消息链中的特定用户潜在影响力及敏感度学习

王超民

2016 年 5 月 25 日

本文研究的问题是从消息链中推测用户的影响力和敏感度，提出了一个数学模型LIS（Latent Influence and Susceptibility）。相较于之前的数学模型，LIS解决了严重的过拟合问题和充分利用了信息传播的上下文信息。

过拟合问题源于在以往的模型中的一个假设：不同用户对之间的人际影响是独立无关的。LIS模型则针对特定用户行为，对用户之间的影响力建模。具体的，用两个d维的向量：影响力向量I_u和敏感度向量S_u，来表示用户u在d个潜在话题的影响力和敏感度。用户对(u,v)之间的人际影响可以通过计算I_u和S_u的内积得到。这种简明的表示方法对于n个用户而言，只需要$2nd(\ll n_2)$个参数。

模型描述

给定一条消息m，记它的转发链C^m为按转发时间先后顺便排列的用户列表$(a_1^m,\ldots,a_N^m)$。如果用户发布或者转发了这条消息，那么我们称用户是活跃的，并且他有一次机会去“激活”其他用户。至于“激活成功与否”，取决于当前时刻的上下文链。

上下文链：当活跃用户$a_i^m(i=1,\ldots,N)$试图去“激活”某用户v时，此时的上下文链为

$$D_{v,i}^m=\left\{a_j^m \mid j\le i,\delta(a_j^m,v)=1\right\}, \tag{1}$$

其中指示器函数$\delta(u,v)$表示来自u的消息是否可以传递至v，我们将用过去所有的转发链和社会影响网络的叠加网络来表示，如图1所示。总之，上下文链代表着用户v此前接收到这条消息的那些用户列表。

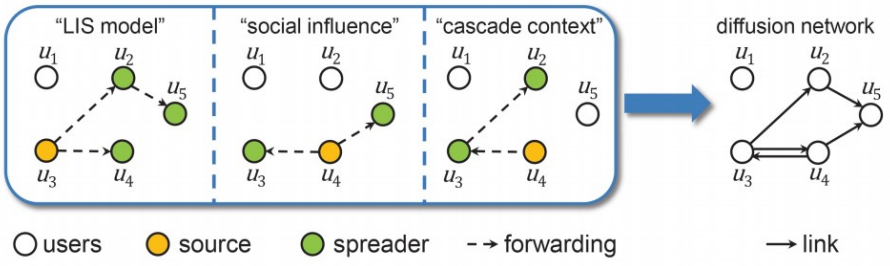

图 1: 图1 转发链的叠加

对于消息m及它的转发链$(a_1^m,\ldots,a_N^m)$，每个用户可以用一个N维的状态向量z_v^m来表示，该状态向量的元素$z_{v,j}^m$代表用户v是否在接收到来自用户a_j^m的消息m后处于活跃状态。如果用

户v从用户a_j^m的消息m后处于活跃状态，那么$z_{v,i}^m = 0 \ \ (1 \le i < j)$，$z_{v,i}^m = 1 \ \ (j \le i < N)$。如果$v$在整个$m$的转发链都处于非活跃状态，那么对于任意的$i$，$z_{v,i}^m = 0$。

z_v^m的似然函数为

$$P(z_v^m \mid \delta) = p(z_{v,0}^m) \prod_{i=1}^{N} p(z_{v,i}^m \mid z_{v,i-1}^m, D_{v,i}^m, \delta) \tag{2}$$

其中$z_{v,0}^m$表示v是否为消息m的始发者

$$p(z_{v,0}^m = 1) = \begin{cases} 1, v \text{ is the source} \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

$$\begin{aligned} &p(z_{v,i}^m = 1 \mid z_{v,i-1}^m = 1, D_{v,i}^m, \delta) = 1 \\ &p(z_{v,i}^m = 1 \mid z_{v,i-1}^m = 0, D_{v,i}^m, \delta) = 1 - exp(-\lambda \delta(a_v^m, v) \sum_{u \in D_{v,i}^m} I_u^T S_v) \\ &p(z_{v,i}^m = 0 \mid z_{v,i-1}^m = 0, D_{v,i}^m, \delta) = 1 - p(z_{v,i}^m = 1 \mid z_{v,i-1}^m = 0, D_{v,i}^m, \delta) \end{aligned} \tag{4}$$

其中λ是比例因子，用来调整上下文链的影响。公式(4)说明了转移概率是如何受到上下文链的影响。

假设转发链不相关，则所有转发链C的似然函数为公式(2)的乘积

$$L(C) = \prod_{m=1}^{|C|} \prod_{v \in V} P(z_v^m \mid \delta) \tag{5}$$

LIS模型的参数通过最小化对数似然代价函数得到

$$Ł(C) = -\sum_{m=1}^{|C|} \sum_{v \in V} \sum_{i=1}^{N} log\ p(z_{v,i}^m \mid z_{v,i-1}^m, D_{v,i}^m, \delta) \tag{6}$$

其中$p(z_{v,0}^m)$总是为1。

参数估计

一般来说，可以直接最小化公式(6)来完成参数I和S的估计。但是带来的过高计算量我们无法承受。同时，一条上下文链会在转发链中重复出现很多次，导致公式(4)的大量重复计算。因此，我们将充分利用在多条转发链中的重叠上下文链来减少重复计算。

记$\Gamma(v)$为关于v在所有转发链中的所有可能的上下文链。对上下文链进行用户分组，重组公式(6)的对数似然函数为

$$Ł(C) = -\sum_{v \in V} \sum_{D_{v,i} \in \Gamma(v)}^{N} (n_{z_{v,i}, D_{v,i}} log\ p(z_{v,i} \mid z_{v,i-1}, D_{v,i}, \delta)) \tag{7}$$

其中$D_{v,i}$表示用户v的一条与特定转发链无关的上下文链，$n_{z_{v,i}, D_{v,i}}$为用户v在状态$z_{v,i}$下，$D_{v,i}$出现在所有转发链的频率。

为了避免过拟合，我们需要对公式(7)增加关于参数向量I和S的正则项，得到最终的参数估计目标函数

$$\begin{aligned} &Ł(C) = -\sum_{v \in V} \sum_{D_{v,i} \in \Gamma(v)}^{N} (n_{z_{v,i}, D_{v,i}} log\ p(z_{v,i} \mid z_{v,i-1}, D_{v,i}, \delta)) + \gamma_I \left\| I \right\|_F^2 + \gamma_S \left\| S \right\|_F^2 \\ &s.t.\ I_{ij} \ge 0, S_{ij} \ge 0, \forall i, j \end{aligned} \tag{8}$$

其中γ_I和γ_S为正则系数，$\|\cdot\|_F$为Frobenius范数。

最后，利用PG(Projected Gradient)法，我们设计了一种迭代算法完成参数估计。关于I和S的梯度

$$\begin{aligned}\frac{\partial \text{Ł}}{I_u} &= -\lambda \sum_{v\in V} S_v \sum_{D_{v,i}\in\Gamma(v)} \Phi_{u\in D_{v,i}}(n_{z_{v,i}=1,D_{v,i}} \frac{1-p_{v,D_{v,i}}}{p_{v,D_{v,i}}} - n_{z_{v,i}=0,D_{v,i}}) + \gamma_I I_u \\ \frac{\partial \text{Ł}}{S_v} &= -\lambda \sum_{D_{v,i}\in\Gamma(v)} \sum_{u\in D_{v,i}} I_u(n_{z_{v,u}=1,D_{v,u}} \frac{1-p_{v,D_{v,u}}}{p_{v,D_{v,u}}} - n_{z_{v,u}=0,D_{v,u}}) + \gamma_S S_v\end{aligned} \tag{9}$$

其中Φ为指示函数，$p_{v,D_{v,i}}$是$p(z_{v,i}=1 \mid z_{v,i-1}, D_{v,i}, \delta)$的简写。参数估计的算法见Algorithm 1。

Algorithm 1 参数估计

Input: Collection of cascades observed in a given time period, the maximum epoch M, and regularization parameters γ_I and γ_S

Output: User-specific influence and susceptibility I, S

Construct diffusion network δ from cascades

Initialize parameters with random values, including I, S

Repeat

For $i = 1$ to n

Calculate gradients $\partial L/\partial I_u$ and $\partial L/\partial S_v$

Update I and S with PG method

Until maximum epoch M is reached or gradients vanish