Санкт-Петербургский государственный университет
Факультет прикладной математики процессов управления

Прикладные задачи машинного обучения на графах

Лабораторная №1. Отчет.

Выполнили:
Студент, каф. теории управления, 2 курс,
Кикенов А.А.

Студент каф. информационных систем, 2 курс,
Касаткин С.М.

Преподаватель:
кандидат физ-мат наук, Доцент каф. мат.моделирования энергитических систем
Е.Б.Воронкова

Санкт-Петербург

2024

Содержание

1 Постановка задачи

Задачей работы является предсказание появления рёбер в темпоральных (временных) графах. Идея задания взята из работы [?]).

2 Используемые материалы

Для работы используются датасеты из репозитория
https://snap.stanford.edu/data/temporal
и представляющие собой реальные сети различной природы (социальные, информационные, технологические). В частности "CollegeMsg"и "sx-mathoverflow". Также для написания кода активно использовался сервис Monica.

3 Свойства сетей

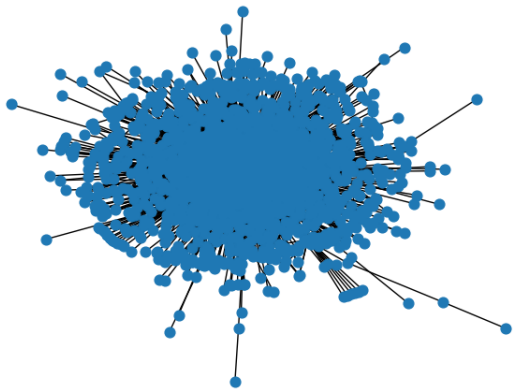
3.1

Были извлечены статические графы CollegeMsg и sx-mathoverflow, для которых были вычислены следующие признаки: число вершин, число ребер, плотность, число компонент слабой связности, доля вершин в максимальной по мощности компоненте связности. Для наибольшей компоненты слабой связности вычислены значения радиуса, диаметра сети, 90 процентиля расстояния (геодезического) между вершинами графа, а также кластерный коэффициент и коэффициент Пирсона.

3.2 Результаты

3.2.1 CollegeMsg

У основного графа G 1899 вершин и 13838 ребер, а у его большей компоненты связности F 1893 вершины и 13835 ребер.
Всего 4 компоненты связности, плотность графа - 0.0077,
Доля вершин в максимальной по мощности компоненте слабой связности: 0.5266,
Минимальная степень узла: 1,
Максимальная степень узла: 255,
Средняя степень узла: 14.6170,
Радиус графа 4,
Диаметр графа 8,
90-й процентиль расстояний: 4.0,
Средний кластерный коэффициент: 0.1097,



Наибольшая компонента связности F

Граф G

Рис. 1: Представление датасета CollegeMsg

Коэф. ассортативности r = -0.1880.

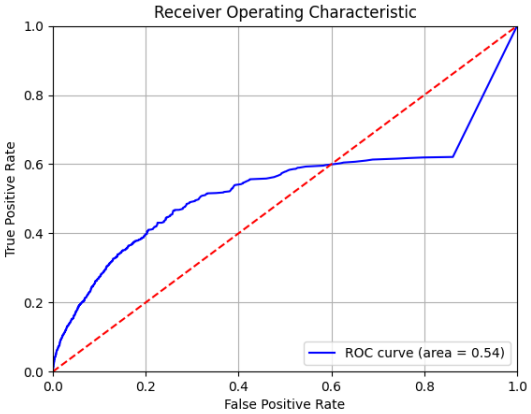
3.2.2 sx-mathoverflo

У основного графа G 24818 вершин и 199973 ребра, а у его его наибольшей компоненты связности F 24668 вершины и 187939 после удаления петель.
Число компонент связности: 104,
Плотность: 0.0006,
Доля вершин в максимальной по мощности компоненте слабой связности: 0.0403,
Минимальная степень узла: 1,
Максимальная степень узла: 2172,
Средняя степень узла: 15.2375.

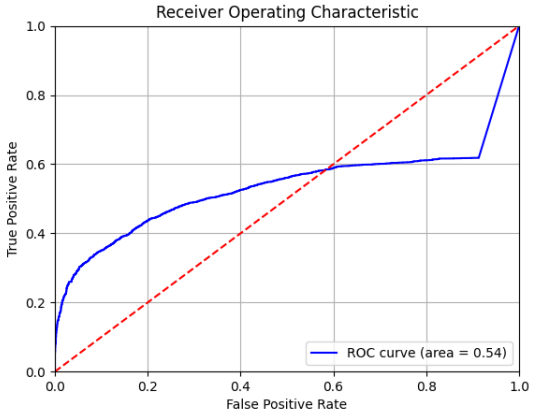
Для вычислени радиуса и диаметра было реализовано 2 алгоритма Снежного кома с ограничением N = 1000 для вершинисследуемого подграфа.

В первом алгоритме берутся все соседи у единственного узла-зерна. Для такого подграфа моно получить следующие числа: Радиус графа 2,
Диаметр графа 3.
90-й процентиль расстояний: 2.0,
Средний кластерный коэффициент: 0.0031
Коэф. ассортативности r = -0.9940

Во втором алгоритме берется каждый сотый сосед: Радиус графа 5,
Диаметр графа 10,
90-й процентиль расстояний: 7.0,
Средний кластерный коэффициент: 0.0023,
Коэф. ассортативности r = -0.3140.



Статический граф



Темпоральный граф

Рис. 2: ROC AUC

4 Предсказание появления ребер в графе

4.1 CollegeMsg

Граф был разбит на 2: граф, на котором происходит обучение, и граф, для которого предсказывается появление ребер. в соотношении 3/1 (75% персентиль)

4.1.1 Статический граф CollegeMsg

Было вычислено 4 статических признака для его. Точность прогнозирования составила 73.1%. Кривая ROC AUC изображена на картинке

4.1.2 Темпоральный граф CollegeMsg

Был вычислен линейный вес, затем было вычислено 4 темпоральных признака для его. Точность прогнозирования составила 62%

5 Заключение

Были рассмотрены два временных графа. Для них были вычислены метрики. Далее для одного из них были построены две модели, предсказывающие появление новых ребер: одна статическая, другая темпоральная. статическая получилась эффективнее темпоральной.
Во время выполнения задания для ускорения работы, а то есть объяснения ошибок, написания чернового кода и для поиска информации, – активно использовался искусственный интеллект Monica.

Список литературы

[1] Лекции https://disk.yandex.ru/d/xtPLW1KFF5Pf9w/lecture-notes-2024

[2] Логистическая рагрессия https://docs.yandex.ru/docs/view?url=ya-disk-public

[3] Данные для работы с графами https://snap.stanford.edu/data/temporal

Основной код:
часть 1:
https://colab.research.google.com/drive/1IMVKQ134YITmPjep89CrRIRN7zF7f59h
часть 2:
https://colab.research.google.com/drive/1Qk7xTJ6EF-W4OjfuyxHXA0GrjBZwF4t4revisionId=0B_g1BEgE6paxbUkyNWFzTUd0VkVPZm5ia1pKeHh4R3k2d2tJPQscrollTo=Iv_roj1VbSvT