Задание по курсу “Теория графов и ее приложения” v.2023

Часть 2. Командный проект

Задание основано на вычислении метрик рёбер и применении их для предсказания появления рёбер в темпоральных (временных) графах. Идея задания взята из статьи

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

Для работы предлагаются датасеты, размещенные на http://konect.cc/networks/ и представляющие собой реальный сети различной природы (социальные, информационные, технологические).

Каждый датасет – это темпоральный (временной) граф, где для каждого ребра указана временная отметка в промежутке от t' до t'' – время появления ребра в графе.

Все графы считать неориентированными.

$H_{[t',t'']}(E, V_H)$ – временной граф с множеством рёбер $E_H = \{(u, v, t),\ u, v \in V, t' \leq t \leq t''\}$.

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

Список датасетов для работы будет размещен в Teams.

1. Свойства сетей (для статических графов)

Для каждой из сетей 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. Оценку провести на основании
   a. вычисления расстояний между 500 (1000) случайно выбранными вершинами из наибольшей компоненты слабой связности;
   b. вычисления расстояний по подграфу "снежный ком" (snowball sample), построенного по следующему принципу: выбирается небольшое начальное множество вершин (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).
3. Для наибольшей компоненты слабой связности вычислить средний кластерный коэффициент сети $\overline{Cl} = \frac{1}{|V|}\sum_{u \in G} Cl_u$, где

$$Cl_u = \begin{cases} \frac{2L_u}{|\Gamma(u)|\cdot|\Gamma(u)-1|}, & |\Gamma(u)| \geqslant 2 \\ 0, & \text{иначе.} \end{cases},$$

где $\Gamma(u)$ – множество соседей вершины, $|\Gamma(u)| = k_u$ –степень вершины, L_u --- число ребер между соседями.

4. Коэффициент ассортативности по степени вершин $-1 \leq r \leq 1$ (коэффициент корреляции Пирсона)

$$r = \frac{\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)k_i k_j}{\sum_{i,j}\left(k_i\delta_{ij} - \frac{k_i k_j}{2m}\right)k_i k_j} = \frac{R_e R_1 - R_2^2}{R_3 R_1 - R_2^2},$$

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

Случаю, когда в сети вершины малой степени чаще соединяются с вершинами большой степени, соответствуют отрицательные значения коэффициента $r < 0$.

2. Предсказания появления ребер в графе

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, т. е. фактически, решается задача бинарной классификации.

Для предсказания необходимо сначала построить признаковое описание для каждого потенциального ребра (вектор/набор признаков) $X_{(u,v)}$, а также ответ $y_{(u,v)}$, который принимает значение $y_{(u,v)} = 1$, если ребро появляется в графе, и $y_{(u,v)} = 0$, иначе.

А. Построение векторов признаков для предсказания появления ребер в графе

В работе [1] предлагается несколько вариантов построения признакового описания, а затем сравнивается качество полученных предсказаний.

Каждая команда должна построить два различных набора признаков: статический (I) и один из наборов, обозначенных в статье (II-A, II-B, III).

Номера наборов для реализации будут указаны в таблице в MS Teams.

Б. Бинарная классификация.

Для обучения модели применяется алгоритм логистической регрессии. Для оценки качества построенной модели используется метрика AUC (Area Under the Receiver Operating Curve) – площадь под ROC AUC – кривой.

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.

В качестве решения будет засчитано следующее:

- Результаты работы алгоритмов на графах малой размерности (тестовые примеры для п. 1, а также вычисление структурных признаков (набор I)).
- Отчет-презентация, содержащий достаточно информации по использованным инструментам и алгоритмам для решения задачи, а также результаты исследования и выводам.
- Необходимое условие для получения зачёта: решение задачи классификации для не менее 4 датасетов (при условии правильной работы кода на тестовых примерах).
- Продумать интерфейс так, чтобы можно было демонстрировать работу алгоритма (вывод результатов, вычисление расстояний между некоторой парой вершин, вычисление локального кластерного коэффициента и т.п.)
- Результаты исследования (сравнений) работ алгоритмов для вычисления диаметра графа (расстояния между парами вершин), метрик рёбер и т.п.

Дополнительные условия

- При выполнении задания запрещено пользоваться готовыми решениями.
- Готовые решения могут быть использованы для импорта/экспорта данных, построения графиков, а также для решения задачи бинарной классификации.

Общие требования к выполнению заданий

1. Допускается работа в группах не более 3-х человек.
    a. После 01.05.2023 студенты, не отметившиеся в файле в составе той или иной команды, считаются работающими самостоятельно.
    b. В одну команду могут входить студенты из разных учебных групп.
    c. Все особые ситуации (например, распад команды из-за непреодолимых разногласий) обсуждаются отдельно с преподавателем до даты защиты проекта (зачёта). Если вы не уверены, является ли ваша ситуация особой, всё равно напишите преподавателям, например, в чате в Telegram, MS Teams, на e-mail.
2. По результатам выполнения проекта каждая команда должна подготовить отчет-презентацию и сделать доклад.
3. Каждая команда должна опубликовать свое решение в репозитории в ветку с номером команды Team_nn.
    Ссылка на GitHub: https://github.com/answerIII/GraphTheory2022-2023/blob/main/README.md .
    Номер команды указывается согласно порядковому номеру в файле Finite-Graphs-20Б11-13-MSTeams.xlsx .
4. При выставлении итоговых баллов учитывается качество и количество проделанной работы (использованные структуры данных, наличие/отсутствие сравнений работы алгоритмов с готовыми решениями, количество обработанных датасетов и т.п.), история коммитов, ответы на вопросы по проекту.

5. При выполнении работы студенты следуют Кодексу универсанта, в частности соблюдают нормы научной этики, уважают права интеллектуальной собственности и не используют недобросовестных методов при прохождении аттестации (п. 4, 6 и 7).
6. Преподаватели оставляют за собой право выдать дополнительное задание, если возникли сомнения в самостоятельности выполнения студентом (группой) задания.

Сроки выполнения задания **25.05.2023**

Воронкова Ева Боруховна e.voronkova@spbu.ru

Вольф Дмитрий Александрович answer.iii@mail.ru