Контролируемая генерация графов

Бишук Антон Юрьевич

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

Москва
2023 г

Цели и задачи

Цели

- Научиться генерировать графы с заранее заданными статистиками.

Задачи

- Предложить модификацию метода генерации графа, которая позволит генерировать графы обладающие заданными свойствами.
- Теоретически обосновать работу предложенного метода.
- Провести сравнение предложенной модификации с существующими методами.

Определения

Пространство признаков (или же статистик) графа на V вершинах - это евклидовое пространство, один радиус вектор в котором характеризует один единственный граф.

Простые признаки (или же простые статистики) в нашем методе - это признаки графа, которые могут быть интерпретированы (например, такие как степень вершин, ребер, и т.п.), а также вычислены алгоритмами с асимптотикой не более квадратичной по времени.

Сложными признаками (или статистиками) мы назовем ортогональное дополнение подпространства простых признаков.

Смешанными признаками назовем любую линейную комбинацию сложных и простых признаков, которая задает граф.

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: $(\mu - \sigma, \mu)$, $(\mu, \mu + \sigma)$, $(\mu - 2\sigma, \mu - \sigma)$, $(\mu + \sigma, \mu + 2\sigma)$, $(\mu - 3\sigma, \mu - 2\sigma)$, $(\mu + 2\sigma, \mu + 3\sigma)$
- Оценка размер наибольшей клики в графе $[O(Vd^2)$, d – максимальная степень вершины] [1]
- Коэффициент кластеризации $[O(V^2)]$ [2]

Теоретические вводные

Стандартный метод формирования скрытого представления в VAE это трюк репараметризации.

Трюк репараметрицащии позволяет гарантировать, что полученное латентное представление графа (смешаные статистики) будут соответствовать многомерному нормальному распределению.

Вектор простых признаков графа будем считать также нормальным вектором.

Теоремы

Theorem (Бишук 2023)

Пусть задано вероятностное пространство, состоящее из статистик графа. Рассмотрим нормальный вектор, принадлежащий подпространству данного вероятностного пространства - будем называть его вектором смешанных статистик. Если заданы простые статистики графа, которые являются компонентами вектора смешанных статистик, то возможно этот вектор смешанных статистик разложить в сумму вектора сложных признаков и вектора простых признаков.

Lemma (Бишук 2023)

Для нахождения сложных статистик в смешанных признаках достаточно найти такой вектор, который получается из смешанных линейным преобразованием и не является скоррелированым с простыми статистиками.

Предложенный метод

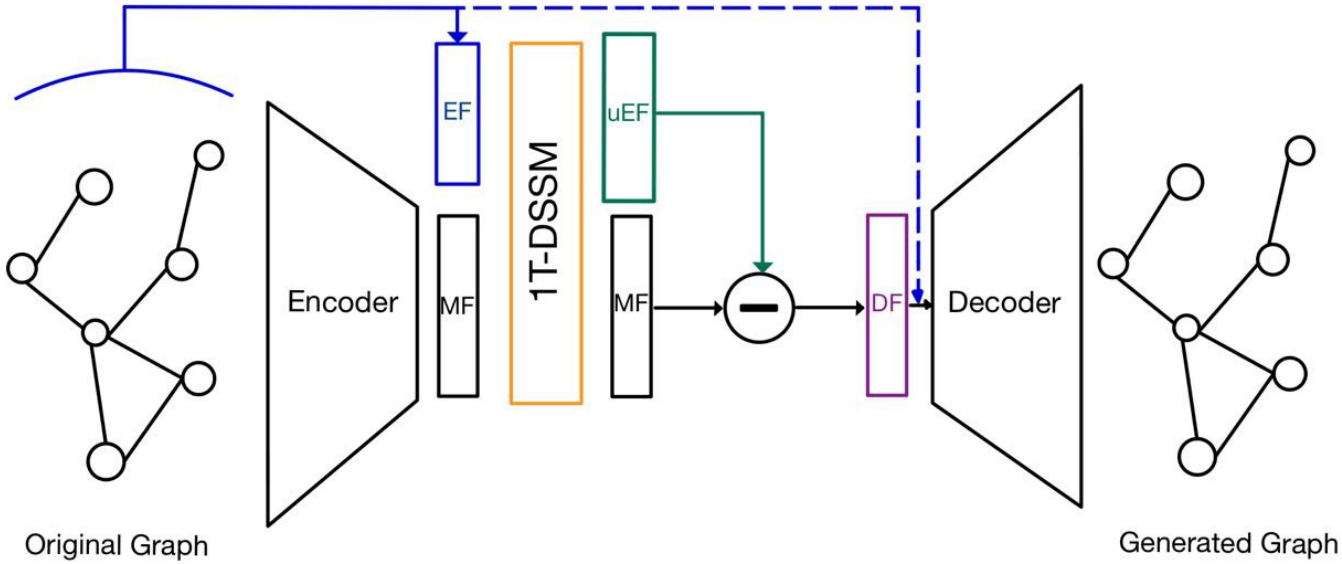

Рис.: Схема модели с генерацией на основе сложных и простых признаков.

Обучение происходит путем уменьшения следующей функции потерь:

$$\text{LOSS}_{method} = \text{BCELoss}(\textit{reconstruction}) + \text{KL-div} + \text{MSELoss}(\textit{similarity})$$

Результаты

Датасеты:

- Cite (2708 статей, разбитых на 7 категорий)
- Citeseer (3327 статьи и 4732 ссылки между ними)

	Dataset	ROC-AUC	AP	MAE (global statistics)
VAE	Cite	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 по классификации наличия ребер.

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

Результаты

Итоги:

Предложен и теоретически обоснован метод, который позволяет генерировать графы с заданными статистиками. Проведены исследования и показаны преимущества метода на стандартных датасетах.

Планы:

- Дополнить множество простых статистик
- Исследовать различные методы агрегации графов в векторе латентного пространства

Список литературы

[1] Bharath Pattabiraman и др. «Fast algorithms for the maximum clique problem on massive sparse graphs». В: *Algorithms and Models for the Web Graph: 10th International Workshop, WAW 2013, Cambridge, MA, USA, December 14-15, 2013, Proceedings 10*. Springer. 2013, с. 156—169.

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