Social Influence Local for Modeling Retweeting Behaviors

Jing Zhang, Biao Liu, Jie Tang, Ting Chen and Juanzi Li
Department of Computer Science, Tsinghua University

Social influence occurs when one's emotions, opinions, or behaviors are affected by others. Influence is local in most cases such as retweet behavior.

Influence locality function:

\[ Q(S_v, G^\tau_v) \] with \( \tau \in \mathbb{N}^+ \)

where \( G^\tau_v \) is v's \( \tau \)-ego network. \( S_v \) is the active neighbors in \( G^\tau_v \).

**Influence Test**

- Sina Weibo Retweet Data: 1,776,950 users, 308,489,739 follow relationships, 300,000 original microblogs, and 23,755,810 retweets.
- Test:
  - Treatment group: Users with \( \geq 1 \) active neighbors.
  - Control group: Users with \( = 1 \) active neighbors.
- To avoid the selection bias:
  - For each user in treatment group, find the most matched user from the original control group.
  - Learn a logistic regression model to estimate the probability of each user being treated. Matching is based on the probability.
- To avoid the confounding bias:
  - Construct the two groups for each microblog and each time span after the microblog being published independently.

**Influence Measure**

\[ Q(S_v, G^\tau_v) = w \times g(S_v, G^\tau_v) + (1 - w) \times f(S_v, G^\tau_v) \]

- **Pairwise Influence**
  - \( F > B \) for influence on v?
  - B only has one path to reach v
  - F has a number of paths to connect v

\[ g(S_v, G^\tau_v) = \sum_{v_i \in S_v} p_{v_i} \]

\( p_{v_i} \) is random walk probability from \( v_i \) to v

- **Structure Influence**
  - \( C + D > A + B \) for influence on v?
  - A and B distribute into different circles
  - C and D reside in the same circle

\[ f(S_v, G^\tau_v) = e^{-\mu|C(S_v)|} \]

\( C(S_v) \) is the number of circles formed by \( S_v \)

**Results**

**Performance of retweet behavior prediction (%)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRC-B</td>
<td>68.11</td>
<td>74.26</td>
<td>71.36</td>
<td>69.74</td>
</tr>
<tr>
<td>LRC-Q</td>
<td>66.82</td>
<td>77.22</td>
<td>71.65</td>
<td>69.44</td>
</tr>
<tr>
<td>LRC-BQ</td>
<td>69.89</td>
<td>77.06</td>
<td>73.30</td>
<td>71.93</td>
</tr>
</tbody>
</table>

LRC-B: Logistic regression classifier with only basic features (e.g., gender, verification status and so on).

LRC-Q: Logistic regression classifier with only influence locality function. \((w=0.5, g=g_6)\)

LRC-BQ: Combine basic features and influence locality function together.

**Performance with and without structure influence (%)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>LRC-Q(w=1)</td>
<td>49.51</td>
<td>51.53</td>
<td>50.50</td>
<td>49.49</td>
</tr>
<tr>
<td>LRC-Q(w=0.5)</td>
<td>51.86</td>
<td>67.70</td>
<td>68.73</td>
<td>52.43</td>
</tr>
</tbody>
</table>

80% instances only have one active neighbors, thus the effect of structure influence can not be presented. So we sample instances with the number of active neighbors larger than 5.

**Performance with different pairwise functions (%)**

The active neighbors with different retweet time exert different influence on retweet behaviors.

The majority of pairwise influence is low and a minority is scattered in a fat right tail, thus geometric mean performs better.