ВЕРОЯТНОСТНОЕ МОДЕЛИРОВАНИЕ ПРОФИЛЕЙ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНОЙ СЕТИ

[1,2]Гомзин А. Г., [1,2,3]Турдаков Д. Ю.

([1]*ИСП РАН, Москва,* [2]*МГУ им. М.В. Ломоносова, Москва*
[3]*НИУ ВШЭ, Москва*)

Профиль пользователя социальной сети содержит значения социо-демографических атрибутов, таких как пол, возраст, семейное положение и т.д. Однако не все значения указываются пользователями или указываются, но неверно. В связи с этим возникает необходимость предсказания скрытых и выявления ложных значений в профилях.

Социальная сеть представляет собой направленный граф (V, E). Вершины графа V представляют различные объекты-сущности: профили, сообщества, сообщения. Рёбра E графа моделируют направленные связи, например, подписки пользователей на других пользователей или сообщества. Каждой вершине $v \in V$ сопоставлен набор атрибутов (в т.ч. социо-демографических) и их значений.

Для предсказания демографических атрибутов обычно используются подходы, основанные на машинном обучении и распространении меток в социальном графе [1][2]. Мы предлагаем подход, основанный на вероятностном моделировании социального графа.

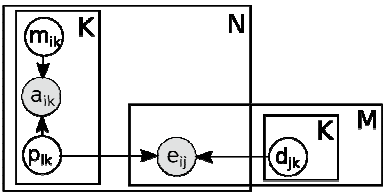

Схема вероятностной модели в графической нотации представлена на рисунке. Рассматриваются M вершин социального графа, N из которых являются пользователями ($N \le M$), K демографических атрибутов. a_{ik} - явно указанное пользователем i значение атрибута k. Если пользователь не указал значение атрибута, то $a_{ik} = \varepsilon$. e_{ij} - наличие или отсутствие направленного ребра между пользователем i и объектом j. Это наблюдаемые случайные величины модели. Далее описывается порождающий процесс и скрытые параметры модели:

1. Для каждого пользователя i и атрибута k генерируется $m_{ik} \in \{T, F, \varepsilon\}$, в зависимости от того, указывает ли пользователь истинное (T), ложное (F), или не указывает (ε) значение атрибута k;

2. Для каждого пользователя i и атрибута k генерируется p_{ik} - истинное значение данного атрибута k пользователя i;

3. Генерируются $a_{ik} \sim P(a_{ik} \mid m_{ik}, p_{ik})$. Если $m_{ik} = \varepsilon$, то $a_{ik} = \varepsilon$; если $m_{ik} = T$, то $a_{ik} = p_{ik}$; иначе $a_{ik} \notin \{p_{ik}, \varepsilon\}$;

4. Для каждого объекта j и атрибута k генерируется d_{ik} - предпочтение, т.е. значение атрибута пользователей, которые предпочитают создавать связи с этим объектом. Принимает одно из значений атрибута k или ε (нет предпочтения).

5. Генерируются рёбра $e_{ij} \sim P(e_{ij} \mid p_{ik}, d_{jk})$, вероятность ребра зависит от истинного профиля пользователя и предпочтений соответствующего объекта.

Для предсказания и исправления значений атрибутов необходимо обозначить априорные распределения при генерации $m_{ik}, p_{ik}, d_{jk}, e_{ij}$ при заданных d_{ik}. Затем необходимо оценить совместную вероятность всех параметров модели и вывести значения скрытых параметров.

Предложенный подход позволяет моделировать ложно указанные значения атрибутов, а также задавать априорное распределение для p_{ik}, учитывающее совместную встречаемость различных атрибутов (например, возраста и уровня образования).

Литература

1. Гомзин А. Г., Кузнецов С. Д. *Методы построения социодемографических профилей пользователей сети Интернет* //Труды Института системного программирования РАН. – 2015. – Т. 27. – №. 4.
2. Phillips L. et al. *Using social media to predict the future: a systematic literature review* //arXiv preprint arXiv:1706.06134. – 2017.