Graph-based Transfer Learning

Jingrui He* Yan Liu† Rick Lawrence†
* Carnegie Mellon University, † IBM Research
Problem Definition

• An example of Transfer Learning

<table>
<thead>
<tr>
<th>Wow. Just wow. And I thought Batman Begins was excellent. This piece of art is PHENOMENAL!! From the scale, to the acting, the atmosphere, the music, the action, it's all art. I have not experienced this level of greatness in the cinema for a long time.</th>
<th>There is just so much to bash this movie about. It’s an obvious cash in, but even fails at being entertaining. It is boring, confusing and the characters are bland. Its just an all round failure and should be buried in the Tomb of the title, never to be re-awakened.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Apple iPhone has a stunning display, a sleek design, and an innovative multitouch user interface. Its Safari browser makes for a superb Web surfing experience, and it offers easy-to-use apps. As an iPod, it shines.</td>
<td>The Apple iPhone has variable call quality and lacks some basic features found in many cell phones, including stereo Bluetooth support and 3G compatibility. Integrated memory is stingy for an iPod, and you have to sync the iPhone to manage music content.</td>
</tr>
</tbody>
</table>
General View

• Source domain
  – Feature space: $X^S$
  – Feature distribution: $D^S$
  – Labeling function: $h^S$

• Target domain
  – Feature space: $X^T$
  – Feature distribution: $D^T$
  – Labeling function: $h^T$

• Scenario 1
  ➢ $X^S = X^T$, $D^S = D^T$, $h^S \neq h^T$
  ➢ E.g. multi-label text classification

• Scenario 2
  ➢ $X^S = X^T$, $D^S \neq D^T$, $h^S \neq h^T$
  ➢ E.g. sentiment classification in different domains; hierarchical document classification

• Scenario 3
  ➢ $X^S \neq X^T$, $D^S \neq D^T$, $h^S \neq h^T$
  ➢ E.g. collaborative filtering; verb argument classification
Roadmap

• Problem Definition
• Related Work
• Learning with Tripartite Graph Only
• Graph-based Learning Framework
• Experimental Results
• Conclusion
Related Work

- **Scenario 1**
  - $X^S = X^T$, $D^S = D^T$, $h^S \neq h^T$

- **Scenario 2**
  - $X^S = X^T$, $D^S \neq D^T$, $h^S \neq h^T$
  - Locally weighted ensemble framework [Gao et al 2008]; spectral regularization framework [Argyriou et al 2007]; etc

- **Scenario 3**
  - $X^S \neq X^T$, $D^S \neq D^T$, $h^S \neq h^T$
  - Meta-level prior [Lee et al 2007]; semi-supervised multitask learning [Liu et al 2007]; etc
Roadmap

• Problem Definition
• Related Work
• Learning with Tripartite Graph Only
• Graph-based Learning Framework
• Experimental Results
• Conclusion
Motivation

• Tripartite graph
  – Transfer feature-level label information

• Compared with existing methods
  – Non-parametric classification function: more flexible
  – Regularization terms in the objective function: prevent over-fitting
  – Feature-level transfer: more reliable
Tripartite Graph

\[ f^L = \begin{bmatrix} \end{bmatrix} \]

\[ f^F = \begin{bmatrix} \end{bmatrix} \]

\[ f^U = \begin{bmatrix} \end{bmatrix} \]

\[ A^{(3,2)} \]

\[ A^{(3,1)} \]

\[ L \]

\[ U \]
Objective Function $Q_1$

$$Q_1(f) = \frac{1}{2} \sum_{i,j=1}^{m+n+d} A_{ij}^{(3)} \left( \frac{f_i}{\sqrt{D_i^{(3)}}} - \frac{f_j}{\sqrt{D_j^{(3)}}} \right)^2 + u \sum_{i=1}^{m+n+d} (f_i - y_i)^2$$

$$f = \begin{bmatrix} f^L \\ f^U \\ f^F \end{bmatrix} \rightarrow m \times 1 \leftarrow \begin{bmatrix} y^L \\ y^U \\ y^F \end{bmatrix} = y \quad A^{(3)} = \begin{bmatrix} 0 & 0 & A^{(3,1)} \\ 0 & 0 & A^{(3,2)} \\ (A^{(3,1)})^T & (A^{(3,2)})^T & 0 \end{bmatrix}$$

$$D^{(3)} = \begin{bmatrix} D^{(3,1)} & 0 & 0 \\ 0 & D^{(3,2)} & 0 \\ 0 & 0 & D^{(3,3)} \end{bmatrix} \quad S^{(3)} = \begin{bmatrix} 0 & 0 & S^{(3,1)} \\ 0 & 0 & S^{(3,2)} \\ (S^{(3,1)})^T & (S^{(3,2)})^T & 0 \end{bmatrix}$$

$$D^{(3)} = \begin{bmatrix} D^{(3,1)} & 0 & 0 \\ 0 & D^{(3,2)} & 0 \\ 0 & 0 & D^{(3,3)} \end{bmatrix} \quad S^{(3)} = \begin{bmatrix} 0 & 0 & S^{(3,1)} \\ 0 & 0 & S^{(3,2)} \\ (S^{(3,1)})^T & (S^{(3,2)})^T & 0 \end{bmatrix}$$
Intuition of $Q_1$

Classification function

$f_i^U > 0 : +$

$f_i^U < 0 : -$  

$Q_1(f) = \frac{1}{2} \sum_{i,j=1}^{m+n+d} A_{ij}^{(3)} \left( \frac{f_i}{\sqrt{D_i^{(3)}}} - \frac{f_j}{\sqrt{D_j^{(3)}}} \right) ^2 + \mu \sum_{i=1}^{m+n+d} (f_i - y_i)^2$

$y^L$: label information

$y^U, y^F$: prior knowledge

Label smoothness

Consistency with label information and prior knowledge
Solving $Q_1$

- **Lemma 1**: If $f^L = y^L$, $Q_1$ is minimized at

$$f^{U*} = \left( I - \alpha^2 S^{(3,2)} (S^{(3,2)})^T \right)^{-1} \left[ (1-\alpha)y^{U} + \alpha(1-\alpha)S^{(3,2)}y^{F} + \alpha^2 S^{(3,2)} (S^{(3,1)})^T y^{L} \right]$$

$$f^{F*} = \left( I - \alpha^2 (S^{(3,2)})^T S^{(3,2)} \right)^{-1} \left[ (1-\alpha)y^{F} + \alpha(S^{(3,1)})^T y^{L} + \alpha(1-\alpha)(S^{(3,2)})^T y^{U} \right]$$

where $\alpha = \frac{1}{1+\mu}$

- Matrix inversion: computationally *expensive* for large data sets
TRITEX: Iterative Algorithm for $Q_1$

- Repeat the following steps until convergence

\[ f^U(t+1) = \alpha S^{(3,2)} f^F(t) + (1-\alpha)y^U \]

\[ f^F(t+1) = \alpha (S^{(3,1)})^T y^L + \alpha (S^{(3,2)})^T f^U(t) + (1-\alpha)y^F \]

- **Theorem 1:** upon convergence

\[ f^U(\infty) = f^{U*} \quad f^F(\infty) = f^{F*} \]
Problem with $Q_1$

Labeled data from both the source domain and the target domain
Roadmap

• Problem Definition
• Related Work
• Learning with Tripartite Graph Only
• Graph-based Learning Framework
• Experimental Results
• Conclusion
Improved Objective Function $Q_2$

$$Q_2(f) = \frac{1}{2} \sum_{i=1}^{m+n+d} \sum_{j=1}^{m+n} A_{ij}^{(3)} \left( f_i \sqrt{D_i^{(3)}} - f_j \sqrt{D_j^{(3)}} \right)^2$$

$$-S = Bf = \begin{bmatrix} 1_{(m+n) \times (m+n)} & 0_{(m+n) \times m} \end{bmatrix}$$

$$A^{(2)} = \begin{bmatrix} A^{(2,1)} & 0 \\ (A^{(2,1)})^T \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$
Solving $Q_2$

- **Lemma 3**: if $f^L = y^L$, $Q_2$ is minimized at

\[
\tilde{f}^{U^*} = \left( (\gamma + \tau + \mu)I - \frac{\gamma^2}{\gamma + \mu} S^{(3,2)}(S^{(3,2)})^T \right)^{-1}.
\]

\[
\left( \mu y^U + \frac{\gamma \mu}{\gamma + \mu} S^{(3,2)} y^F + \frac{\gamma^2}{\gamma + \mu} S^{(3,2)}(S^{(3,1)})^T y^L + \tau (S^{(2,1)})^T y^L \right)
\]

\[
f^{U^*} = \left( I - \alpha^2 S^{(3,2)}(S^{(3,2)})^T \right)^{-1} \left( (1 - \alpha) y^U + \alpha (1 - \alpha) S^{(3,2)} y^F + \alpha^2 S^{(3,2)}(S^{(3,1)})^T y^L \right)
\]
Relationship between $Q_1$ and $Q_2$

- **Theorem 2**: if $f^L = y^L$, then $\tilde{f}^U^*$ can be obtained by minimizing $Q_1$ with the following parametrization

Prior knowledge of unlabeled data

- TRITER Algorithm for Minimizing $Q_2$
Roadmap

- Problem Definition
- Related Work
- Learning with Tripartite Graph Only
- Graph-based Learning Framework
- Experimental Results
- Conclusion
Settings

- **3 Areas**
  - Sentiment Classification (SC): movie and product review data set
  - Document Classification (DC): 20 newsgroups data set
  - Intrusion Detection (ID): KDD Cup 99 data set

- **Competitors**
  - Target Only: learning from the target domain only
  - Source Only: learning from the source domain only
  - Source+Target: linear combination of source and target domains
  - Semi-supervised: manifold ranking algorithm
  - BTL: transfer learning toolkit by UC Berkeley
  - TBoost: boosting-based transfer learning method
Parameter Study

\[
\gamma = 1 \quad \tau = 5
\]
Comparison Results

Sentiment Classification

Intrusion Detection

Document Classification
Conclusion

• Graph-based Transfer Learning Framework
  – Tripartite Graph
    • 3 types of nodes: labeled, unlabeled, features
    • Propagate the label information via the features
  – Bipartite Graph
    • 2 types of nodes: labeled, unlabeled
    • Domain related smoothness constraint
  – Objective Function $Q_2$
    • Label smoothness on the tripartite graph
    • Label smoothness on the bipartite graph
    • Consistency with the label information and the prior knowledge
  – Iterative Algorithm: TRITER
Thank You!