Контролируемая генерация графов

Бишук Антон Юрьевич

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Москва
2023 г

Цели и задачи

Цель

Научиться генерировать графы с заданными статистиками.

Задачи

- Предложить модификацию метода генерации графа, которая позволит задавать некоторые стандартные числовые характеристики в явном виде.
- Теоретически обосновать работу предложенного метода.
- Провести сравнение предложенной модификации с существующими методами.

Определения

Простые признаки

Простые признаки (или же простые статистики) в нашем методе - это числовые характеристики используемые в теории графов, которые могут быть вычислены не более, чем за квадратичное время.

Смешанными признаки

Смешанными признаками назовем любой способ численно описать граф.

В нашей работе мы будем рассматривать не все возможные численные представления графа, а только те, которые можно получить в модели VAE в скрытом представлении.

Сложными признаки

Сложными статистиками назовем вектор $\vec{d}$, такой что каждая компонента вектора $\vec{d}$ независима от компонент вектора простых статистик $\vec{s}$ и при этом вектор смешанных статистик m выражается через $\vec{d}$ и $\vec{s}$ линейно.

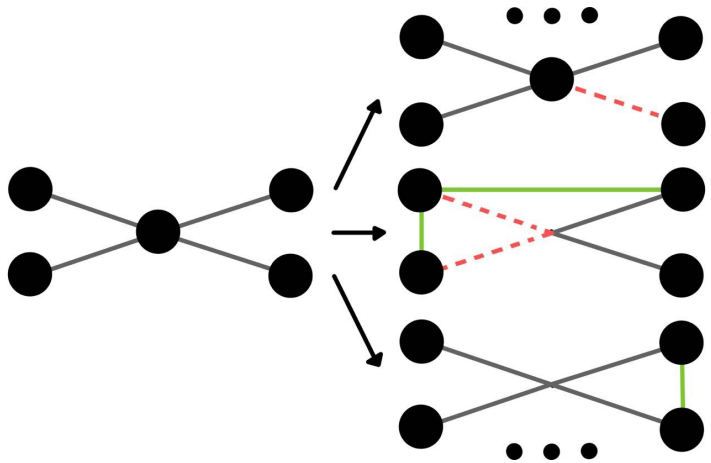
Простые статистики

- Размерные показатели $[O(1)]$:
  - Число ребер
  - Число вершин
- Вершины специального вида $[O(V)]$:
  - Изолированные вершины – вершины без единого ребра
  - Висячие вершины – вершины с одним ребром
  - Промежуточные вершины – вершины с двумя ребрами
  - Вершины, связанные с каждой вершиной графа
- Статистики на степенях вершин $[O(V)]$:
  - Максимальная степень вершины
  - Средняя степень вершины
  - Медианная степень вершины
  - Модальная степень вершины
  - Стандартное отклонение степеней вершин в графе
- Гистограмма степеней вершин графа $[O(V)]$ (здесь μ – средняя степень вершин в графе, σ – среднеквадратичное отклонение степеней вершин в графе): Доля вершин со степенью на интервалах: $(\mu - \sigma, \mu)$, $(\mu, \mu + \sigma)$, $(\mu - 2\sigma, \mu - \sigma)$, $(\mu + \sigma, \mu + 2\sigma)$, $(\mu - 3\sigma, \mu - 2\sigma)$, $(\mu + 2\sigma, \mu + 3\sigma)$
- Коэффициент кластеризации $[O(V^2)]$ [1]

Построения распределения графа

Пусть у нас есть граф G и соответствующая ему матрица смежности A.

Проведем над ним следующую операцию: n раз удалим либо добавим случайное число ребер в граф, получив тем самым набор из n графов, близких к исходному.



Для каждого полученного графа рассчитаем простые статистики и скрытые представления, пропуская его через энкодер. Дополнительно преобразуем простые статистики при помощи линейной свертки, приближая их к скрытым представлениям графа.

Стандартизируем скрытые представления. Имеем n векторов представления графа, полученных при помощи *графовых сверток* и n векторов представления графа, полученных при помощи *простых статистик*.

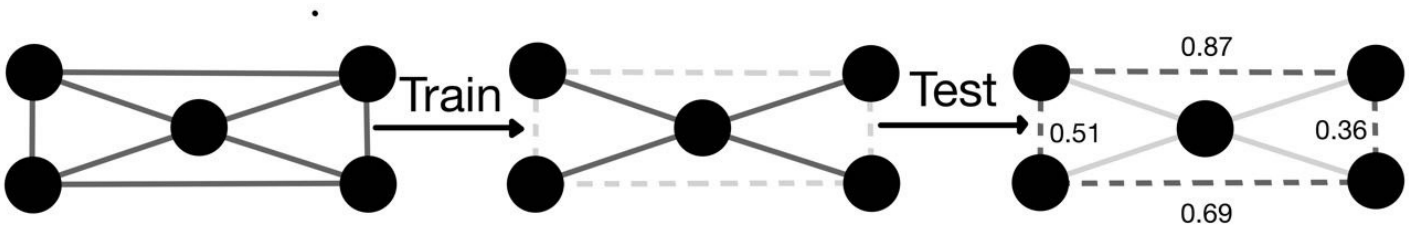
Постановка задачи реконструкции

Дано:

- Граф G с n вершинами и матрицей смежности A размера (n, n), где $A_{ij} = 1$, если ребро (i, j) существует в графе, и 0 в противном случае.
- Набор маркированных ребер $E = (i, j)$ существующих в графе.

Задача:

Предсказать вероятность существования ребер между всеми парами вершин в графе, включая немаркированные ребра. Однако особый интерес представляет именно предсказание наличия маркированных ребер.



Предложенный метод

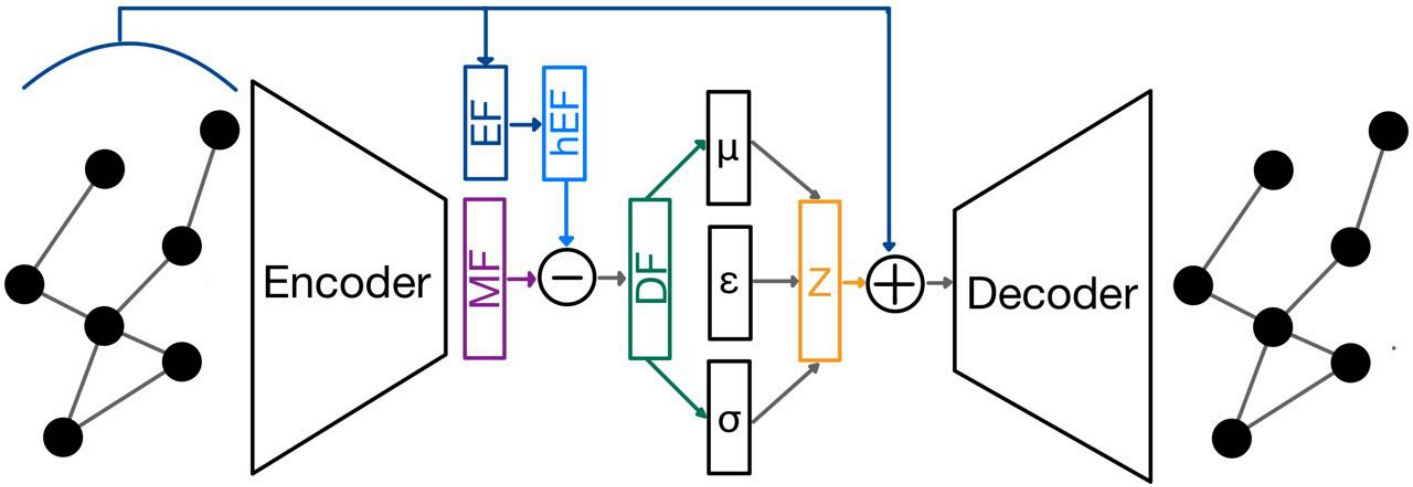


Рис.: Схема предложенного метода. Здесь MF – матрица смешанных статистик графа, EF – вектор простых статистик, hEF – скрытое представление вектора простых статистик, DF – матрица сложных статистик, – случайная величина $\in N(0,1)$, а Z – матрица из распределения $N(\mu, \sigma)$

Обучение происходит путем уменьшения следующей функции потерь:

$$\text{LOSS}_{method} = \text{BCELoss (реконструкция графа)}$$

$$+\text{KL-div (нормальность скрытого представления)}$$

$$+\text{MSELoss (приближение смешанных статистик простыми)}$$

Теоретическое обоснование

Hypothesis

Существует линейное отображение вектора смешанных статистик в вектор простых статистик.

Theorem (Бишук 2023)

Отображение из смешанных статистик в простые можно дополнить до невырожденного преобразования путем добавления строк, ортогональных исходным. Получившееся преобразование будет разбивать пространство статистик на простые и сложные.

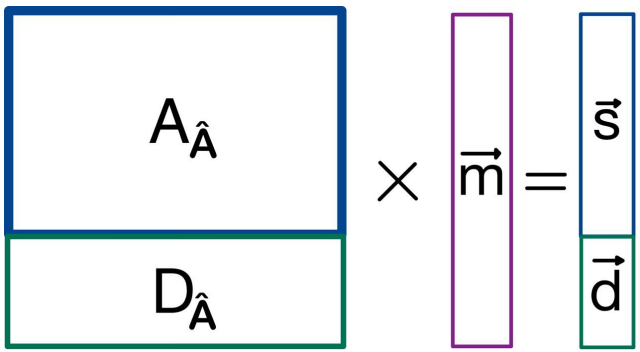


Рис.: Построенное линейное преобразование смешанных статистик

Идея доказательства

Remark

Независимость элементов вектора простых статистик мы можем гарантировать по построению.

Lemma

Пусть дан набор независимых, одинаково распределенных случайных величин $p_1, p_2, \dots p_n$. *Случайная величина* $\xi = a_1 p_1 + a_2 p_2 + \dots + a_n p_n$ *статистически зависима от каждой из случайных величин* p_i, *коэффициент перед которой* $a_i \neq 0$.

Lemma

Матрица $A_{\hat{A}}$ *имеет полный ранг.*

Добавим дополнительное условие на независимость компонент вектора $\vec{d}$ от компонент вектора $\vec{s}$ и между собой.

Для нормальных векторов условие независимости эквивалентна нескоррелированности, которая в свою очередь дает нам условие, что элементы вектора простых и сложных статистик независимы статистически, если независимы линейно строки матрицы преобразования.

Результаты

Датасеты:

- Cora (2708 статей и 10556 ссылки между ними)
- Citeseer (3327 статьи и 9228 ссылки между ними)

Метрики:

	Dataset	ROC-AUC	AP	MAE (GT statistics)
VAE	Cora	75.18 %	**75.81%**	0.066
Our method		**76.68 %**	75.18 %	**0.046 (-30%)**
VAE	Citeseer	**82.09 %**	**79.94 %**	0.072
Our method		76.28 %	76.44 %	**0.060 (-17%)**

Таблица: Результаты вычислительного эксперимента в задаче предсказания существования маркированных рёбер.

Предложенный метод восстановил граф хуже, чем стандартное VAE, однако точность простых статистик была увеличена.

Итоги

Итоги работы:

Предложен и теоретически обоснован метод к генерации графов, который позволяет получать графы с заранее заданными структурными свойствами. Методы был имплементирован и протестирован на данных из известных наборов. Были проведены исследования и показаны преимущества предложенного метода в задаче генерации графов с фиксированными свойствами.

Планы на будущие работы:

- Расширить множество простых статистик;
- Исследовать влияние фиксировании различных статистик графа на их разнообразие;
- Исследовать различные методы агрегации графов в векторе латентного пространства;
- Исследовать различные подходы к выделению сложных признаков и подмешивания простых;
- Рассмотреть другие функционалы качества.

Список литературы

[1] Jari Saramäki и др. «Generalizations of the clustering coefficient to weighted complex networks». В: *Physical Review E* 75.2 (2007), с. 027105.