Предобработка графов в задачах кластеризации с высокой плотностью связей

студент: Фаляхов И.Р.

научный руководитель: д.т.н. Матвеев И.А.

Москва 2020

Процессы кластеризации графов с высокой плотностью связей и сложной природой данных часто непредсказуемы: качество результатов узнается пост-фактум.

Для улучшения качества кластеризации используются различные алгоритмы предобработки, но не всегда понятно в каких случаях они улучшат результат.

При решении задачи выявления рабочих групп сотрудников компании столкнулся с подобной проблемой, что и послужило мотивацией к анализу.

Постановка задачи

Процесс кластеризации состоит из трех основных пунктов:

1. Метрика для алгоритма кластеризации;
2. Алгоритм кластеризации;
3. **Методы предобработки;**

Задача: провести анализ ряда методов предобработки и их различных комбинаций на разных данных. Понять, в каких случаях методы предобработки помогают улучшить кластеризацию.

Определения

- Степень вершины (d) – количество соседей вершины;
- Ассортативность графа: $assort_{xy} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \overline{x})^2 \sum_{i=1}^{m}(y_i - \overline{y})^2}}$
  где $x^m = \{x_1, x_2, \dots, x_m\}$,
  $y^m = \{y_1, y_2, \dots, y_m\}$ – степени вершин,
  имеющих общее ребро (между x_i и y_i есть общее ребро),
  $\bar{x}$ – выборочное среднее по x^m;
- Степень связности двух узлов: $S(v, w) = \frac{|N(v) \cap N(w)|}{|N(v) \cup N(w)|}$
  где $N(v)$ - множество соседей
  узла v;

-Треугольник (T) – случай, когда 3 узла попарно соединены рёбрами;

-Граф: $G = (V, E, W)$;

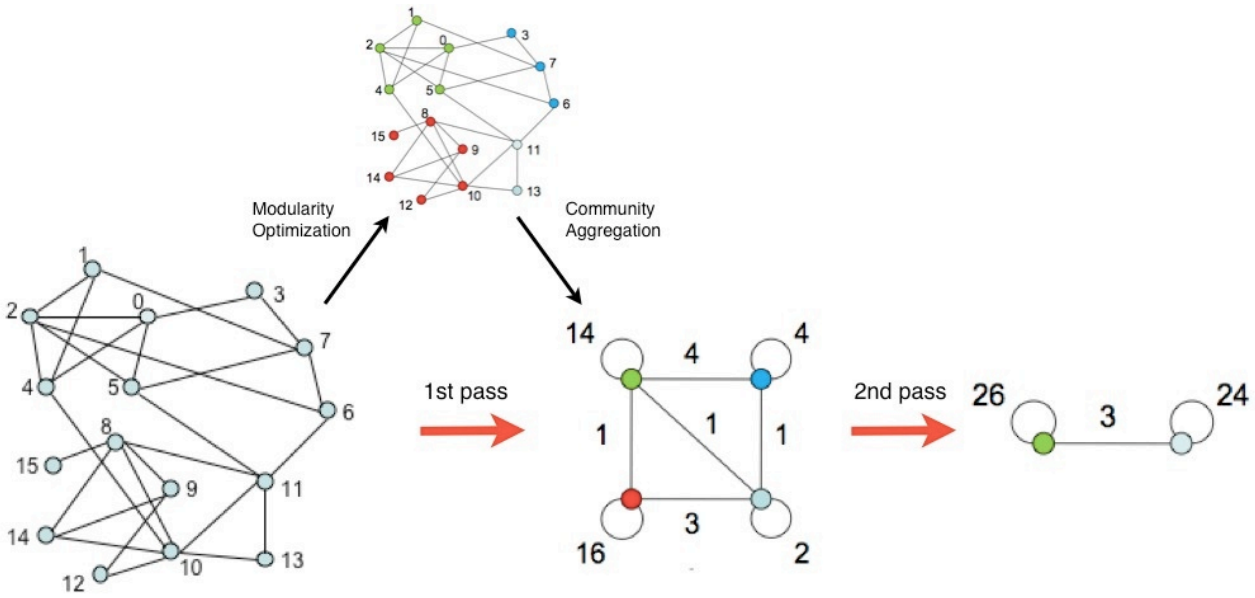
Разбиение на кластера: $C = \{c_1, \dots, c_s\},\ c_i \subseteq V,\ \sqcup\, c_i = V$;

Метрика алгоритма и алгоритм кластеризации

1. Метрика – модулярность. $$Q = \frac{1}{2m}\sum_{vw}\left[A_{vw} - \frac{k_v k_w}{2m}\right]\delta(c_v, c_w)$$

   v, w – узлы, A_{vw} – вес ребра инцинд. и v и w, $m = |E|$, k_v – сумма рёбер, инцинд. v, c_v – кластер узла v.

2. Лувенский алгоритм.
   Сложность: $O(n \log n)$



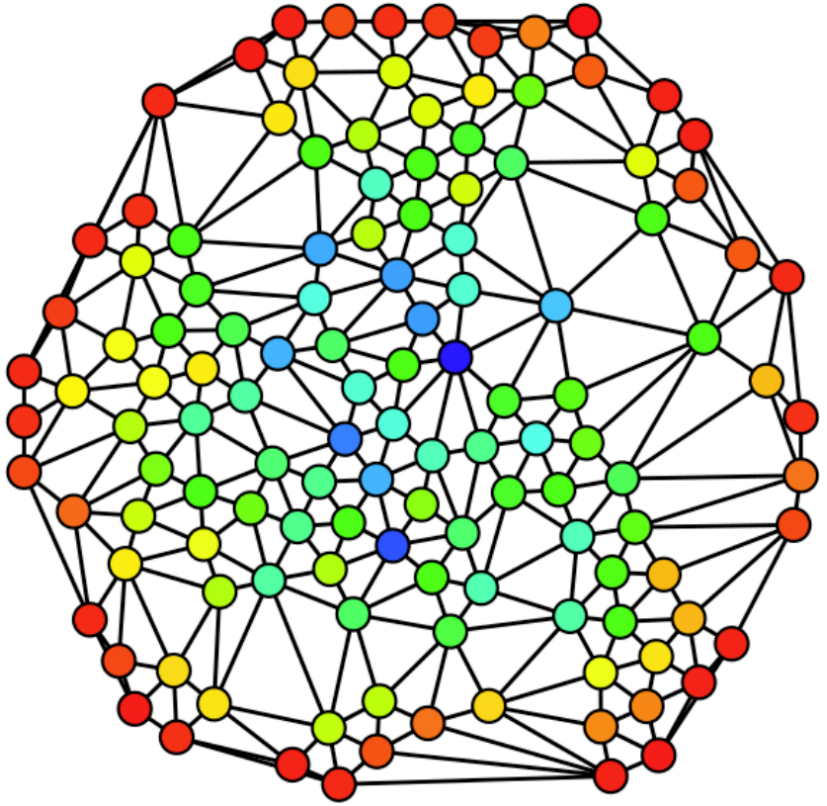
Методы предобработки

1. Удаление вершин с высоким значением метрики центральности (Betwenness Centrality).

$$BC(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

s, v, t – узлы; σ_{st} – количество кратчайших путей из s в t;
$\sigma_{st}(v)$ – количество кратчайших путей из s в t, проходящих через v.

Сложность: $O(nm + n^2 \log n)$



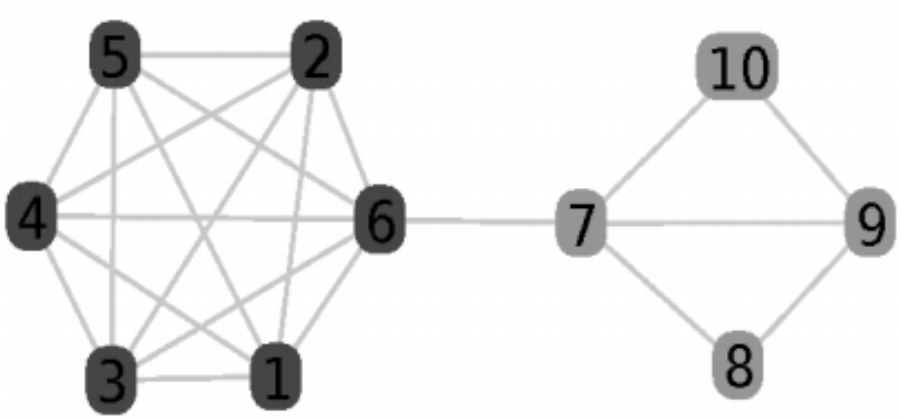
Методы предобработки

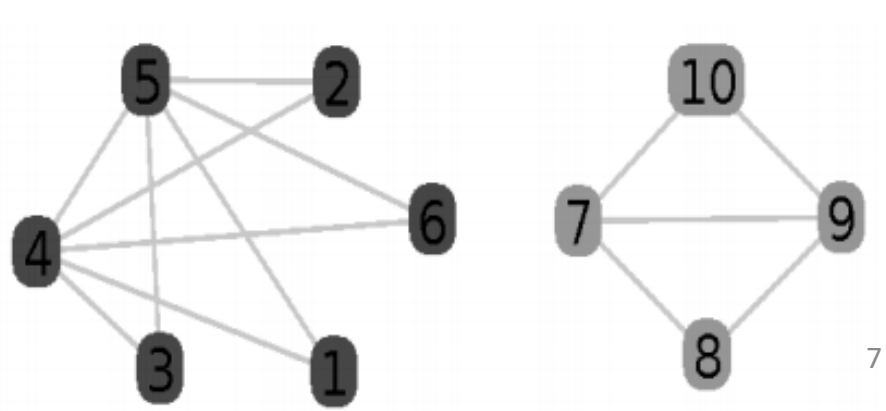
2. L-SPAR с параметром разреженности $\omega \in [0,1]$.

   Для каждой вершины графа v выполняем:

   - Вычисляем степень связанности соседей с вершиной v и сортируем по степени связанности в порядке невозрастания;
   - Оставляем рёбра с первыми $|N(v)|^{\omega}$ соседями по степени связанности, где $N(v)$ – множество соседей вершины v;

   Сложность: $O(nd_{mean})$





Методы предобработки

3. Усредненный L-SPAR.

Для каждой вершины графа v выполняем:

- Вычисляем степень связанности соседей с вершиной v;
- Оставляем рёбра с теми соседями, у которых степень связанности больше или равна их усредненной степени связанности;

Сложность: $O(nd_{mean})$

Метрики качества

1. Чистота (Purity).
2. Нормализованная взаимная информация (NMI).
3. Исправленный индекс Рэнда (ARI).
4. Стабильность разбиений

Метрики качества

4. Стабильность разбиений:
   1) Строится 2 разбиения;
   2) Вычисляется мера Жаккара для каждого кластера этих разбиений и усредняется;
   3) Предыдущие 2 шага повторяются 10 раз;
   4) Получившиеся усредненные меры Жаккара усредняются;

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

Данные

1. Данные встреч сотрудников компании Qiwi. Узлы – сотрудники, рёбра – встречи друг с другом.
   No. nodes = 1440, No. edges = 16972, d_{max}= 147, d_{mean} = 24,
   T_{max}= 1689, T_{mean}= 153, $assort$ = 0.19;
2. Данные о ссылках внутри научных стаей на другие статьи Cora. Узлы – статьи, рёбра – ссылки.
   No. nodes = 2708, No. edges = 5278, d_{max}= 168, d_{mean}= 4,
   T_{max}= 200, T_{mean}= 75, $assort$ = -0.06;
3. Данные о ссылках внутри научных статей на другие статьи Citeseer. Узлы – статьи, рёбра – ссылки.
   No. nodes = 3264, No. edges = 4536, d_{max}= 99, d_{mean}= 3,
   T_{max}= 85, T_{mean}= 2, $assort$ = 0.05;
4. Данные эго-сетей сотрудничества исследователей COLLAB. Узлы – исследователи, рёбра – участие в исследованиях.
   No. nodes = 372.5 тыс., No. edges = 12.3 млн., d_{max}= 491, d_{mean}= 66,
   T_{max}= 39.5 тыс., T_{mean}= 5016, $assort$ = 0.92;

Вычислительные эксперименты

Были проведены эксперименты с применением описанных методов предобработки по отдельности и в различных комбинациях:

- удаление центральных вершин + L-SPAR с параметром разреженности;
- удаление центральных вершин + усреднённый L-SPAR;
- L-SPAR с параметром разреженности + удаление центральных вершин;
- Усреднённый L-SPAR + удаление центральных вершин;

Так же был проведены эксперименты с отсутствием методов предобработки.

Результаты

Таблица 1: Метрики качества на наборе данных Qiwi

Algo	PU	NMI	ARI	STAB
$LOUV$	0.64	0.38	0.25	0.6
BD	0.69	0.41	0.29	0.69
LSP	0.75	0.63	0.5	0.75
$AverLSP$	0.74	0.63	0.51	0.76
$BD + LSP$	**0.88**	**0.79**	**0.65**	**0.89**
$BD + AverLSP$	0.87	**0.79**	**0.65**	**0.89**
$LSP + BD$	0.81	0.76	0.59	**0.89**
$AverLSP + BD$	0.81	0.76	0.59	**0.89**

Таблица 2: Метрики качества на наборе данных Cora

Algo	PU	NMI	ARI	STAB
$LOUV$	0.69	0.4	0.19	0.65
BD	0.73	0.41	0.21	0.7
LSP	0.74	0.4	0.2	0.7
$AverLSP$	0.74	0.39	0.2	0.7
$BD + LSP$	**0.78**	0.46	**0.31**	**0.91**
$BD + AverLSP$	**0.78**	**0.47**	0.3	**0.91**
$LSP + BD$	0.75	0.42	0.24	**0.91**
$AverLSP + BD$	0.75	0.41	0.24	**0.9**

Результаты

Таблица 3: Метрики качества на наборе данных Citeseer

Algo	PU	NMI	ARI	STAB
$LOUV$	**0.74**	0.37	**0.1**	**0.95**
BD	0.73	0.39	0.08	**0.95**
LSP	**0.74**	0.38	0.07	**0.95**
$AverLSP$	**0.74**	0.37	0.07	**0.95**
$BD + LSP$	**0.74**	**0.42**	**0.1**	**0.95**
$BD + AverLSP$	**0.74**	0.41	0.09	**0.95**
$LSP + BD$	**0.74**	0.41	0.08	**0.95**
$AverLSP + BD$	0.73	0.4	0.08	**0.95**

Таблица 4: Метрики качества на наборе данных COLLAB

Algo	PU	NMI	ARI	STAB
$LOUV$	**1**	**1**	**1**	**1**
BD	**1**	**1**	**1**	**1**
LSP	0.91	0.87	0.75	**1**
$AverLSP$	0.91	0.87	0.75	**1**
$BD + LSP$	0.85	0.74	0.64	**1**
$BD + AverLSP$	0.84	0.74	0.59	**1**
$LSP + BD$	0.83	0.68	0.5	**1**
$AverLSP + BD$	0.83	0.68	0.5	**1**

Заключение

- Усредненный L-SPAR имеет сравнимые значения метрик с L-SPAR с подобранным параметром разреженности;
- На наборах данных с малым значением параметра ассортативности и большим среднем количеством треугольников лучше всего себя показывают алгоритмы с последовательным применением удаления центральных вершин и алгоритмом L-SPAR, с параметром разреженности и усреднённый;
- В наборах данных с таким же малым значением ассортативности, но при этом малым средним количеством треугольников, нет чёткой картины. В таких случаях отсутствие предобработки может показать себя лучше.
- В наборах данных с значением ассортативности близким к единице методы предобработки могут быть излишни: алгоритм без применения методов предобработки показал себя максимально эффективным.

СПАСИБО ЗА ВНИМАНИЕ