Redefining Matrices

\[ W_{uv}' = p_u^{\text{cont}} \times W_{uv} \]

\[ S_{uu} = \sqrt{p_u^{\text{inj}}} \]

\[ R_{u\top} = p_u^{\text{abnd}}, \text{ and 0 for non-dummy labels} \]
Modified Adsorption (MAD)
[Talukdar and Crammer, ECML 2009]
Modified Adsorption (MAD)
[Talukdar and Crammer, ECML 2009]

\[
\arg\min_{\hat{Y}} \sum_{l=1}^{m+1} \left[ \|S\hat{Y}_l - SY_l\|^2 + \mu_1 \sum_{u,v} M_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \|\hat{Y}_l - R_l\|^2 \right]
\]

- \(m\) labels, +1 dummy label
- \(M = W^\top + W'\) is the symmetrized weight matrix
- \(\hat{Y}_{vl}\): weight of label \(l\) on node \(v\)
- \(Y_{vl}\): seed weight for label \(l\) on node \(v\)
- \(S\): diagonal matrix, nonzero for seed nodes
- \(R_{vl}\): regularization target for label \(l\) on node \(v\)
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

\[
\arg\min_{\hat{Y}} \sum_{l=1}^{m+1} \left[ \| S\hat{Y}_l - SY_l \|^2 \right] + \mu_1 \sum_{u,v} M_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \| \hat{Y}_l - R_l \|^2
\]

- \(m\) labels, +1 dummy label
- \(M = W^\dagger + W'\) is the symmetrized weight matrix
- \(\hat{Y}_{vl}\): weight of label \(l\) on node \(v\)
- \(Y_{vl}\): seed weight for label \(l\) on node \(v\)
- \(S\): diagonal matrix, nonzero for seed nodes
- \(R_{vl}\): regularization target for label \(l\) on node \(v\)
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

\[
\text{arg min}_{\hat{Y}} \sum_{l=1}^{m+1} \left[ \| S\hat{Y}_l - SY_l \|^2 \right] + \mu_1 \sum_{u,v} M_{uv}(\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \| \hat{Y}_l - R_l \|^2
\]

- \( m \) labels, +1 dummy label
- \( M = W^\dagger + W' \) is the symmetrized weight matrix
- \( \hat{Y}_{vl} \): weight of label \( l \) on node \( v \)
- \( Y_{vl} \): seed weight for label \( l \) on node \( v \)
- \( S \): diagonal matrix, nonzero for seed nodes
- \( R_{vl} \): regularization target for label \( l \) on node \( v \)
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{Y}} \sum_{l=1}^{m+1} \left( \|S\hat{Y}_l - SY_l\|^2 \right) + \mu_1 \sum_{u,v} M_{uv} (\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \|\hat{Y}_l - R_l\|^2$$

- $m$ labels, +1 dummy label
- $M = W^T + W'$ is the symmetrized weight matrix
- $\hat{Y}_{vl}$: weight of label $l$ on node $v$
- $Y_{vl}$: seed weight for label $l$ on node $v$
- $S$: diagonal matrix, nonzero for seed nodes
- $R_{vl}$: regularization target for label $l$ on node $v$
Modified Adsorption (MAD)
[Talukdar and Crammer, ECML 2009]

\[
\arg\min_{\hat{Y}} \sum_{l=1}^{m+1} \left[ \| S\hat{Y}_l - SY_l \|_2^2 + \mu_1 \sum_{u,v} M_{uv} (\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \| \hat{Y}_l - R_l \|_2^2 \right]
\]

- \( m \) labels, +1 dummy label
- \( M = W \) is the symmetrized weight matrix
- \( \hat{Y}_{vl} \): weight of label \( l \) on node \( v \)
- \( Y_{vl} \): seed weight for label \( l \) on node \( v \)
- \( S \): diagonal matrix, nonzero for seed nodes
- \( R_{vl} \): regularization target for label \( l \) on node \( v \)

Match Seeds (soft)  
Smooth  
Match Priors (Regularizer)

\( \hat{Y}_v \) used to indicate none-of-the-above label
Modified Adsorption (MAD)

[Talukdar and Crammer, ECML 2009]

$$\arg\min_{\hat{Y}} \sum_{l=1}^{m+1} \left( \left\| S \hat{Y}_l - S Y_l \right\|^2 \right) + \mu_1 \sum_{u,v} M_{uv} (\hat{Y}_{ul} - \hat{Y}_{vl})^2 + \mu_2 \left\| \hat{Y}_l - R_l \right\|^2$$

- $m$ labels, +1 dummy label
- $M = W$ is the symmetrized weight matrix
- $\hat{Y}_{vl}$: weight of label $l$ on node $v$
- $Y_{vl}$: seed weight for label $l$ on node $v$
- $S$: diagonal matrix, nonzero for seed nodes
- $R_{vl}$: regularization target for label $l$ on node $v$

Match Seeds (soft)

Smooth

Match Priors (Regularizer)

MAD has extra regularization compared to LP-ZGL (Zhu et al, ICML 03); similar to QC (Bengio et al, 2006)
Inputs $Y, R : |V| \times (|L| + 1), W : |V| \times |V|, S : |V| \times |V|$ diagonal

$\hat{Y} \leftarrow Y$

$M = W' + W^\dagger$

$Z_v \leftarrow S_{vv} + \mu_1 \sum_{u \neq v} M_{vu} + \mu_2 \quad \forall v \in V$

repeat
  for all $v \in V$ do
    $\hat{Y}_v \leftarrow \frac{1}{Z_v} \left( (SY)_v + \mu_1 M_{v}.\hat{Y} + \mu_2 R_v \right)$
  end for
until convergence
MAD (contd.)

Inputs $Y, R: |V| \times (|L| + 1), W: |V| \times |V|, S: |V| \times |V|$ diagonal
$\hat{Y} \leftarrow Y$
$M = W^t + W^\dagger$
$Z_v \leftarrow S_{vv} + \mu_1 \sum_{u \neq v} M_{vu} + \mu_2 \quad \forall v \in V$
repeat
  for all $v \in V$ do
    $\hat{Y}_v \leftarrow \frac{1}{Z_v} \left( (SY)_v + \mu_1 M_{v.} \hat{Y} + \mu_2 R_v \right)$
  end for
until convergence

- Extends Adsorption with well-defined optimization
- Importance of a node can be discounted
- Easily Parallelizable: Scalable
MapReduce Implementation of MAD
MapReduce Implementation of MAD

• Map
  – Each node send its current label assignments to its neighbors
MapReduce Implementation of MAD

• Map
  – Each node send its current label assignments to its neighbors

• Reduce
  – Each node updates its own label assignment using messages received from neighbors, and its own information (e.g., seed labels, reg. penalties etc.)
MapReduce Implementation of MAD

- **Map**
  - Each node sends its current label assignments to its neighbors

- **Reduce**
  - Each node updates its own label assignment using messages received from neighbors, and its own information (e.g., seed labels, reg. penalties etc.)

- **Repeat until convergence**
Experiments
Text Classification
Text Classification

PRBEP (macro-averaged) on WebKB Dataset, 3148 test instances
Sentiment Classification
Sentiment Classification

Precision on 3568 Sentiment test instances
Class-Instance Acquisition
Class-Instance Acquisition

Graph with 303k nodes, 2.3m edges.
Class-Instance Acquisition

Freebase-2 Graph, 192 WordNet Classes

Mean Reciprocal Rank (MRR)

Amount of Supervision

LP-ZGL Adsorption MAD

Graph with 303k nodes, 2.3m edges.
## New (Class, Instance) Pairs Found

<table>
<thead>
<tr>
<th>Class</th>
<th>A few non-seed Instances found by Adsorption</th>
</tr>
</thead>
<tbody>
<tr>
<td>NFL Players</td>
<td>Tony Gonzales, Thabiti Davis, Taylor Stubblefield, Ron Dixon, Rodney Hannan, …</td>
</tr>
</tbody>
</table>

Total classes: **9081**
When is MAD most effective?
When is MAD most effective?

![Graph showing relative increase in MRR by MAD over LP-ZGL vs. average degree.](image-url)

- Relative Increase in MRR by MAD over LP-ZGL
- Average Degree

The graph indicates that MAD is most effective when the average degree is high, specifically around 30, where the relative increase in MRR is maximized.
When is MAD most effective?

MAD seems to be more effective in graphs with high average degree, where there is greater need for regularization.
Benefits of Having an Optimization Framework
Extension to Dependent Labels

- Labels are not always mutually exclusive.
Extension to Dependent Labels

- Labels are not always mutually exclusive.
• Labels are not always mutually exclusive.
MAD with Dependent Labels (MADDL)

\[
\text{min} \quad \text{Seed Label Loss} \quad + \quad \text{Edge Smoothness Loss} \quad + \quad \text{Label Prior Loss (e.g. prior on dummy label)}
\]
MAD with Dependent Labels (MADDL)

\[
\min \quad \text{Seed Label Loss (if any)} + \text{Edge Smoothness Loss} + \text{Label Prior Loss (e.g. prior on dummy label)} + \text{Dependent Label Loss}
\]
MAD with Dependent Labels (MADDL)

MADDL Objective

\[ \min \text{ Seed Label Loss (if any)} + \text{ Edge Smoothness Loss} + \text{ Label Prior Loss (e.g. prior on dummy label)} + \text{ Dependent Label Loss} \]
MAD with Dependent Labels (MADDL)

MADDL Objective

\[
\text{min} \quad \text{Seed Label Loss (if any)} + \text{Edge Smoothness Loss} + \text{Label Prior Loss (e.g. prior on dummy label)} + \text{Dependent Label Loss}
\]

Penalize if similar labels are assigned different scores on a node
MAD with Dependent Labels (MADDL)

MADDL Objective

\[
\min \quad \text{Seed Label Loss (if any)} + \text{Edge Smoothness Loss} + \text{Label Prior Loss (e.g. prior on dummy label)} + \text{Dependent Label Loss}
\]

Penalize if similar labels are assigned different scores on a node
MAD with Dependent Labels (MADDL)

MADDL Objective

\[
\min \text{ Seed Label Loss (if any)} + \text{ Edge Smoothness Loss} + \text{ Label Prior Loss (e.g. prior on dummy label)} + \text{ Dependent Label Loss}
\]

MADDL objective results in a scalable iterative update, with convergence guarantee.

Penalize if similar labels are assigned different scores on a node.

BrownAle 1.0 Ale
Smooth Sentiment Ranking

rank 1

smooth predictions
Smooth Sentiment Ranking

rank 1
smooth predictions

rank 4

non-smooth predictions
Smooth Sentiment Ranking

Prefer smooth predictions over non-smooth predictions.

Rank 1: ⭐⭐⭐⭐⭐

Rank 4: ⭐⭐⭐⭐
Smooth Sentiment Ranking

Prefer smooth predictions over non-smooth predictions

MADDL Label Constraints
Smooth Sentiment Ranking

Count of Top Predicted Pair in MAD Output
Smooth Sentiment Ranking

Count of Top Predicted Pair in MAD Output

Count of Top Predicted Pair in MADDL Output
Smooth Sentiment Ranking

MADDL generates smoother ranking, while preserving accuracy of prediction.
Other Methods and Graph Construction
Other Methods and Graph Construction

• Measure Propagation [Subramanya and Bilmes, NIPS 2009]
Other Methods and Graph Construction

• Measure Propagation [Subramanya and Bilmes, NIPS 2009]
  - KL-based loss
Other Methods and Graph Construction

- **Measure Propagation** [Subramanya and Bilmes, NIPS 2009]
  - KL-based loss
  - Results over 120m node graph
Other Methods and Graph Construction

• **Measure Propagation** [Subramanya and Bilmes, NIPS 2009]
  - KL-based loss
  - Results over 120m node graph

• Chapter 21 of the SSL book [Chapelle et al.]
Other Methods and Graph Construction

• **Measure Propagation** [Subramanya and Bilmes, NIPS 2009]
  - KL-based loss
  - Results over 120m node graph

• **Chapter 21 of the SSL book** [Chappelle et al.]

• **Graph Construction**
Other Methods and Graph Construction

- **Measure Propagation** [Subramanya and Bilmes, NIPS 2009]
  - KL-based loss
  - Results over 120m node graph

- **Chapter 21 of the SSL book** [Chapelle et al.]

- **Graph Construction**
  - **Task-specific Graph Construction using Metric Learning** [Dhillon, Talukdar, and Crammer, ACL 2010]
Graph-based SSL: Final Thoughts
Graph-based SSL: Final Thoughts

- Provide flexible representation
  - for both relational and IID data
Graph-based SSL: Final Thoughts

• Provide flexible representation
  – for both relational and IID data

• Scalable: MR/Hadoop-based implementation
Graph-based SSL: Final Thoughts

• Provide flexible representation
  – for both relational and IID data

• Scalable: MR/Hadoop-based implementation

• Can handle labeled as well as unlabeled data
Graph-based SSL: Final Thoughts

• Provide flexible representation
  – for both relational and IID data

• Scalable: MR/Hadoop-based implementation

• Can handle labeled as well as unlabeled data

• Can handle multiple-classes
Graph-based SSL: Final Thoughts

• Provide flexible representation
  – for both relational and IID data
• Scalable: MR/Hadoop-based implementation
• Can handle labeled as well as unlabeled data
• Can handle multiple-classes
• Effective in practice
Code in Junto Label Propagation Toolkit
(includes Hadoop-based implementation)

http://code.google.com/p/junto/
Code in Junto Label Propagation Toolkit (includes Hadoop-based implementation)

http://code.google.com/p/junto/

Thanks!