Теория графов и её приложения

Командный проект

Воронкова Ева Боруховна, Вольф Дмитрий Александрович

24/04/2023

Содержание

Проект

Командный проект

Тема Предсказание появления рёбер в темпоральных (временных) графах

Кратко Предсказать, появится ли ребро между парой вершин (u, v) к моменту времени t'', если на момент времени t^* ребро между этими вершинами отсутствовало.

Идея Bruin, Gerrit Jan de; Veenman, Cor J.; van den Herik, H. Jaap; Takes, Frank W. (2021): Supervised temporal link prediction in large-scale real-world networks. In Soc. Netw. Anal. Min. 11 (1). DOI: 10.1007/s13278-021-00787-3.

Данные Сети с http://konect.cc/networks/

Предварительные замечания

- Все графы считаются неориентированными (направления ребер игнорируются).
- Для графов с "положительными" и "отрицательными" связями учитываются только "положительные" рёбра.
- Связи в графах могут рассматриваться, как непрерывные (persistent relations) или дискретные (discrete events). В последнем случае сеть представляет собой мультиграф.

Обозначения

- Темпоральный (временной) — граф, для каждого ребра которого указана временная отметка в промежутке от t' до t'' (время появления ребра в графе):

  $H_{[t',t'']} = (V, E_H)$ с множеством рёбер
  $E_H = \{(u, v, t) \mid u, v \in V,\ t' \leqslant t \leqslant t''\}$.

- Статический граф $G = (V, E_G)$ — полученный из графа $H_{[t',t'']}$, путем сохранения всех рёбер, появившихся в графе от t' до t'' (без учета кратности рёбер).

- Связи в графах могут рассматриваться, как непрерывные (persistent relations) или дискретные (discrete events). В последнем случае сеть представляет собой мультиграф.

Свойства сетей

Свойства сетей (для статических графов) i

Для каждого *статического* графа вычислить следующие характеристики:

1. Число вершин, число рёбер, плотность (отношение числа рёбер к максимально возможному числу рёбер), число компонент слабой связности, долю вершин в максимальной по мощности компоненте слабой связности.
2. Для наибольшей компоненты слабой связности вычислить/оценить значения радиуса, диаметра сети, 90 процентиля расстояния (геодезического) между вершинами графа. Для больших графов оценку провести на основании:
   - 2.1 вычисления расстояний между 500(1000) случайно выбранными вершинами из наибольшей компоненты слабой связности;
   - 2.2 вычисления расстояний по подграфу "снежный ком"(snowball sample), построенного по следующему принципу: выбирается небольшое начальное множество вершин (1, 2 или 3), затем в граф добавляются все их соседи, затем соседи соседей и т.д., пока число вершин в подграфе не станет равным (примерно) заданному значению (например, 500 или 1000).

Свойства сетей (для статических графов) ii

3. Вычислить средний кластерный коэффициент $\bar{Cl}$, где

$$\bar{Cl} = \frac{1}{|V|}\sum_{i \in V} Cl_i, \quad Cl_i = \begin{cases} \dfrac{2e_i}{k_i(k_i-1)}, & k_i \geqslant 2 \\ 0, & \text{иначе.} \end{cases},$$

где Cl_i — локальный кластерный коэффициент вершины v_i, k_i — степень вершины, e_i — число ребер между соседями v_i.

4. Коэффициент ассортативности по степени вершины $-1 \leqslant r \leqslant 1$ (коэф. корреляции Пирсона):

$$r = \frac{\text{cov}}{\text{var}} = \frac{\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})k_i k_j}{\sum_{ij}(k_i\delta_{ij} - \frac{k_i k_j}{2m})k_i k_j} = \frac{R_e R_1 - R_2^2}{R_3 R_1 - R_2^2}$$

$$R_1 = \sum_i k_i = 2m, \quad R_2 = \sum_i k_i^2, \quad R_3 = \sum_i k_i^3, \quad R_e = \sum_{ij} A_{ij} k_i k_j.$$

Предсказания появления ребер в графе

Задача предсказания появления ребра

- Основная задача проекта — предсказать, появится ли ребро между парой вершин (u, v) к моменту времени t'', если на момент времени $t*$ ребро между этими вершинами отсутствовало.
- Рассматриваются все несмежные пары вершин (u, v) в сети $H_{[t_{q=0}, t_{q=s}]}$ и предсказывается появление ребра в сети $H_{[t_{q=s}, t_{q=1}]}$.
- $t_{q=0}, t_{q=1}$ — время появления первого и последнего ребра в сети, $t_{q=s}, 0 < s < 1$ — q-тый процентиль временных отметок для сети.
- Каждой паре несмежных ребер сопоставляется вектор признаков $X_{(u,v)}$, а также ответ $y_{(u,v)} = 1$, если вершины стали смежными, и $y_{(u,v)} = 0$, если ребро не появилось.
- Задача бинарной классификации

Признаки. Static topological features (I)

- Статические признаки (Static topological features) (п. 4.1.1)
- 4 признака для каждой пары вершин: Common Neighbours (CN); Adamic-Adar (AA); Jaccard Coefficient (JC); Preferential Attachment (PA)
- Каждая команда вычисляет эти признаки!

Признаки. Temporal topological features (II)

II Темпоральные признаки (Temporal topological features) (п. 4.1.2)

A Вычисление весов (Temporal weighting): пересчет временных отметок в веса по формулам

B Учет событий прошлого (The proposed approach of past event aggregation)

! только для сетей с дискретными событиями (discrete events)

C Вычисление взвешенных темпоральных признаков (Computation of weighted topological features)

Признаки. Temporal topological features (II)

II Темпоральные признаки (Temporal topological features) (п. 4.1.2)

B Учет событий прошлого (The proposed approach of past event aggregation)

! только для сетей с дискретными событиями (discrete events)

Для каждого ребра с $t \leqslant t_{q=s}$ вычисляются веса (см. п. A), а затем данные агрегируются, т.е. вычисляются 8 функций: $0, 1, 2, 3, 4$ квартили, сумма, среднее и дисперсия.

C По каждому из 8 значений вычисляются 4 взвешенных признака.

Признаки. Node activity features (III)

III Учет активности вершин (Node activity features) (п. 4.1.3)

1. Вычисление весов (Temporal weighting) (см. п. A из 4.1.2)
2. Агрегация данных для каждого узла, т.е. вычисляются 7 функций по смежным заданной вершине ребрам:

   Для каждого ребра с $t \leqslant t_{q=s}$ вычисляются веса (см. п. A), а затем данные агрегируются, т.е. вычисляются 7 функций: $0, 1, 2, 3, 4$ квартили, сумма, среднее.
3. Для каждой пары несмежных вершин (потенциального ребра) вычисляется сумма, модуль разности, минимальное и максимальное значения.

Задача бинарной классификации

Задача бинарной классификации

- Задача бинарной классификации (появится ребро или нет)

  грубо: в пространстве признаков $\mathbb{R}^p$ строим гиперплоскость такую, что "выше" лежат объекты, которые относятся к классу 1, ниже — к классу 0).

- Для обучения модели применяется алгоритм логистической регрессии — статичтическая модель для прогнозирования вероятности возникновения некоторого события путём подгонки данных к логистической кривой.

- Классификатор выдаёт для каждого объекта вероятность того, что объект принадлежит к определенному классу. Далее, по принятому порогу объекты делятся на классы.

- Для оценки качества построенной модели используется метрика AUC (Area Under the Receiver Operating Curve) – площадь под ROC AUC – кривой.

Задача бинарной классификации

- Задача бинарной классификации (появится ребро или нет).
- ROC AUC - кривая строится на основании соотношения доли верно классифицированных объектов, обладающих некоторым свойством (TPR true positive rate) и доли объектов, не обладающих свойством, но ошибочно классифицированных как обладающие этим свойством (FPR false positive rate), при различных уровнях порога принятия решения.

TPR Доля рёбер, которые появились в графе $y_{(u,v)} = 1$ и которые классификатор отметил, как появившиеся в графе.

FPR Доля рёбер, которые не появились в графе, но которые были отмечены, как появившиеся.

Задача бинарной классификации

- Для больших сетей для выборки
- выбираем только вершины, находящиеся на расстоянии 2 друг от друга;
- выбираем (с заменой) по 10000 пар вершин, между которыми образуется ребро, и тех, между которыми ребро не образуется.
- Полученную выборку делим на обучающую (75%) и тестовую.
- Для обучения используем готовые решения например, из Python scikit-learn package со стандартным набором параметров.

Полезные ссылки и материалы

- Bruin, Gerrit Jan de; Veenman, Cor J.; van den Herik, H. Jaap; Takes, Frank W. (2021): Supervised temporal link prediction in large-scale real-world networks. In Soc. Netw. Anal. Min. 11 (1). DOI: 10.1007/s13278-021-00787-3.
- Данные http://konect.cc/
- Логистическая регрессия (пример).
- Лекции по МО (2016б ВШЭ).
  Лекция 04 — метрики качества, 05 — логистическая регрессия.