ОБ ОДНОЙ ЗАДАЧЕ КЛАСТЕРИЗАЦИИ НА ГРАФЕ

Ильев В.П., Ильева С.Д.

Омский государственный университет

Задачи кластеризации

В *задаче кластеризации* требуется разбить заданное множество объектов на несколько подмножеств (*кластеров*) на основе сходства объектов друг с другом.

Мера сходства определяется по-разному в разных задачах.

В теории распознавания образов и в машинном обучении задачи кластеризации относят к разделу *обучения без учителя*.

Наряду с этим рассматриваются также *задачи кластеризации с частичным обучением*, в которых часть объектов изначально распределена по кластерам.

Задача кластеризации на графе (задача аппроксимации графа)

Одна из наглядных формализаций задачи кластеризации взаимосвязанных объектов — *задача кластеризации на графе* (или *задача аппроксимации графа*):
– вершины графа соответствуют объектам;
– пары вершин, соответствующие похожим объектам, связаны посредством ребер.

Требуется разбить множество вершин на попарно непересекающиеся группы (кластеры) с учетом реберной структуры графа.

Цель — минимизация числа связей между кластерами и числа недостающих связей внутри кластеров.

Определения и обозначения

Обыкновенный граф называется *кластерным графом*, если каждая его компонента связности является полным графом.

- $\mathcal{M}(V)$ – множество всех кластерных графов на множестве вершин V,
- $\mathcal{M}_k(V)$ – множество всех кластерных графов на множестве V, имеющих ровно k непустых компонент связности, $2 \leqslant k \leqslant |V|$,
- $\mathcal{M}_{1,k}(V)$ – множество всех кластерных графов на множестве V, имеющих не более k компонент связности, $2 \leqslant k \leqslant |V|$.

Расстояние между графами $G_1 = (V, E_1)$, $G_2 = (V, E_2)$:

$$\rho(G_1, G_2) = |E_1 \triangle E_2| = |E_1 \setminus E_2| + |E_2 \setminus E_1|.$$

Постановки задач

Задача A. Дан граф $G = (V, E)$. Найти такой граф $M^* \in \mathcal{M}(V)$, что

$$\rho(G, M^*) = \min_{M \in \mathcal{M}(V)} \rho(G, M)$$

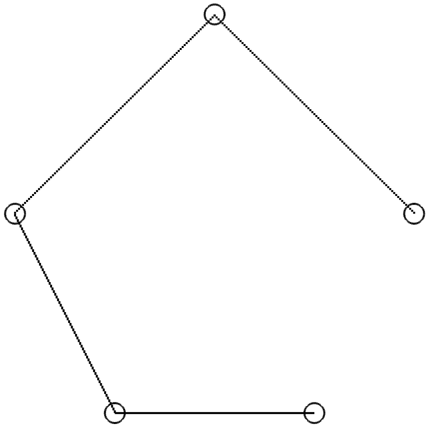
Задача A$_\mathbf{k}$. Дан граф $G = (V, E)$ и целое число k, $2 \leqslant k \leqslant |V|$. Найти такой граф $M^* \in \mathcal{M}_k(V)$, что

$$\rho(G, M^*) = \min_{M \in \mathcal{M}_k(V)} \rho(G, M)$$

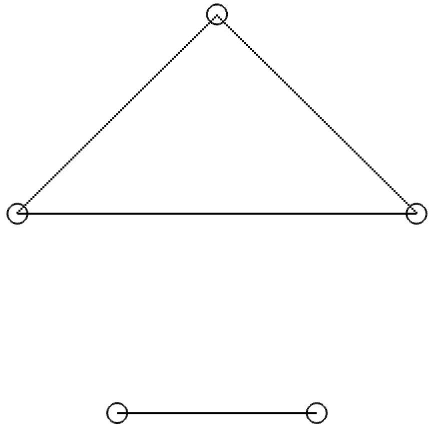
Задача A$_\mathbf{1,k}$. Дан граф $G = (V, E)$ и целое число k, $2 \leqslant k \leqslant |V|$. Найти такой граф $M^* \in \mathcal{M}_{1,k}(V)$, что

$$\rho(G, M^*) = \min_{M \in \mathcal{M}_{1,k}(V)} \rho(G, M)$$

Пример. Задачи $\mathbf{A}$, $\mathbf{A}_2$, $\mathbf{A}_{1,2}$



G



M^*

Другие названия

- **Задача аппроксимации графов**

[Zahn C.T. *Approximating symmetric relations by equivalence relations* // J. of the Society for Industrial and Applied Mathematics. 1964. V. 12. P. 840–847.]

- **Correlation Clustering**

[Bansal N., Blum A., Chawla S. *Correlation Clustering* // Machine Learning. 2004. V. 56. P. 89–113.]

- **Cluster Editing**

[Shamir R., Sharan R., Tsur D., *Cluster graph modification problems* // Discrete Appl. Math. 2004. V. 144. P. 173–182.]

Точные алгоритмы

- Задача **A** полиномиально разрешима для графов специального вида

  [Zahn, 1964]

- Задача **A** для графов без треугольников сведена к задаче о наибольшем паросочетании

  [Фридман, 1971]

- Точный алгоритм решения задачи **A** для графов, не содержащих четырехвершинных подграфов ровно с пятью ребрами

  [Вейнер, 1971]

Вычислительная сложность

- Задача $\mathbf{A}$ NP-трудна [Křivánek-Morávek, 1986; Bansal-Blum-Chawla, 2004; Shamir-Sharan-Tsur, 2004]

- Задача $\mathbf{A_k}$ NP-трудна при любом фиксированном $k \geqslant 2$ [Shamir-Sharan-Tsur, 2004; Giotis-Guruswami, 2006]

- Задачи $\mathbf{A_{1,2}}$ и $\mathbf{A_2}$ NP-трудны на кубических графах.

Отсюда выводится, что все варианты задач являются NP-трудными, включая $\mathbf{A_{1,k}}$ при любом фиксированном $k \geqslant 2$ [Агеев-Ильев-Кононов-Талевнин, 2006]

Алгоритмы приближенного решения

- 3-приближенный алгоритм для задачи $\mathbf{A_{1,2}}$ [Bansal-Blum-Chawla, 2004]
- 4-приближенный алгоритм для задачи $\mathbf{A}$ [Charikar-Guruswami-Wirth, 2005]
- Рандомизированная ППС для задачи $\mathbf{A_{1,k}}$ при любом фиксированном $k \geqslant 2$ [Giotis-Guruswami, 2006]
- 2-приближенный алгоритм для задачи $\mathbf{A_{1,2}}$ [Coleman-Saunderson-Wirth, 2008]
- 2,5-приближенный алгоритм для задачи $\mathbf{A}$ [Ailon-Charikar-Newman, 2008]
- 3-приближенный алгоритм для задачи $\mathbf{A_2}$ [Ильев-Ильева-Навроцкая, 2011]

Задача кластеризации с частичным обучением

Задача $\mathbf{A}_\mathbf{k}^+$. Дан граф $G = (V, E)$ и целое число k, $2 \leqslant k \leqslant |V|$. Выделено множество попарно различных вершин $X = \{x_1, \ldots, x_k\} \subseteq V$. Найти такой граф $M^* \in \mathcal{M}_k(V)$, что

$$\rho(G, M^*) = \min_{M \in \mathcal{M}_k(V)} \rho(G, M),$$

где минимум берется по всем кластерным графам $M \in \mathcal{M}_k(V)$, в которых никакие две вершины множества $X = \{x_1, \ldots, x_k\}$ не принадлежат одному и тому же кластеру (т. е. множеству вершин одной компоненты связности графа M).

- Задача $\mathbf{A}_\mathbf{k}^+$ NP-трудна при любом фиксированном $k \geqslant 2$.

Приближенный алгоритм для $\mathbf{A}_2^+$

Задача $\mathbf{A}_2^+$. Дан граф $G=(V,E)$ и множество $X=\{x_1,x_2\}$, где $x_1,x_2\in V$, $x_1\neq x_2$. Найти такой граф $M^*\in\mathcal{M}_2(V)$, что

$$\rho(G,M^*)=\min_{M\in\mathcal{M}_2(V)}\rho(G,M),$$

причем минимум берется по всем кластерным графам $M\in\mathcal{M}_2(V)$, таким, что $x_1\in V_1$, $x_2\in V_2$, где V_1,V_2 – кластеры (множества вершин компонент связности графа M).

Процедура построения кластерного графа.
Пусть v – произвольная вершина графа G, $N(v)$ – множество вершин графа G, смежных с v. Построим кластерный граф $M_v=M(V_1,V_2)\in\mathcal{M}_2(V)$ по следующим правилам.

Правила построения графа $M_v = M(V_1, V_2) \in \mathcal{M}_2(V)$

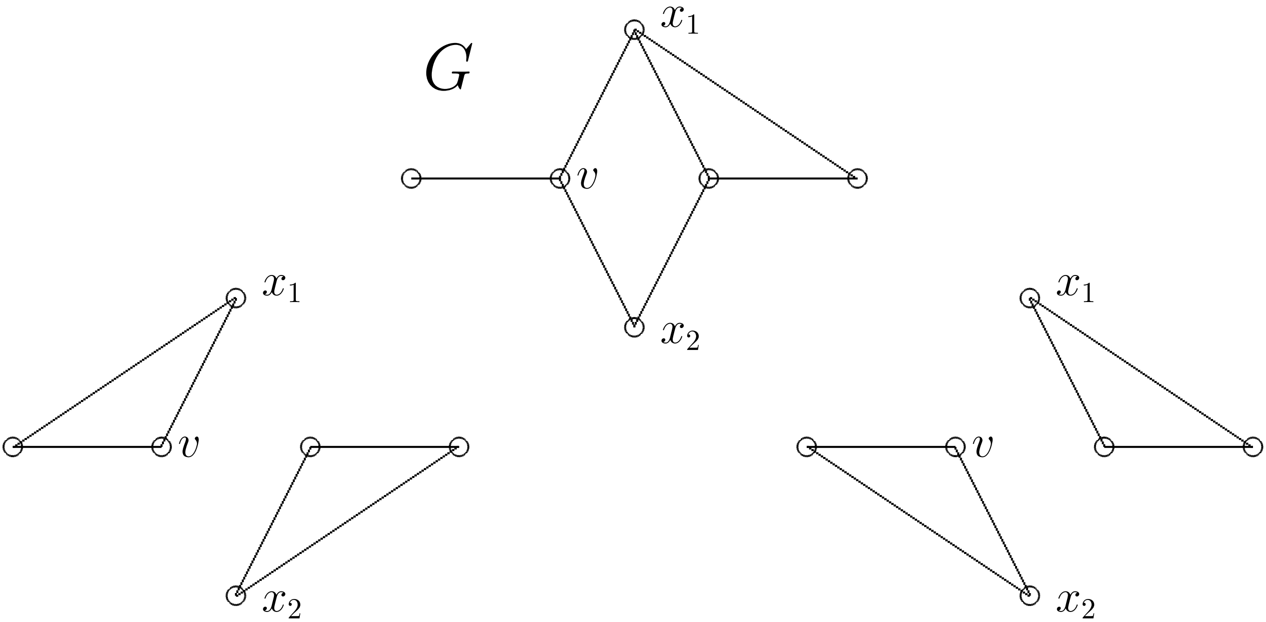
(а) Если в графе G вершина v смежна ровно с одной из вершин x_1, x_2, то $V_1 = \{v\} \cup N(v)$, $V_2 = V \setminus V_1$.

(б) Если в графе G вершина v смежна с обеими вершинами x_1, x_2, то

$M' = M(V_1', V_2')$, где $V_1' = (\{v\} \cup N(v)) \setminus \{x_2\}$, $V_2' = V \setminus V_1'$,
$M'' = M(V_1'', V_2'')$, где $V_1'' = (\{v\} \cup N(v)) \setminus \{x_1\}$, $V_2'' = V \setminus V_1''$.

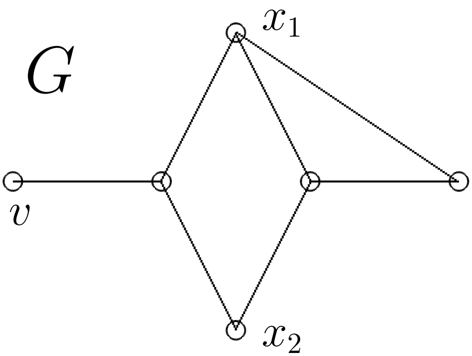
Если при этом $\rho(G, M') \leqslant \rho(G, M'')$, то $M_v = M'$, в противном случае $M_v = M''$.

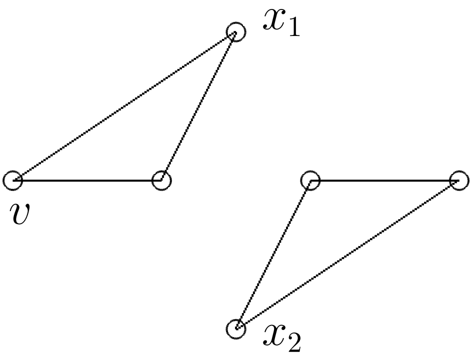
Пример. Правило (б)

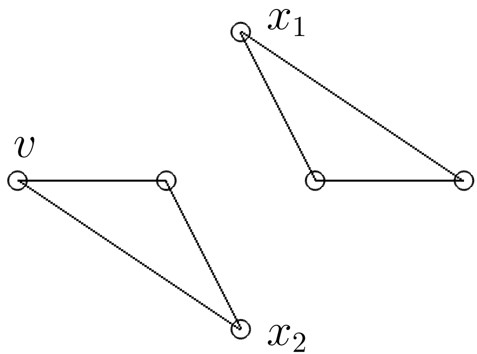


$\rho(G, M') = 5$ $\rho(G, M'') = 3$

Пример. Правило (в)







$\rho(G, M') = 5$ $\qquad$ $\rho(G, M'') = 3$

Правила построения графа $M_v = M(V_1, V_2) \in \mathcal{M}_2(V)$

(в) Если в графе G вершина v несмежна с обеими вершинами x_1, x_2, то

$M' = M(V_1', V_2')$, где $V_1' = \{v\} \cup N(v) \cup \{x_1\}$, $V_2' = V \setminus V_1'$,
$M'' = M(V_1'', V_2'')$, где $V_1'' = \{v\} \cup N(v) \cup \{x_2\}$, $V_2'' = V \setminus V_1''$.

Если при этом $\rho(G, M') \leqslant \rho(G, M'')$, то $M_v = M'$, в противном случае $M_v = M''$.

(г) Если вершина v совпадает с одной из вершин x_1, x_2, то $V_1 = (\{v\} \cup N(v)) \setminus \{x\}$, $V_2 = V \setminus V_1$,
где $x = x_1$, если $v = x_2$, и $x = x_2$, если $v = x_1$.

Приближенный алгоритм

Рассмотрим следующий алгоритм приближенного решения задачи $\mathbf{A_2^+}$.

Алгоритм ALG_2^+

Шаг 1. Для каждой вершины $v \in V$ построить кластерный граф $M_v \in \mathcal{M}_2(V)$ по правилам (а)–(г).

Шаг 2. Среди всех графов M_v выбрать такой граф M_2^+, что

$$\rho(G, M_2^+) = \min_{v \in V} \rho(G, M_v).$$

Конец алгоритма.

Гарантированная оценка точности алгоритма ALG_2^+

Теорема 2. При $n \geqslant 3$ для любого n-вершинного графа $G = (V, E)$ и любого множества $X = \{x_1, x_2\} \subset V$ $(x_1 \neq x_2)$ алгоритм ALG_2^+ находит такой кластерный граф $M_2^+ \in \mathcal{M}_2(V)$, что

$$\rho(G, M_2^+) \leqslant \left(3 - \frac{6}{n}\right) \rho(G, M^*),$$

где $M^* \in \mathcal{M}_2(V)$ – оптимальное решение задачи $\mathbf{A_2^+}$ на графе G.