Combine Link and Content for Community Detection: A Discriminative Approach

KDD 09

Tianbao Yang, Rong Jin, Yun Chi, Shenghuo Zhu

Michigan State University & NEC Laboratories America

1. R97944018 李庭嫣
2. R97922028 賴弘哲
Outline

• Problems
• Related Works
• Model
• Experiments
• Conclusion
Problem to Solve

- Community detection from web content
  - By Clustering (Unsupervised learning)
Traditional Solutions

Link Analysis

Hybrid

Content Analysis
Link Analysis

• Measured Based
  – Edge-Removal: **Modularity, Bridge-Cut**
  – Partitioning: **Kernighan-Lin**
  – Hierarchical: **Hierarchical Clustering**
  – Density-based: **SCAN**
  – Ranking-based: **C-Rank, Page-Rank, HITS**

• Probabilistic model
  – Model-based: **EM, PHITS**
Content Analysis

• Topic Model
  – PLSA
  – LDA

• Generative Model
  – co-occurrence → irrelevant word
Hybrid

- PHITS-PLSA
- LDA-Link-Word

Use generative model for content analysis part
→ Discriminative model
Generative VS Discriminative

Generative model
From Wikipedia, the free encyclopedia

In statistics, a generative model is a model for randomly generating observable data, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Generative models are used in machine learning for either modeling data directly (i.e., modeling observed draws from a probability density function), or as an intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through the use of Bayes' rule.

Discriminative model
From Wikipedia, the free encyclopedia

Discriminative models are a class of models used in machine learning for modeling the dependence of an unobserved variable $\mathbf{y}$ on an observed variable $\mathbf{x}$. Within a statistical framework, this is done by modeling the conditional probability distribution $P(\mathbf{y} | \mathbf{x})$, which can be used for predicting $\mathbf{y}$ from $\mathbf{x}$.

Ex. For a general word, eg. “fun”, 2 groups: “Sports” and “Entertainment”
Generative Model: “fun” will be in the model for both “Sports” and “Entertainment”
Discriminative Model
Model
Flow

Link Analysis
- Popularity-Based Conditional Link Model (PCL)

Content Analysis
- Discriminative Content Model (DC)

Optimization: EM
Notations(1/2)

- **Node space** \( \mathcal{V} = \{1, \cdots, n\} \)
- **Link space** \( \mathcal{E} = \{(i \rightarrow j) | s_{ij} \neq 0\} \)
  - Link \( s_{ij} \) can be \( \{0, 1\} \) or \( \mathbb{N}^+ \)
  - \( s_{ij} \neq 0 \) means a link from \( i \) to \( j \) or node \( i \) cites \( j \)
  - Link-in space \( \mathcal{LI}(i) \in \mathcal{V} \): all nodes possibly cite \( i \)
  - Link-out space \( \mathcal{LO}(i) \in \mathcal{V} \): all nodes possible cited by \( i \)
  - \( \mathcal{LI}(i) = \mathcal{LO}(i) = \mathcal{V} \) in general case, but in citation network: \( \mathcal{LI}(i) \) is all papers newer than \( i \)
Notations (2/2)

- $\mathcal{I}(i) = \{j | s_{ji} \neq 0\}$ nodes actually cite i
- $\mathcal{O}(i) = \{j | s_{ij} \neq 0\}$ nodes actually cited by i

- **Degree**
  - $d_{in}(i) = |\mathcal{I}(i)|$ indegree of node i
  - $d_{out}(i) = |\mathcal{O}(i)|$ outdegree of node i

- **K**: the number of community
Popularity-based Conditioned Link Model (PCL) (1/3)

• Our target: $\Pr(j|i)$, the probability of linking node $i$ to $j$ ($j$ is among $\mathcal{L}(i)$)

• For each node:
  
  – $z_i \in \{1, \cdots, K\}$ denote the community of $i$
  
  – $b_i \geq 0$, the popularity variable of node $i$, the higher popularity means $i$ will more likely be cited by other nodes

  – $\gamma_{ik}$ community membership of node $i$ in community $k$

  $$\Pr(j|i; z_i, b) = \frac{\gamma_{jz_i} b_j}{\sum_{j' \in \mathcal{L}(i)} \gamma_{j'z_i} b_{j'}}$$
Popularity-based Conditioned Link Model (PCL) (2/3)

\[
\Pr(j|i; \gamma, b) = \frac{\gamma_{jz_i} b_j}{\sum_{j' \in \mathcal{O}(i)} \gamma_{j'z_i} b_{j'}}
\]

- Assume \( z_i \sim \text{Mult} (\gamma_{i1}, \cdots, \gamma_{iK}) \), multinomial distribution

\[
\Pr(j|i; \gamma, b) = \sum_k \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j' \in \mathcal{O}(i)} \gamma_{j'k} b_{j'}}
\] (2)
Maximum likelihood estimate for PCL model (1/3)

- Observe on $\mathcal{E} = \{(i \rightarrow j)|s_{ij} \neq 0\}$
  - Likelihood
  $$\log \mathcal{L} = \sum_{(i \rightarrow j) \in \mathcal{E}} \hat{s}_{ij} \log \sum_k \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j' \in \mathcal{O}(i)} \gamma_{j'k} b_{j'}}$$
  - $\hat{s}_{ij}$ is normalized $s_{ij}$, $\sum_{j \in \mathcal{O}(i)} \hat{s}_{ij} = 1$

- MLE for optimal $\gamma$ and $b$
  $$\max_{\gamma, b} \sum_{(i \rightarrow j) \in \mathcal{E}} \hat{s}_{ij} \log \sum_k \gamma_{ik} \frac{\gamma_{jk} b_j}{\sum_{j' \in \mathcal{O}(i)} \gamma_{j'k} b_{j'}}$$
  s.t. $\sum_k \gamma_{ik} = 1$, $\gamma_{ik} \geq 0$, $b_i \geq 0$
Maximum likelihood estimate for PCL model (2/3)

- By apply Jenson’s inequality and \(-\log x \geq 1 - x\)

\[
\max_{\gamma, b} \sum_{(i \rightarrow j) \in E} \hat{s}_{ij} \sum_k q_{ijk} \left( \log \gamma_{ik} \gamma_{jk} b_j - \sum_{j' \in \mathcal{O}(i)} \frac{\gamma_{j'k} b_{j'}}{\tau_{ik}} \right)
\]

\[
\text{s.t. } \sum_k \gamma_{ik} = 1, \gamma_{ik} \geq 0, b_i \geq 0
\]

\[
\tau_{ik} = \sum_{j' \in \mathcal{O}(i)} \gamma_{j'k}^{t-1} b_{j'}^{t-1}
\]

\[
q_{ijk} \propto \gamma_{ik}^{t-1} \frac{\gamma_{jk}^{t-1} b_j^{t-1}}{\sum_{j' \in \mathcal{O}(i)} \gamma_{j'k}^{t-1} b_{j'}^{t-1}} \quad \text{s.t. } \sum_k q_{ijk} = 1
\]
Maximum likelihood estimate for PCL model (3/3)

- Optimal solution

\[ \forall i, d_{out}(i) \neq 0, d_{in}(i) \neq 0, \quad \gamma_{ik} = \frac{n(i, k)}{m(i, k)b_i + n_{out}(i)} , \quad b_i = \frac{n_{in}(i)}{\sum_k m(i, k)\gamma_{ik}} \]  \hspace{1cm} (9)

\[ \forall i, d_{out}(i) = 0, d_{in}(i) \neq 0, \quad \gamma_{ik} \propto \frac{n_{in}(i, k)}{m(i, k)}, \quad b_i = \frac{n_{in}(i)}{\sum_k m(i, k)\gamma_{ik}} \]

\[ \forall i, d_{out}(i) \neq 0, d_{in}(i) = 0, \quad \gamma_{ik} = \frac{n_{out}(i, k)}{\sum_k n_{out}(i, k)}, \quad b_i = 0 \]

\[ \forall i, d_{out}(i) = 0, d_{in}(i) = 0, \quad \gamma_{ik} \text{ is any non-negative value such that } \sum_k \gamma_{ik} = 1, \quad b_i = 0 \]

\[ n_{in}(i, k) = \sum_{j \in \mathcal{I}(i)} \hat{s}_{ji}q_{jik} \]

\[ n_{in}(i) = \sum_k n_{in}(i, k) \]

\[ n(i, k) = n_{in}(i, k) + n_{out}(i, k) \]

\[ n_{out}(i, k) = \sum_{j \in \mathcal{O}(i)} \hat{s}_{ij}q_{ijk} \]

\[ n_{out}(i) = \sum_k n_{out}(i, k) \]

\[ m(i, k) = \sum_{j \in \mathcal{L}(i)} \frac{n_{out}(j, k)}{\tau_{jk}} \]
Discriminative Content (DC) model

$$\Pr(z_i = k) = y_{ik} = \frac{\exp(w_k^T x_i)}{\sum_l \exp(w_l^T x_i)}$$

- Content vector: $x_i \in \mathbb{R}^d$
- Weight vector: $w_k \in \mathbb{R}^d$

• Replace $\gamma_{ik}$ by $y_{ik}$

$$\Pr(j| i; b, w) = \sum_k y_{ik} \frac{y_{jk} b_j}{\sum_{j' \in \mathcal{O}(i)} y_{j'k} b_{j'}}$$

$$\log \mathcal{L} = \sum_{(i \rightarrow j) \in \mathcal{E}} \hat{s}_{ij} \log \sum_k y_{ik} \frac{y_{jk} b_j}{\sum_{j' \in \mathcal{O}(i)} y_{j'k} b_{j'}}$$
Two stage method for optimization (1/4)

- **Expectation**
  - Compute $\tau_{ik}$ and $q_{ijk}$ from $Y$ and $b$

- **Maximization**
  - Maximize over $y \in \Delta$ rather than $W$

\[
\max_{\Delta, b} \sum_{(i, j) \in \mathcal{E}} \hat{s}_{ij} \sum_{k} q_{ijk} \left( \log y_{ik} y_{jk} b_j - \sum_{j' \in \mathcal{O}(i)} \frac{y_{j'k} b_{j'}}{\tau_{ik}} \right)
\]

\[
\Delta = \left\{ y \mid \exists w, y_{ik} = \frac{\exp(w_k^T \phi(x_i))}{\sum_l \exp(w_l^T \phi(x_i))} \right\}
\]

- $\phi(x) : \mathbb{R}^d \rightarrow \mathbb{R}^m$ a transformation on $x$
Two stage method for optimization (2/4)

• First Stage
  – Ignore $\Delta$ constrain, and use only “Sum-to-One” on $y_{ik}$ and $b_i \geq 0$ to derive $\tilde{y}_{ik}$ and $b_i$

• Second Stage
  – Minimizing the KL divergence between $\tilde{y}_{ik}$ and $y \in \Delta$

$$\max_w \sum_i \sum_k \tilde{y}_{ik} \log y_{ik} = \sum_i \sum_k \tilde{y}_{ik} \log \frac{\exp(w_k^T \phi(x_i))}{\sum_l \exp(w_l^T \phi(x_i))}$$

  – Add the regularization term

$$\max_w \sum_i \sum_k \tilde{y}_{ik} \log \frac{\exp(w_k^T \phi(x_i))}{\sum_l \exp(w_l^T \phi(x_i))} - \frac{\lambda}{2} \sum_k w_k^T w_k \quad (15)$$
Two stage method for optimization (3/4)

– It’s a multi-class logistic regression problem, can be maximized efficiently by the Newton-Raphson method
Two stage method for optimization (4/4)

- The full algorithm

**Algorithm 1** Algorithm for maximizing the log-likelihood

1. **Input** the number of iterations or convergence rate
2. Initialize $w_k$ to zeros, $b_i$ randomly, $\lambda$ to a fixed value
3. in the E-step, compute $\tau_{ik}$ and $q_{ijk}$ as in Eq. (6) and (7) using $y_{ik}$ rather than $\gamma_{ik}$
4. in the M-step,
   - compute $\gamma_{ik}$, and $b_i$ as in Theorem 4
   - update $w_k$ by maximizing the objective in Eq. (15) with $\gamma_{ik}$ in place of $\tilde{y}_{ik}$, and then compute $y_{ik}$
5. repeat Step 3 and 4 until the input number of iterations is exceeded or convergence rate is satisfied.
6. **Output** $\gamma_{ik}$ or $y_{ik}$ as the final membership
Extensions

- Combine link and content
  - PCL + PLSA
  - PHITS + DC

- Use above as baseline in experiments
Experiment
Dataset

- Political Blog Data Set
  - 1490 blogs, labeled as conservative or liberal (K=2)

- Wikipedia Data Set
  - 105 nodes and 799 links, no label (Set K=20 [13])

- Cora Data Set
  - 2708 nodes and 5429 links, 7 kind of labels (K=7)

- Citeseer Data Set
  - 3312 nodes and 4732 links, 6 classes (K=6)
Performance Metrics

• Recall
  – Report from positions 1 to 20 in the rank
• Normalized Mutual Information (NMI)

\[
NMI(C, C') = \frac{\text{MI}(C, C')}{\max(H(C), H(C'))}
\]

• Pairwise F-measure (PWF)

\[
\text{precision} = \frac{|S \cap T|}{|S|}, \quad \text{recall} = \frac{|S \cap T|}{|T|}
\]

\[
PWF = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

• Modularity (Modu)

\[
Modu(C) = \sum_k \left[ \frac{\text{Cut}(C_k, C_k)}{\text{Cut}(C, C)} - \left( \frac{\text{Cut}(C_k, C)}{\text{Cut}(C, C)} \right)^2 \right]
\]

• Normalized Cut (NCut)

\[
\text{NCut}(C_1, \ldots, C_k) = \sum_{i=1}^{K} \frac{\text{Cut}(C_i, \bar{C}_i)}{\text{vol}(C_i)}
\]
Link Prediction

• Hide some link in the network
  – The result also show the importance of popularity variable “b”

(a) Recall on Political Blog  (b) Recall on Wikipedia
(c) Recall on Cora         (d) Recall on Citeseer
## Community Detection (1/2)

### Table 1: Partition Measure on Cora and Citeseer Dataset

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>NMI</th>
<th>PWF</th>
<th>Modu</th>
<th>NCut</th>
<th>NMI</th>
<th>PWF</th>
<th>Modu</th>
<th>NCut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cora</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHITS</td>
<td>0.0570</td>
<td>0.1894</td>
<td>0.3929</td>
<td>3.2466</td>
<td>0.0101</td>
<td>0.1773</td>
<td>0.4588</td>
<td>2.2370</td>
</tr>
<tr>
<td>LDA-Link</td>
<td>0.0762</td>
<td>0.2278</td>
<td>0.2189</td>
<td>4.5687</td>
<td>0.0356</td>
<td>0.2363</td>
<td>0.2911</td>
<td>3.7457</td>
</tr>
<tr>
<td>PCL</td>
<td>0.0884</td>
<td>0.2055</td>
<td>0.5903</td>
<td>1.9391</td>
<td>0.0315</td>
<td>0.1927</td>
<td>0.6436</td>
<td>1.1181</td>
</tr>
<tr>
<td>NCUT</td>
<td>0.1715</td>
<td>0.2864</td>
<td>0.2701</td>
<td><strong>0.2732</strong></td>
<td>0.1833</td>
<td>0.3252</td>
<td>0.6577</td>
<td><strong>0.1490</strong></td>
</tr>
<tr>
<td>Citeseer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLSA</td>
<td>0.2107</td>
<td>0.2864</td>
<td>0.2682</td>
<td>4.2686</td>
<td>0.0965</td>
<td>0.2298</td>
<td>0.2885</td>
<td>3.2294</td>
</tr>
<tr>
<td>LDA-Word</td>
<td>0.2310</td>
<td>0.2774</td>
<td>0.2970</td>
<td>3.7820</td>
<td>0.1342</td>
<td>0.2880</td>
<td>0.3022</td>
<td>3.0165</td>
</tr>
<tr>
<td>NCUT(RBF kernel)</td>
<td>0.1317</td>
<td>0.2457</td>
<td>0.1839</td>
<td>4.7775</td>
<td>0.0976</td>
<td>0.2386</td>
<td>0.2133</td>
<td>3.7078</td>
</tr>
<tr>
<td>NCUT(pp kernel)</td>
<td>0.1804</td>
<td>0.2912</td>
<td>0.2487</td>
<td>4.6612</td>
<td>0.1986</td>
<td>0.3282</td>
<td>0.4802</td>
<td>1.8118</td>
</tr>
<tr>
<td>Link + Content</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PHITS-PLSA</td>
<td>0.3140</td>
<td>0.3526</td>
<td>0.3956</td>
<td>3.2880</td>
<td>0.1188</td>
<td>0.2596</td>
<td>0.3863</td>
<td>2.7397</td>
</tr>
<tr>
<td>LDA-Link-Word</td>
<td>0.3587</td>
<td>0.3969</td>
<td>0.4576</td>
<td>2.8906</td>
<td>0.1920</td>
<td>0.3045</td>
<td>0.5058</td>
<td>2.0369</td>
</tr>
<tr>
<td>LCF</td>
<td>0.1227</td>
<td>0.2456</td>
<td>0.1664</td>
<td>4.8101</td>
<td>0.0934</td>
<td>0.2361</td>
<td>0.2011</td>
<td>3.6721</td>
</tr>
<tr>
<td>NCUT(RBF kernel)</td>
<td>0.2444</td>
<td>0.3062</td>
<td>0.3703</td>
<td>1.6585</td>
<td>0.1592</td>
<td>0.2957</td>
<td>0.4280</td>
<td>1.7592</td>
</tr>
<tr>
<td>NCUT(pp kernel)</td>
<td>0.3866</td>
<td>0.4214</td>
<td>0.5158</td>
<td>0.7903</td>
<td>0.1986</td>
<td>0.3282</td>
<td>0.4802</td>
<td>1.8118</td>
</tr>
<tr>
<td>PCL-PLSA</td>
<td>0.3900</td>
<td>0.4233</td>
<td>0.5503</td>
<td>2.1575</td>
<td>0.2207</td>
<td>0.3334</td>
<td>0.5505</td>
<td>1.6786</td>
</tr>
<tr>
<td>PHITS-DC</td>
<td>0.4359</td>
<td>0.4526</td>
<td>0.6384</td>
<td>1.5165</td>
<td>0.2063</td>
<td>0.3295</td>
<td>0.6117</td>
<td>1.2074</td>
</tr>
<tr>
<td>PCL-DC</td>
<td><strong>0.5123</strong></td>
<td><strong>0.5450</strong></td>
<td><strong>0.6976</strong></td>
<td><strong>1.0093</strong></td>
<td><strong>0.2921</strong></td>
<td><strong>0.3876</strong></td>
<td><strong>0.6857</strong></td>
<td><strong>0.7505</strong></td>
</tr>
</tbody>
</table>
Community Detection (2/2)

- $\lambda = 5$ achieve the highest performance

(a) Cora
(b) Citeseer

Figure 2: Partition Measure of PCL-DC vs. $\lambda$
Conclusion
Conclusion

• Link Analysis → popularity → Conditional Model

• Content Analysis → irrelevant attribute → Discriminative Model
Open Question

• In this paper, they claimed that the probability of a node to be linked is by its popularity. Do you agree with that?
• Can you defined more attribute that may also affect the probability?
• Please also defined the conditional probability and the likelihood function
Hypertext Induced Topic Search

- Query based Search
- Authority & Hub Rank
  - A good hub points to many good authorities
  - A good authority is pointed to by many good hubs