ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Июль–сентябрь 2020. Т. 27, № 3. С. 88–108

УДК 519.8 DOI 10.33048/daio.2020.27.680

2-ПРИБЛИЖЁННЫЕ АЛГОРИТМЫ ДЛЯ ДВУХ ЗАДАЧ КЛАСТЕРИЗАЦИИ НА ГРАФАХ

В. П. Ильев[1,2,a], *С. Д. Ильева*[1], *А. В. Моршинин*[2,b]

[1] Омский гос. университет им. Ф. М. Достоевского,
пр. Мира, 55а, 644077 Омск, Россия

[2] Омский филиал Института математики им. С. Л. Соболева,
ул. Певцова, 13, 644043 Омск, Россия

E-mail: [a] iljev@mail.ru, [b] morshinin.alexander@gmail.com

Аннотация. Изучаются вариант задачи 2-кластеризации на графе и соответствующая задача с частичным обучением. В этих задачах для данного графа требуется найти ближайший 2-кластерный граф, т. е. граф на том же множестве вершин, имеющий ровно 2 непустые компоненты связности, каждая из которых является полным графом. Расстояние между графами равно числу их несовпадающих рёбер. Обе рассматриваемые задачи NP-трудны. В 2008 г. Коулман, Саундерсон и Вирт предложили полиномиальный 2-приближённый алгоритм для аналогичной задачи, в которой число кластеров не превосходит 2. К сожалению, метод доказательства гарантированной оценки точности алгоритма Коулмана, Саундерсона и Вирта неприменим для задачи 2-кластеризации на графе, в которой число кластеров в точности равно 2. Мы предлагаем полиномиальный 2-приближённый алгоритм для задачи 2-кластеризации на произвольном графе. В отличие от доказательства Коулмана, Саундерсона и Вирта, наше доказательство гарантированной оценки точности этого алгоритма не использует техники переключений. Кроме того, предложен аналогичный 2-приближённый алгоритм для соответствующей задачи с частичным обучением. Библиогр. 9.

Ключевые слова: граф, кластеризация, NP-трудная задача, приближённый алгоритм, гарантированная оценка точности.

Введение

В задачах кластеризации требуется разбить данное множество объектов на несколько подмножеств (кластеров) только на основе сходства

объектов друг с другом. Изучаются также варианты задач кластеризации с частичным обучением, в которых фиксированное подмножество объектов изначально распределено по кластерам. В задачах кластеризации на графах отношение сходства объектов задано посредством неориентированного графа, вершины которого взаимно однозначно соответствуют объектам, а рёбра соединяют похожие объекты.

Мы рассматриваем вариант задачи кластеризации на графе, эквивалентный хорошо известной задаче 2-Correlation Clustering [1, 2], в которой для данного графа требуется найти ближайший 2-кластерный граф, т. е. граф на том же множестве вершин, имеющий ровно 2 непустые компоненты связности, каждая из которых является полным графом. Расстояние между графами равно числу их несовпадающих рёбер.

Будем рассматривать только *обыкновенные* графы, т. е. графы без петель и кратных рёбер. Обыкновенный граф называется *кластерным графом*, если каждая его компонента связности является полным графом [3]. Обозначим через $\mathcal{M}_k(V)$ множество всех кластерных графов на множестве вершин V, имеющих ровно k непустых компонент связности, $\mathcal{M}_{\leqslant k}(V)$ — множество всех кластерных графов на множестве V, имеющих не более k компонент связности, $2 \leqslant k \leqslant |V|$.

Если $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — обыкновенные графы с нумерованными вершинами на одном и том же множестве вершин V, то *расстояние* $\rho(G_1, G_2)$ между ними определяется так:

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

т. е. $\rho(G_1, G_2)$ — число несовпадающих рёбер в графах G_1 и G_2.

Следующие варианты задачи кластеризации на графе изучались в литературе под различными наименованиями: задача аппроксимации графа [4–6], k-Correlation Clustering [1, 2], MinDisAgree[k] [7], k-Cluster Editing [3], и т. д.

Задача $\mathbf{GC}_k$. *Для произвольного графа $G = (V, E)$ и натурального числа k, $2 \leqslant k \leqslant |V|$, найти граф $M^* \in \mathcal{M}_k(V)$ такой, что*

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

Задача $\mathbf{GC}_{\leqslant k}$. *Для произвольного графа $G = (V, E)$ и натурального числа k, $2 \leqslant k \leqslant |V|$, найти граф $M^* \in \mathcal{M}_{\leqslant k}(V)$ такой, что*

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

В 2004 г. Шамир, Шаран и Цур [3] доказали, что задача GC_k NP-трудна для любого фиксированного $k \geqslant 2$. В 2006 г. Гиотис и Гурусвами [7] опубликовали более простое доказательство того же результата. В том же году Агеев, Ильев, Кононов и Талевнин [6] независимо доказали, что

задачи GC_2 и $\mathrm{GC}_{\leqslant 2}$ NP-трудны уже на 3-регулярных (кубических) графах, откуда вывели, что обе задачи GC_k и $\mathrm{GC}_{\leqslant k}$ на произвольных графах NP-трудны для любого фиксированного $k \geqslant 2$.

Методы решения задач кластеризации составляют важный раздел теории распознавания образов и машинного обучения. 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 (как правило, небольшая) изначально распределена по кластерам [8, 9].

Рассмотрим следующую формализацию задачи кластеризации с частичным обучением.

Задача SGC_k. *Даны обыкновенный граф $G = (V, E)$ и целое число k, $2 \leqslant k \leqslant |V|$. Выделено множество попарно различных вершин $Z = \{z_1, \ldots, z_k\} \subset V$. Требуется найти граф $M^* \in \mathcal{M}_k(V)$ такой, что*

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

причём минимум берётся по всем кластерным графам $M = (V, E_M) \in \mathcal{M}_k(V)$, в которых $z_i z_j \notin E_M$ для любых $i, j \in \{1, \ldots, k\}$; другими словами, никакие две вершины множества $Z = \{z_1, \ldots, z_k\}$ не принадлежат одной и той же компоненте связности (т. е. одному кластеру) графа M.

Несложно свести по Тьюрингу задачу GC_k к SGC_k и тем самым показать, что задача SGC_k тоже NP-трудна.

В 2004 г. Бансал, Блюм и Чаула [1] предложили простой полиномиальный 3-приближённый алгоритм для задачи $\mathrm{GC}_{\leqslant 2}$. В 2006 г. Гиотис и Гурусвами [7] разработали рандомизированную полиномиальную приближённую схему (PTAS) для задачи MinDisAgree[k], эквивалентной задаче $\mathrm{GC}_{\leqslant k}$ (для любого фиксированного $k \geqslant 2$). В 2008 г. Коулман, Саундерсон и Вирт [2] указали, что сложность PTAS из [7] лишает её перспективы практического использования, и предложили 2-приближённый алгоритм для задачи $\mathrm{GC}_{\leqslant 2}$, применив процедуру локального поиска к каждому допустимому решению, полученному с помощью 3-приближённого алгоритма из [1].

К сожалению, метод доказательства гарантированной оценки точности алгоритма Коулмана, Саундерсона и Вирта неприменим для задачи GC_2. Коулман, Саундерсон и Вирт используют технику переключений, которая позволяет свести кластеризацию вершин любого графа к эквивалентной задаче, оптимальное решение которой является полным графом, т. е. кластерным графом, состоящим из единственного кластера. В задаче GC_2 любое оптимальное решение должно состоять из двух непустых кластеров, поэтому нужны другой приближённый алгоритм и другой способ доказательства его гарантированной оценки точности.

В разд. 1 мы предлагаем модифицированный 2-приближённый алгоритм для задачи GC_2. В отличие от доказательства Коулмана, Саундерсона и Вирта, наше доказательство гарантированной оценки точности этого алгоритма не использует техники переключений. В разд. 2 рассматривается вариант задачи кластеризации на графе с частичным обучением. Для случая $k = 2$ предложен полиномиальный 2-приближённый алгоритм.

1. Задача $\mathbf{GC}_2$

1.1. Постановка задачи и вспомогательные утверждения. Рассмотрим частный случай задачи GC_k при $k = 2$.

Задача $\mathbf{GC}_2$. *Для произвольного графа $G = (V, E)$ найти такой граф $M^* \in \mathcal{M}_2(V)$, что*

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

Через $N_G(v)$ обозначим окрестность вершины v, т. е. множество вершин графа $G = (V, E)$, смежных с вершиной v. Через $\overline{N}_G(v)$ обозначим множество вершин графа G, не смежных с v: $\overline{N}_G(v) = V \setminus (N_G(v) \cup \{v\})$.

Для непустых множеств $V_1, V_2 \subseteq V$, образующих разбиение V, т. е. $V_1 \cap V_2 = \varnothing$ и $V_1 \cup V_2 = V$, обозначим через $M(V_1, V_2)$ кластерный граф из множества $\mathcal{M}_2(V)$ с компонентами связности, порождёнными множествами V_1, V_2. Сами множества V_1, V_2 будем называть *кластерами*.

Пусть $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — два графа с нумерованными вершинами на множестве V, $n = |V|$. Через $D(G_1, G_2)$ обозначим граф на множестве вершин V с множеством рёбер $E_1 \triangle E_2$. Заметим, что $\rho(G_1, G_2)$ равно количеству рёбер в графе $D(G_1, G_2)$.

Следующее утверждение легко доказывается с помощью леммы о рукопожатиях.

Лемма 1. *Пусть $d_{\min}$ — минимум степеней вершин в графе $D(G_1, G_2)$. Тогда*

$$\rho(G_1, G_2) \geqslant \frac{n d_{\min}}{2}.$$

Пусть $G = (V, E)$ — произвольный граф. Для вершины $v \in V$ и множества $A \subseteq V$ через A_v^+ обозначим число таких вершин $u \in A$, что $vu \in E$, а через A_v^- — число таких вершин $u \in A$, что $vu \notin E$.

Следующая лемма справедлива для произвольного графа $G = (V, E)$ и всех кластерных графов из множества $\mathcal{M}_2(V)$.

Лемма 2. *Пусть $G = (V, E)$ — произвольный граф, а $M_1 = M(X_1, Y_1)$ и $M_2 = M(X_2, Y_2)$ — два произвольных кластерных графа из множества $\mathcal{M}_2(V)$. Тогда*

$$\rho(G,M_1)-\rho(G,M_2)$$
$$=\sum_{u\in X_1\cap Y_2}((X_1\cap X_2)_u^- - (X_1\cap X_2)_u^+ + (Y_1\cap Y_2)_u^+ - (Y_1\cap Y_2)_u^-)$$
$$+\sum_{u\in Y_1\cap X_2}((Y_1\cap Y_2)_u^- - (Y_1\cap Y_2)_u^+ + (X_1\cap X_2)_u^+ - (X_1\cap X_2)_u^-).$$

Доказательство. Заметим, что $X_2=(X_1\cap X_2)\cup(Y_1\cap X_2)$ и $Y_2=(X_1\cap Y_2)\cup(Y_1\cap Y_2)$. Тогда расстояние между G и M_2 равно

$$\begin{aligned}\rho(G,M_2)=&\frac{1}{2}\sum_{u\in X_1\cap X_2}(X_1\cap X_2)_u^- + \frac{1}{2}\sum_{u\in Y_1\cap X_2}(Y_1\cap X_2)_u^-\\ &+\frac{1}{2}\sum_{u\in X_1\cap Y_2}(X_1\cap Y_2)_u^- + \frac{1}{2}\sum_{u\in Y_1\cap Y_2}(Y_1\cap Y_2)_u^-\\ &+\sum_{u\in X_1\cap Y_2}(X_1\cap X_2)_u^+ + \sum_{u\in Y_1\cap X_2}(Y_1\cap Y_2)_u^+ + \sum_{u\in Y_1\cap X_2}(X_1\cap X_2)_u^-\\ &+\sum_{u\in Y_1\cap X_2}(X_1\cap Y_2)_u^+ + \sum_{u\in Y_1\cap Y_2}(X_1\cap X_2)_u^+ + \sum_{u\in X_1\cap Y_2}(Y_1\cap Y_2)_u^-. \quad (1)\end{aligned}$$

Аналогично расстояние между G и M_1 равно

$$\begin{aligned}\rho(G,M_1)=&\frac{1}{2}\sum_{u\in X_1\cap X_2}(X_1\cap X_2)_u^- + \frac{1}{2}\sum_{u\in Y_1\cap X_2}(Y_1\cap X_2)_u^-\\ &+\frac{1}{2}\sum_{u\in X_1\cap Y_2}(X_1\cap Y_2)_u^- + \frac{1}{2}\sum_{u\in Y_1\cap Y_2}(Y_1\cap Y_2)_u^-\\ &+\sum_{u\in X_1\cap Y_2}(X_1\cap X_2)_u^- + \sum_{u\in Y_1\cap X_2}(Y_1\cap Y_2)_u^- + \sum_{u\in Y_1\cap X_2}(X_1\cap X_2)_u^+\\ &+\sum_{u\in Y_1\cap X_2}(X_1\cap Y_2)_u^+ + \sum_{u\in Y_1\cap Y_2}(X_1\cap X_2)_u^+ + \sum_{u\in X_1\cap Y_2}(Y_1\cap Y_2)_u^+. \quad (2)\end{aligned}$$

Вычитая (1) из (2), получим требуемое равенство. Лемма 2 доказана.

1.2. Процедура локального поиска. Рассмотрим следующую процедуру локального поиска.

Процедура LS(M,X,Y,x,y)

Вход: граф $G=(V,E)$, кластерный граф $M=M(X,Y)\in\mathcal{M}_2(V)$, $x\in X$, $y\in Y$.

Выход: кластерный граф $M'=M(X',Y')\in\mathcal{M}_2(V)$.

Итерация 0. Положим $X_0=X$, $Y_0=Y$.

Итерация k, $k \geqslant 1$.

Шаг 1. Для каждой вершины $u \in V \setminus \{x, y\}$ вычислить следующую величину $\delta_k(u)$ (изменение значения целевой функции при переносе вершины u в другой кластер). При $\delta_k(u) > 0$ эту величину будем называть *локальным улучшением вершины u на итерации k*:

$$\delta_k(u) = \begin{cases} (X_{k-1})_u^- - (X_{k-1})_u^+ + (Y_{k-1})_u^+ - (Y_{k-1})_u^- & \text{для } u \in X_{k-1} \setminus \{x\}, \\ (Y_{k-1})_u^- - (Y_{k-1})_u^+ + (X_{k-1})_u^+ - (X_{k-1})_u^- & \text{для } u \in Y_{k-1} \setminus \{y\}. \end{cases}$$

Шаг 2. Выбрать вершину $u_k \in V \setminus \{x, y\}$ такую, что

$$\delta_k(u_k) = \max_{u \in V \setminus \{x,y\}} \delta_k(u).$$

Шаг 3. Если $\delta_k(u_k) \leqslant 0$, то Стоп. Положить $X' = X_{k-1}$, $Y' = Y_{k-1}$, $M' = M(X', Y')$. Конец.

Шаг 4. Если $u_k \in X_{k-1}$, то положить $X_k = X_{k-1} \setminus \{u_k\}$, $Y_k = Y_{k-1} \cup \{u_k\}$. Если же $u_k \in Y_{k-1}$, то положить $X_k = X_{k-1} \cup \{u_k\}$, $Y_k = Y_{k-1} \setminus \{u_k\}$.

Перейти на итерацию $k + 1$.

Замечание 1. Кластерный граф M', который возвращает процедура LS, всегда принадлежит множеству $\mathcal{M}_2(V)$.

Это утверждение очевидно, поскольку вершины $x \in X$ и $y \in Y$ всегда лежат в разных кластерах.

Замечание 2. Трудоёмкость процедуры LS оценивается как $O(n^4)$.

Оценим трудоёмкость одной итерации процедуры LS. На k-й итерации для каждой вершины $u \in V \setminus \{x, y\}$ вычисляется величина $\delta_k(u)$. Для её вычисления необходимо определить смежность u со всеми вершинами множества $V \setminus \{u\}$. Таким образом, трудоёмкость вычисления величины $\delta_k(u)$ имеет порядок $O(n)$. Следовательно, трудоёмкость одной итерации процедуры LS — $O(n^2)$.

Теперь оценим количество итераций процедуры локального поиска LS. На k-й итерации процедура либо переносит вершину $u \in V \setminus \{x, y\}$ с максимальным значением $\delta_k(u)$ в противоположный кластер, уменьшая значение целевой функции на величину $\delta_k(u) \geqslant 1$, либо (если $\delta_k(u) \leqslant 0$ для всех вершин $u \in V \setminus \{x, y\}$) возвращает в качестве найденного локального оптимума граф $M'(X', Y')$, где $X' = X_{k-1}$, $Y' = Y_{k-1}$. Заметим, что значение целевой функции $\rho(G, M)$ удовлетворяет неравенству $0 \leqslant \rho(G, M) \leqslant \frac{n(n-1)}{2}$. Следовательно, общее количество итераций не превосходит по порядку n^2, а значит, трудоёмкость процедуры LS оценивается величиной порядка $O(n^4)$.

Отметим, что трудоёмкость процедуры LS можно уменьшить, если не пересчитывать на k-й итерации все величины $\delta_k(u)$, а использовать предыдущие значения $\delta_{k-1}(u)$ и корректировать их относительно перенесённой на итерации $k-1$ вершины u_{k-1}.

1.3. 2-Приближённый алгоритм для задачи GC_2. Прежде чем приступить к описанию 2-приближённого алгоритма для задачи GC_2, докажем вспомогательное утверждение.

Лемма 3. *Пусть $M^* = M(X^*, Y^*) \in \mathcal{M}_2(V)$ — оптимальное решение задачи GC_2 на n-вершинном графе $G = (V, E)$, где $|X^*| \geqslant 2$, $|Y^*| \geqslant 2$. Тогда для любой вершины $v \in V$ справедливо неравенство*

$$d_D(v) \leqslant \frac{n}{2},$$

где $D = D(G, M^*)$, $d_D(v)$ — степень вершины v в графе D.

Доказательство. Предположим, напротив, что существует такая вершина $w \in V$, что $d_D(w) > \frac{n}{2}$, т. е. $d_D(w) = \frac{n}{2} + c$, где $c > 0$, $\frac{n}{2} + c \in \mathbb{N}$ и $\frac{n}{2} + c \leqslant n - 1$. Рассмотрим граф $\widetilde{M} \in \mathcal{M}_2(V)$, полученный из графа M^* путём переноса вершины w в другой кластер. При этом, очевидно, $\widetilde{M}$ будет допустимым решением задачи GC_2 на G, поскольку $|X^*| \geqslant 2$, $|Y^*| \geqslant 2$. Рассмотрим граф $\widetilde{D} = D(G, \widetilde{M})$. Нетрудно заметить, что степень вершины w в графе $\widetilde{D}$ равна «нестепени» вершины w в графе D:

$$d_{\widetilde{D}}(w) = n - 1 - \frac{n}{2} - c = \frac{n}{2} - 1 - c.$$

Очевидно, что графы D и $\widetilde{D}$ отличаются лишь рёбрами вида $wu, u \in V$, остальные рёбра у них совпадают. Следовательно,

$$\begin{aligned}\rho(G, \widetilde{M}) - \rho(G, M^*) = |N_{\widetilde{D}}(w)| - |N_D(w)| = d_{\widetilde{D}}(w) - d_D(w)\\ = \frac{n}{2} - 1 - c - \frac{n}{2} - c = -(1 + 2c) < 0.\end{aligned}$$

Из полученного неравенства следует, что M^* не является оптимальным решением, что противоречит условию. Лемма 3 доказана.

Если убрать ограничение $|X^*| \geqslant 2$, $|Y^*| \geqslant 2$, то утверждение леммы 3 может стать неверным. Действительно, пусть дан полный граф K_n, $n \geqslant 3$. Оптимальное решение задачи GC_2 на K_n — граф M^*, в котором один кластер состоит из произвольной вершины графа K_n, а другой кластер порождён оставшимися вершинами. Нетрудно заметить, что в графе $D(K_n, M^*)$ степень вершины, образующей отдельный кластер, будет равна $n - 1$. При этом степени всех оставшихся вершин всё равно будут не больше чем $\frac{n}{2}$ (это неравенство доказывается так же, как и в лемме 3).

Рассмотрим алгоритм приближённого решения задачи GC_2.

Алгоритм A_1

Вход: граф $G = (V, E)$.

Выход: кластерный граф $M_1 = M(X, Y) \in \mathcal{M}_2(V)$.

Шаг 1. Для каждой упорядоченной пары вершин $(v, w) \in V^2$, $v \neq w$, выполнить

Шаг 1.1. Построить кластерный граф $M_{v,w} = M(X, Y) \in \mathcal{M}_2(V)$, где $X = \{v\} \cup (N_G(v) \setminus \{w\})$, $Y = V \setminus X$.

Шаг 1.2. Запустить процедуру локального поиска $\mathrm{LS}(M_{v,w}, X, Y, v, w)$. Обозначить полученный граф через $M'_{v,w}$.

Шаг 2. Среди всех локальных оптимумов $M'_{v,w}$, построенных на шаге 1.2, выбрать ближайший к G кластерный граф M_1:

$$\rho(G, M_1) = \min_{\substack{(v,w)\in V^2,\\ v\neq w}} \rho(G, M'_{v,w}).$$

Конец.

Поскольку количество упорядоченных пар вершин графа G не превосходит по порядку n^2, с учётом замечания 2 несложно оценить трудоёмкость алгоритма.

Замечание 3. Трудоёмкость алгоритма A_1 оценивается как $O(n^6)$.

Теорема 1. *Пусть $G = (V, E)$ — произвольный граф, а $M^* = M(X^*, Y^*)$ — оптимальное решение задачи GC_2 на G. Тогда среди всех графов, построенных алгоритмом A_1 на шаге* 1.1*, всегда существует такой граф $M_{v,w} = M(X, Y)$, что*

1) *$M_{v,w}$ может быть получен из графа M^* путём переноса не более чем $d_{\min} = d_D(v)$ вершин (здесь $D = D(G, M^*)$),*

2) *$v \in X \cap X^*$, $w \in Y \cap Y^*$.*

Доказательство. Рассмотрим в качестве v вершину минимальной степени в графе $D = D(G, M^*)$, т. е. $d_D(v) = d_{\min}$. Без ограничения общности будем считать, что $v \in X^*$. Тогда, очевидно, существует вершина w такая, что $w \in Y^*$. Рассмотрим кластерный граф $M_{v,w} = M(X, Y)$, построенный алгоритмом A_1 на шаге 1.1. Очевидно, что для этого графа выполняется п. 2 утверждения теоремы. Докажем, что этот граф может быть получен из графа M^* путём переноса не более чем $d_{\min}$ вершин в другой кластер (другими словами, для графа $M_{v,w}$ выполняется п. 1 утверждения теоремы).

По определению графа D

$$X^* = \{v\} \cup (N_G(v) \setminus N_D(v)) \cup (\overline{N}_G(v) \cap N_D(v)). \tag{3}$$

Нетрудно заметить, что

$$N_G(v) = (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)).$$

Возможны два случая.

Случай 1. Вершины v и w не смежны в G, т. е. $w \in \overline{N}_G(v) \cap \overline{N}_D(v)$. Тогда $N_G(v) \setminus \{w\} = N_G(v)$. Вычислим мощность множества $X^* \triangle X$. По определению графа $M_{v,w}$ имеем

$$\begin{aligned} X = \{v\} \cup (N_G(v) \setminus \{w\}) &= \{v\} \cup N_G(v) \\ &= \{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)). \end{aligned}$$

Тогда, используя (3), получим

$$\begin{aligned} X^* \triangle X &= (X^* \setminus X) \cup (X \setminus X^*) \\ &= (\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v)) = N_D(v). \end{aligned}$$

Таким образом, $|X^* \triangle X| = |N_D(v)| = d_D(v) = d_{\min}$, а значит, граф $M_{v,w}$ может быть получен из графа M^* путём переноса $d_{\min}$ вершин множества $N_D(v)$ в другой кластер.

Случай 2. Вершины v и w смежны в G, т. е. $w \in N_G(v) \cap N_D(v)$. Тогда $d_{\min} \geqslant 1$. Вычислим мощность множества $X^* \triangle X$. По определению графа $M_{v,w}$

$$\begin{aligned} X = \{v\} \cup (N_G(v) \setminus \{w\}) &= (\{v\} \cup N_G(v)) \setminus \{w\} \\ &= (\{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{w\}. \end{aligned}$$

Тогда, используя (3) и включение $w \in N_G(v) \cap N_D(v)$, получим

$$X^* \triangle X = ((\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{w\} = N_D(v) \setminus \{w\}.$$

Таким образом, $|X^* \triangle X| = |N_D(v)| - 1 = d_{\min} - 1$, а значит, граф $M_{v,w}$ может быть получен из графа M^* путём переноса $d_{\min} - 1$ вершин множества $N_D(v) \setminus \{w\}$ в другой кластер.

Итак, мы показали, что граф $M_{v,w}$ может быть получен из графа M^* путём переноса не более чем $d_{\min}$ вершин в другой кластер. Теорема 1 доказана.

Используя теорему 1, можем убрать ограничения $|X^*| \geqslant 2$, $|Y^*| \geqslant 2$ из леммы 3 и переписать её в других обозначениях.

Лемма 4. *Пусть $G = (V, E)$ — произвольный n-вершинный граф, $n \geqslant 3$, $M^* = M(X^*, Y^*)$ — оптимальное решение задачи GC_2 на графе G, а $M_{v,w}$ — кластерный граф, для которого справедлива теорема 1. Тогда для любой вершины $u \in V \setminus \{v, w\}$ справедливы следующие неравенства:*

1) *если $u \in X^*$, то $(X^*)^-_u + (Y^*)^+_u \leqslant \frac{n}{2}$,*
2) *если $u \in Y^*$, то $(X^*)^+_u + (Y^*)^-_u \leqslant \frac{n}{2}$.*

Доказательство. Возможны три случая.

Случай 1: $|X^*| \geqslant 2$, $|Y^*| \geqslant 2$. Доказательство этого случая аналогично доказательству леммы 3.

Случай 2: $|X^*| = 1$, $|Y^*| \geqslant 2$. Тогда в графе M^* кластер X^* состоит из единственной вершины v. Поскольку утверждение леммы должно быть справедливо для всех вершин $u \in V \setminus \{v, w\}$, доказательство этого случая также аналогично доказательству леммы 3.

Случай 3: $|X^*| \geqslant 2$, $|Y^*| = 1$. Этот случай доказывается аналогично случаю 2 путём замены вершины v вершиной w. Лемма 4 доказана.

Теперь можем приступить к доказательству основного утверждения этого раздела — гарантированной оценки точности алгоритма A_1.

Теорема 2. *Для любого графа $G = (V, E)$ верно неравенство*

$$\rho(G, M_1) \leqslant 2\rho(G, M^*),$$

где $M_1 \in \mathcal{M}_2(V)$ — решение, построенное алгоритмом A_1, а $M^* \in \mathcal{M}_2(V)$ есть оптимальное решение задачи GC_2 на графе G.

Доказательство. Пусть $M^* = M(X^*, Y^*)$ и v — вершина минимальной степени в графе $D = D(G, M^*)$. Согласно теореме 1 среди всех графов, построенных алгоритмом A_1 на шаге 1.1, всегда существует такой граф $M_{v,w} = M(X, Y)$, что

1) $M_{v,w}$ получен из графа M^* путём переноса не более чем $d_D(v) = d_{\min}$ вершин,

2) $v \in X \cap X^*, w \in Y \cap Y^*$.

Рассмотрим поведение процедуры $\mathrm{LS}(M_{v,w}, X, Y, v, w)$ на этом графе. Очевидно, что

$$|X \cap Y^*| \cup |Y \cap X^*| \leqslant d_{\min}.$$

Процедура локального поиска LS начинает свою работу с множеств $X_0 = X$ и $Y_0 = Y$. На каждой итерации k процедура LS либо переносит некоторую вершину $u_k \in V \setminus \{v, w\}$ в другой кластер, либо не переносит ни одной из вершин и заканчивает свою работу.

Рассмотрим итерацию $t + 1$, обладающую следующими свойствами:

• на каждой из итераций $k \in \{1, \ldots, t\}$ процедура LS выбирала некоторую вершину $u_k \in (X \cap Y^*) \cup (Y \cap X^*)$;

• на итерации $t + 1$ процедура LS либо выбирает вершину $u_{t+1} \in ((X \cap X^*) \cup (Y \cap Y^*)) \setminus \{v, w\}$, либо итерация $t + 1$ является последней для процедуры LS.

Введём следующие величины:

$$\alpha_{t+1}(u) = \begin{cases} (X_t \cap X^*)_u^- - (X_t \cap X^*)_u^+ + (Y_t \cap Y^*)_u^+ - (Y_t \cap Y^*)_u^-, \\ \qquad\qquad \text{если } u \in X_t \cap Y^*, \\ (Y_t \cap Y^*)_u^- - (Y_t \cap Y^*)_u^+ + (X_t \cap X^*)_u^+ - (X_t \cap X^*)_u^-, \\ \qquad\qquad \text{если } u \in Y_t \cap X^*, \end{cases}$$

$$\beta_{t+1}(u) = \begin{cases} (X_t \cap Y^*)_u^- - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^-, \\ \qquad\qquad \text{если } u \in X_t \cap Y^*, \\ (Y_t \cap X^*)_u^- - (Y_t \cap X^*)_u^+ + (X_t \cap Y^*)_u^+ - (X_t \cap Y^*)_u^-, \\ \qquad\qquad \text{если } u \in Y_t \cap X^*. \end{cases}$$

Тогда для каждой вершины $u \in (X_t \cap Y^*) \cup (Y_t \cap X^*)$ величина $\delta_{t+1}(u)$ раскладывается в сумму величин $\alpha_{t+1}(u)$ и $\beta_{t+1}(u)$:

$$\delta_{t+1}(u) = \alpha_{t+1}(u) + \beta_{t+1}(u). \tag{4}$$

Действительно, если $u \in X_t \cap Y^*$, то

$$\begin{aligned} \delta_{t+1}(u) &= (X_t)_u^- - (X_t)_u^+ + (Y_t)_u^+ - (Y_t)_u^- \\ &= (X_t \cap X^*)_u^- - (X_t \cap X^*)_u^+ + (Y_t \cap Y^*)_u^+ - (Y_t \cap Y^*)_u^- \\ &\quad + (X_t \cap Y^*)_u^- - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^- \\ &= \alpha_{t+1}(u) + \beta_{t+1}(u). \end{aligned}$$

Для вершин $u \in Y_t \cap X^*$ равенство (4) доказывается аналогично.

Рассмотрим кластерный граф $M_t = M(X_t, Y_t)$. Согласно лемме 2

$$\rho(G, M_t) - \rho(G, M^*) = \sum_{u \in X_t \cap Y^*} \alpha_{t+1}(u) + \sum_{u \in Y_t \cap X^*} \alpha_{t+1}(u).$$

Поскольку на итерациях $k \in \{1, \ldots, t\}$ процедурой LS перемещались только вершины из множества $(X \cap Y^*) \cup (Y \cap X^*)$, то

$$|X_t \cap Y^*| + |Y_t \cap X^*| = r \leqslant d_{\min}. \tag{5}$$

Тогда

$$\rho(G, M_t) - \rho(G, M^*) \leqslant r \max\{\alpha_{t+1}(u) \mid u \in (X_t \cap Y^*) \cup (Y_t \cap X^*)\}. \tag{6}$$

Для каждой вершины $u \in V \setminus \{v, w\}$ оценим величину её локального улучшения $\delta_{t+1}(u)$ (другими словами, как изменится значение целевой функции при переносе вершины u в другой кластер).

(а) Докажем, что для всех вершин $u \in ((X_t \cap X^*) \cup (Y_t \cap Y^*)) \setminus \{v, w\}$ справедливо неравенство

$$\delta_{t+1}(u) \leqslant 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1. \tag{7}$$

Приведём доказательство для вершин $u \in (X_t \cap X^*) \setminus \{v\}$. Для вершин $u \in (Y_t \cap Y^*) \setminus \{w\}$ неравенство доказывается аналогично с помощью симметричной замены X_t и X^* на Y_t и Y^*.

Заметим, что

$$(Y_t \cap X^*)_u^+ + (Y_t \cap X^*)_u^- + (X_t \cap Y^*)_u^+ + (X_t \cap Y^*)_u^- = |Y_t \cap X^*| + |X_t \cap Y^*|, \quad (8)$$

$$(X_t \cap X^*)_u^+ + (X_t \cap X^*)_u^- + (Y_t \cap Y^*)_u^+ + (Y_t \cap Y^*)_u^- = n - 1 - |Y_t \cap X^*| - |X_t \cap Y^*|. \quad (9)$$

Согласно лемме 4

$$(X_t \cap X^*)_u^- + (Y_t \cap Y^*)_u^+ \leqslant (X^*)_u^- + (Y^*)_u^+ \leqslant \frac{n}{2}. \quad (10)$$

По определению для вершины u из множества $(X_t \cap X^*) \setminus \{v\}$ величина $\delta_{t+1}(u)$ равна

$$\begin{aligned}\delta_{t+1}(u) &= (X_t)_u^- - (X_t)_u^+ + (Y_t)_u^+ - (Y_t)_u^- \\ &= (X_t \cap X^*)_u^- - (X_t \cap X^*)_u^+ + (Y_t \cap Y^*)_u^+ - (Y_t \cap Y^*)_u^- \\ &\quad + (X_t \cap Y^*)_u^- - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^-.\end{aligned}$$

Добавим и вычтем величину $(X_t \cap X^*)_u^- + (Y_t \cap Y^*)_u^+$. Тогда

$$\begin{aligned}\delta_{t+1}(u) &= 2((X_t \cap X^*)_u^- + (Y_t \cap Y^*)_u^+) \\ &\quad - (X_t \cap X^*)_u^- - (Y_t \cap Y^*)_u^+ - (X_t \cap X^*)_u^+ - (Y_t \cap Y^*)_u^- \\ &\quad + (X_t \cap Y^*)_u^- - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^-.\end{aligned}$$

Используя (9) и (10), получим

$$\begin{aligned}\delta_{t+1}(u) &\leqslant 2\frac{n}{2} - ((X_t \cap X^*)_u^- + (Y_t \cap Y^*)_u^+ + (X_t \cap X^*)_u^+ + (Y_t \cap Y^*)_u^-) \\ &\quad + (X_t \cap Y^*)_u^- - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^- \\ &= n - n + 1 + |Y_t \cap X^*| + |X_t \cap Y^*| + (X_t \cap Y^*)_u^- \\ &\quad - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^-.\end{aligned}$$

Поскольку все члены в правой части неравенства неотрицательны, то

$$\begin{aligned}(X_t \cap Y^*)_u^- - (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ - (Y_t \cap X^*)_u^- \\ \leqslant (X_t \cap Y^*)_u^- + (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ + (Y_t \cap X^*)_u^-.\end{aligned}$$

Тогда, используя (8), получим

$$\begin{aligned}\delta_{t+1}(u) &\leqslant |Y_t \cap X^*| + |X_t \cap Y^*| + 1 \\ &\quad + (X_t \cap Y^*)_u^- + (X_t \cap Y^*)_u^+ + (Y_t \cap X^*)_u^+ + (Y_t \cap X^*)_u^-\end{aligned}$$

$$\leqslant |Y_t \cap X^*| + |X_t \cap Y^*| + 1 + |Y_t \cap X^*| + |X_t \cap Y^*| = 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1.$$

(b) Теперь докажем, что для всех вершин $u \in (Y_t \cap X^*) \cup (X_t \cap Y^*)$

$$\delta_{t+1}(u) \leqslant 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1. \tag{11}$$

Действительно, если итерация $t+1$ является последней итерацией процедуры LS, то

$$\delta_{t+1}(u) \leqslant 0 < 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1.$$

Если же итерация $t+1$ не последняя, то процедура LS выбрала для переноса некоторую вершину $u_{t+1} \in ((X \cap X^*) \cup (Y \cap Y^*)) \setminus \{v, w\}$. Поскольку на итерациях $k \in \{1, \dots t\}$ процедурой LS перемещались только вершины из $(X \cap Y^*) \cup (Y \cap X^*)$, то $X \cap X^* \subset X_t \cap X^*$ и $Y \cap Y^* \subset Y_t \cap Y^*$. Следовательно, $u_{t+1} \in ((X_t \cap X^*) \cup (Y_t \cap Y^*)) \setminus \{v, w\}$, а значит, согласно (7) имеем

$$\delta_{t+1}(u) \leqslant \delta_{t+1}(u_{t+1}) \leqslant 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1.$$

Далее докажем, что на итерации $t+1$ для каждой вершины u из множества $(X_t \cap Y^*) \cup (Y_t \cap X^*)$ справедливо неравенство

$$\alpha_{t+1}(u) \leqslant \frac{n}{2}. \tag{12}$$

Приведём доказательство для вершин $u \in Y_t \cap X^*$. Для вершин $u \in X_t \cap Y^*$ неравенство доказывается аналогично с помощью симметричной замены X_t и X^* на Y_t и Y^*.

Предположим, напротив, что существует такая вершина $p \in Y_t \cap X^*$, что $\alpha_{t+1}(p) > \frac{n}{2}$. Согласно (4) имеем

$$\beta_{t+1}(p) = \delta_{t+1}(p) - \alpha_{t+1}(p) < \delta_{t+1}(p) - \frac{n}{2}.$$

Тогда в силу (11) $\delta_{t+1}(p) \leqslant 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1$, а значит,

$$\beta_{t+1}(p) < 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1 - \frac{n}{2}. \tag{13}$$

Так как $d_{\min}$ — наименьшая из степеней графа $D = D(G, M^*)$, то

$$d_D(p) = (Y_t \cap X^*)_p^- + (X_t \cap X^*)_p^- + (X_t \cap Y^*)_p^+ + (Y_t \cap Y^*)_p^+ \geqslant d_{\min}.$$

Тогда, используя (5), получим

$$(Y_t \cap X^*)_p^- + (X_t \cap X^*)_p^- + (X_t \cap Y^*)_p^+ + (Y_t \cap Y^*)_p^+ \geqslant |Y_t \cap X^*| + |X_t \cap Y^*|. \tag{14}$$

Поскольку $p \in Y_t \cap X^*$, то

$$(Y_t \cap X^*)_p^+ + (Y_t \cap X^*)_p^- + (X_t \cap Y^*)_p^+ + (X_t \cap Y^*)_p^- = |Y_t \cap X^*| + |X_t \cap Y^*| - 1, \tag{15}$$

$$(X_t \cap X^*)_p^+ + (X_t \cap X^*)_p^- + (Y_t \cap Y^*)_p^+ + (Y_t \cap Y^*)_p^- = n - |Y_t \cap X^*| - |X_t \cap Y^*|. \quad (16)$$

Используя (15), получим

$$\begin{aligned}\beta_{t+1}(p) &= (Y_t \cap X^*)_p^- - (Y_t \cap X^*)_p^+ + (X_t \cap Y^*)_p^+ - (X_t \cap Y^*)_p^- \\ &= (Y_t \cap X^*)_p^- + (X_t \cap Y^*)_p^+ + (Y_t \cap X^*)_p^- \\ &\quad + (X_t \cap Y^*)_p^+ - |Y_t \cap X^*| - |X_t \cap Y^*| + 1 \\ &= 2((Y_t \cap X^*)_p^- + (X_t \cap Y^*)_p^+) - |Y_t \cap X^*| - |X_t \cap Y^*| + 1.\end{aligned}$$

Из (14) следует, что

$$(Y_t \cap X^*)_p^- + (X_t \cap Y^*)_p^+ \geqslant |Y_t \cap X^*| + |X_t \cap Y^*| - (X_t \cap X^*)_p^- - (Y_t \cap Y^*)_p^+,$$

поэтому

$$\begin{aligned}\beta_{t+1}(p) &\geqslant 2(|Y_t \cap X^*| + |X_t \cap Y^*| - (X_t \cap X^*)_p^- - (Y_t \cap Y^*)_p^+) \\ &\quad - |Y_t \cap X^*| - |X_t \cap Y^*| + 1 = |Y_t \cap X^*| + |X_t \cap Y^*| + 1 \\ &\quad - 2(X_t \cap X^*)_p^- - 2(Y_t \cap Y^*)_p^+.\end{aligned}$$

Добавим и вычтем величину $(X_t \cap X^*)_p^+ + (Y_t \cap Y^*)_p^-$. Тогда

$$\begin{aligned}\beta_{t+1}(p) &\geqslant |Y_t \cap X^*| + |X_t \cap Y^*| + 1 \\ &\quad + (Y_t \cap Y^*)_p^- - (Y_t \cap Y^*)_p^+ + (X_t \cap X^*)_p^+ - (X_t \cap X^*)_p^- \\ &\quad - ((Y_t \cap Y^*)_p^- + (Y_t \cap Y^*)_p^+ + (X_t \cap X^*)_p^+ + (X_t \cap X^*)_p^-).\end{aligned}$$

Используя (16), получим

$$\begin{aligned}\beta_{t+1}(p) &\geqslant |Y_t \cap X^*| + |X_t \cap Y^*| + 1 \\ &\quad + (Y_t \cap Y^*)_p^- - (Y_t \cap Y^*)_p^+ + (X_t \cap X^*)_p^+ - (X_t \cap X^*)_p^- - n \\ &\quad + |Y_t \cap X^*| + |X_t \cap Y^*| = 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1 - n \\ &\quad + (Y_t \cap Y^*)_p^- - (Y_t \cap Y^*)_p^+ + (X_t \cap X^*)_p^+ - (X_t \cap X^*)_p^-.\end{aligned}$$

Остаётся заметить, что поскольку $p \in Y_t \cap X^*$, то

$$\alpha_{t+1}(p) = (Y_t \cap Y^*)_p^- - (Y_t \cap Y^*)_p^+ + (X_t \cap X^*)_p^+ - (X_t \cap X^*)_p^-,$$

а значит,

$$\begin{aligned}\beta_{t+1}(p) &\geqslant 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1 - n + \alpha_{t+1}(p) \\ &> 2(|Y_t \cap X^*| + |X_t \cap Y^*|) + 1 - \frac{n}{2}.\end{aligned}$$

Последнее неравенство противоречит (13). Значит, в силу произвольности вершины p для каждой вершины $u \in Y_t \cap X^*$ имеет место неравенство (12).

Используя (5), (6), (12), а также лемму 1, выводим

$$\begin{aligned}\rho(G, M'_{v,w}) - \rho(G, M^*) &\leqslant \rho(G, M_t) - \rho(G, M^*) \\ &\leqslant r \max\{\alpha_{t+1}(u) \mid u \in (X_t \cap Y^*) \cup (Y_t \cap X^*)\} \\ &\leqslant r\,\frac{n}{2} \leqslant d_{\min}\,\frac{n}{2} \leqslant \rho(G, M^*).\end{aligned}$$

Значит, $\rho(G, M'_{v,w}) \leqslant 2\rho(G, M^*)$.

Граф $M'_{v,w}$ будет построен на шаге 1.2 алгоритма A_1, откуда получим

$$\rho(G, M_1) \leqslant \rho(G, M'_{v,w}) \leqslant 2\rho(G, M^*).$$

Теорема 2 доказана.

2. 2-Приближённый алгоритм для задачи $\mathbf{SGC}_2$

Рассмотрим следующий вариант задачи кластеризации на графе с частичным обучением.

Задача $\mathbf{SGC}_2$. *Для произвольных графа $G = (V, E)$ и множества $Z = \{z_1, z_2\} \subset V$ найти граф $M^* \in \mathcal{M}_2(V)$ такой, что*

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

где минимум берётся по всем таким графам $M = (V, E_M) \in \mathcal{M}_2(V)$, что $z_1 z_2 \notin E_M$, т. е. z_1, z_2 лежат в разных компонентах связности графа M.

В этом разделе мы покажем, что по аналогии с задачей GC_2 можно построить 2-приближённый алгоритм для задачи SGC_2. Для этого понадобится ряд вспомогательных утверждений.

Лемма 5. *Пусть $M^* = M(X^*, Y^*) \in \mathcal{M}_2(V)$ — оптимальное решение задачи SGC_2 на n-вершинном графе $G = (V, E)$, $Z = \{z_1, z_2\} \subset V$. Тогда для любой вершины $v \in V \setminus Z$ справедливы следующие неравенства:*

1) *если $v \in X^*$, то $(X^*)_v^- + (Y^*)_v^+ \leqslant \frac{n}{2}$,*
2) *если $v \in Y^*$, то $(X^*)_v^+ + (Y^*)_v^- \leqslant \frac{n}{2}$.*

Доказательство аналогично доказательству леммы 4.

Рассмотрим алгоритм приближённого решения задачи SGC_2.

Алгоритм $\mathbf{A}_2$

Вход: граф $G = (V, E)$, $\{z_1, z_2\} \subset V$.

Выход: кластерный граф $M_2 = M(X, Y) \in \mathcal{M}_2(V)$, вершины z_1, z_2 лежат в разных кластерах.

Шаг 1. Для каждой вершины $v \in V$ выполнить:

Шаг 1.1. (а) Если $v \notin \{z_1, z_2\}$, то построить два кластерных графа $\overline{M}_v = M(\overline{X}, \overline{Y})$ и $\overline{\overline{M}}_v = M(\overline{\overline{X}}, \overline{\overline{Y}})$, где

$$\overline{X} = \{v\} \cup ((N_G(v) \cup \{z_1\}) \setminus \{z_2\}), \quad \overline{Y} = V \setminus \overline{X},$$
$$\overline{\overline{X}} = \{v\} \cup ((N_G(v) \cup \{z_2\}) \setminus \{z_1\}), \quad \overline{\overline{Y}} = V \setminus \overline{\overline{X}}.$$

(б) Если $v \in \{z_1, z_2\}$, то построить граф $M_v = M(X, Y)$, где

$$X = \{v\} \cup (N_G(v) \setminus \{z\}), Y = V \setminus X.$$

Здесь $z = z_1$, если $v = z_2$, или $z = z_2$, если $v = z_1$.

Шаг 1.2. (а) Если $v \notin \{z_1, z_2\}$, то дважды применить процедуру локального поиска LS: $\mathrm{LS}(\overline{M}_v, \overline{X}, \overline{Y}, z_1, z_2)$ и $\mathrm{LS}(\overline{\overline{M}}_v, \overline{\overline{X}}, \overline{\overline{Y}}, z_1, z_2)$. Полученные графы обозначить через M'_v и M''_v.

(б) Если $v \in \{z_1, z_2\}$, то применить процедуру локального поиска $\mathrm{LS}(M_v, X, Y, z_1, z_2)$. Обозначить полученный граф через M_v.

Шаг 2. Среди всех локальных оптимумов, построенных на шаге 1.2, выбрать ближайший к G кластерный граф $M_2 = M(X, Y)$.

Конец.

Поскольку количество графов, построенных на шаге 1, оценивается величиной порядка $O(n)$, то с учётом замечания 2 несложно оценить трудоёмкость алгоритма.

Замечание 4. Трудоёмкость алгоритма A_2 оценивается как $O(n^5)$.

Докажем теорему, являющуюся аналогом теоремы 1 для алгоритма A_2.

Теорема 3. *Пусть $G = (V, E)$ — произвольный граф, $\{z_1, z_2\} \subset V$, а $M^* = M(X^*, Y^*)$ — оптимальное решение задачи SGC_2 на G. Тогда среди всех графов, построенных алгоритмом A_2 на шаге* 1.1, *существует такой граф $M_v = M(X, Y)$, что*

1) *M_v может быть получен из графа M^* путём переноса не более чем $d_D(v) = d_{\min}$ вершин (здесь $D = D(G, M^*)$),*

2) *если $z_1 \in X^*$, $z_2 \in Y^*$, то $z_1 \in X \cap X^*$, $z_2 \in Y \cap Y^*$; если же $z_1 \in Y^*$, $z_2 \in X^*$, то $z_1 \in Y \cap Y^*$, $z_2 \in X \cap X^*$.*

Доказательство. Рассмотрим в качестве v вершину минимальной степени в графе $D = D(G, M^*)$, т. е. $d_D(v) = d_{\min}$. Без ограничения общности будем считать, что $v \in X^*$. Тогда по определению графа $D = D(G, M^*)$ имеем

$$X^* = \{v\} \cup (N_G(v) \setminus N_D(v)) \cup (\overline{N}_G(v) \cap N_D(v)), \tag{17}$$
$$N_G(v) = (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)). \tag{18}$$

Возможны четыре случая.

Случай 1: $v = z_1$. Тогда $M_v = M(X, Y)$ согласно п. (б) шага 1.1, где

$$X = \{v\} \cup (N_G(v) \setminus \{z_2\}), \quad Y = V \setminus X.$$

П. 2 утверждения теоремы очевиден по построению графа M_v. Докажем п. 1. Для этого достаточно оценить мощность $X^* \triangle X$.

Если $z_2 \notin N_G(v)$, то согласно (18) имеем

$$\begin{aligned} X = \{v\} \cup (N_G(v) \setminus \{z_2\}) &= \{v\} \cup N_G(v) \\ &= \{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)). \end{aligned}$$

Тогда, используя (17), получим

$$\begin{aligned} X^* \triangle X &= (X^* \setminus X) \cup (X \setminus X^*) \\ &= (\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v)) = N_D(v). \end{aligned}$$

Если же $z_2 \in N_G(v)$, то в силу (18) имеем

$$\begin{aligned} X = \{v\} \cup (N_G(v) \setminus \{z_2\}) &= (\{v\} \cup N_G(v)) \setminus \{z_2\} \\ &= (\{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{z_2\}. \end{aligned}$$

Тогда ввиду (17) получим

$$X^* \triangle X = ((\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{z_2\} = N_D(v) \setminus \{z_2\}.$$

Отсюда следует, что $|X^* \triangle X| \leqslant |N_D(v)| = d_{\min}$. Значит, граф M_v может быть получен из графа M^* путём переноса не более чем $d_{\min}$ вершин множества $N_D(v)$, т. е. п. 1 утверждения теоремы выполнен.

Случай 2: $v = z_2$. Доказательство этого случая аналогично доказательству случая 1.

Случай 3: $v \notin \{z_1, z_2\}$, $z_1 \in X^*$, $z_2 \in Y^*$. Тогда согласно п. (а) шага 1.1 в качестве $M_v = M(X, Y)$ рассмотрим граф $\overline{M_v} = M(\overline{X}, \overline{Y})$, где

$$\overline{X} = \{v\} \cup ((N_G(v) \cup \{z_1\}) \setminus \{z_2\}), \quad \overline{Y} = V \setminus \overline{X}.$$

П. 2 утверждения теоремы очевиден по построению графа M_v, а п. 1 может быть доказан оценкой мощности $X^* \triangle X$.

Если $z_1 \in N_G(v)$, $z_2 \in N_G(v)$, то согласно (18) имеем

$$\begin{aligned} X = \{v\} \cup ((N_G(v) \cup \{z_1\}) \setminus \{z_2\}) &= (\{v\} \cup N_G(v)) \setminus \{z_2\} \\ &= (\{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{z_2\}. \end{aligned}$$

Тогда, используя (17), получим

$$X^* \triangle X = ((\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{z_2\} = N_D(v) \setminus \{z_2\}.$$

Если $z_1 \notin N_G(v)$, $z_2 \notin N_G(v)$, то в силу (18) имеем

$$X = \{v\} \cup ((N_G(v) \cup \{z_1\}) \setminus \{z_2\}) = \{v\} \cup N_G(v) \cup \{z_1\}$$
$$= \{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)) \cup \{z_1\}.$$

Тогда ввиду (17) получим

$$X^* \triangle X = ((\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v))) \setminus \{z_1\} = N_D(v) \setminus \{z_1\}.$$

Если $z_1 \in N_G(v)$, $z_2 \notin N_G(v)$, то согласно (18) имеем

$$X = \{v\} \cup ((N_G(v) \cup \{z_1\}) \setminus \{z_2\}) = \{v\} \cup N_G(v)$$
$$= \{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)).$$

Тогда, используя (17), получим

$$X^* \triangle X = (\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v)) = N_D(v).$$

Если $z_1 \notin N_G(v)$, $z_2 \in N_G(v)$, то в силу (18) имеем

$$X = \{v\} \cup ((N_G(v) \cup \{z_1\}) \setminus \{z_2\}) = (\{v\} \cup N_G(v) \cup \{z_1\}) \setminus \{z_2\}$$
$$= (\{v\} \cup (N_G(v) \setminus N_D(v)) \cup (N_G(v) \cap N_D(v)) \cup \{z_1\}) \setminus \{z_2\}.$$

Тогда ввиду (17) получим

$$X^* \triangle X = ((\overline{N}_G(v) \cap N_D(v)) \cup (N_G(v) \cap N_D(v)))$$
$$\setminus \{z_1, z_2\} = N_D(v) \setminus \{z_1, z_2\}.$$

Отсюда следует, что $|X^* \triangle X| \leqslant |N_D(v)| = d_{\min}$. Значит, граф M_v может быть получен из графа M^* путём переноса не более чем $d_{\min}$ вершин множества $N_D(v)$, т. е. п. 1 утверждения теоремы выполнен.

Случай 4: $v \notin \{z_1, z_2\}$, $z_1 \in Y^*$, $z_2 \in X^*$. Доказательство этого случая аналогично доказательству случая 3. Теорема 3 доказана.

Из леммы 5 и теоремы 3 следует гарантированная оценка точности алгоритма A_2.

Теорема 4. *Для произвольных графа $G = (V, E)$ и множества $\{z_1, z_2\} \subset V$ верно неравенство*

$$\rho(G, M) \leqslant 2\rho(G, M^*),$$

где $M \in \mathcal{M}_2(V)$ — решение, построенное алгоритмом A_2, *а $M^* \in \mathcal{M}_2(V)$ есть оптимальное решение задачи* SGC_2 *на графе G.*

Доказательство теоремы 4 аналогично доказательству теоремы 2, только вместо леммы 4 и теоремы 1 используются лемма 5 и теорема 3.

ЛИТЕРАТУРА

1. **Bansal N., Blum A., Chawla S.** Correlation clustering // Mach. Learn. 2004. Vol. 56, No. 1–3. P. 89–113.
2. **Coleman T., Saunderson J., Wirth A.** A local-search 2-approximation for 2-correlation-clustering // Algorithms – ESA 2008. Proc. 16th Annu. Eur. Symp. (Karlsruhe, Germany, Sept. 15–17, 2008). Heidelberg: Springer, 2008. P. 308–319. (Lect. Notes Comput. Sci.; Vol. 5193).
3. **Shamir R., Sharan R., Tsur D.** Cluster graph modification problems // Discrete Appl. Math. 2004. Vol. 144, No. 1–2. P. 173–182.
4. **Zahn C. T.** Approximating symmetric relations by equivalence relations // J. Soc. Ind. Appl. Math. 1964. Vol. 12, No. 4. P. 840–847.
5. **Ильев В. П., Фридман Г. Ш.** К задаче аппроксимации графами с фиксированным числом компонент // Докл. АН СССР. 1982. Т. 264, № 3. С. 533–538.
6. **Агеев А. А., Ильев В. П., Кононов А. В., Талевнин А. С.** Вычислительная сложность задачи аппроксимации графов // Дискрет. анализ и исслед. операций. Сер. 1. 2006. Т. 13, № 1. С. 3–11.
7. **Giotis I., Guruswami V.** Correlation clustering with a fixed number of clusters // Theory Comput. 2006. Vol. 2, No. 1. P. 249–266.
8. **Chapelle O., Schölkopf B., Zein A.** Semi-supervised learning. Cambridge, MA: MIT Press, 2006.
9. **Bair E.** Semi-supervised clustering methods // Wiley Interdisciplinary Reviews: Computational Statistics. 2013. Vol. 5, No. 5. P. 349–361.

Ильев Виктор Петрович
Ильева Светлана Диадоровна
Моршинин Александр Владимирович

Статья поступила
10 января 2020 г.
После доработки —
6 мая 2020 г.
Принята к публикации
25 мая 2020 г.

DISKRETNYI ANALIZ I ISSLEDOVANIE OPERATSII
/DISCRETE ANALYSIS AND OPERATIONS RESEARCH/
July–September 2020. Vol. 27, No. 3. P. 88–108

UDC 519.8 DOI 10.33048/daio.2020.27.680

2-APPROXIMATION ALGORITHMS FOR TWO GRAPH CLUSTERING PROBLEMS

V. P. Il'ev[1,2,a], *S. D. Il'eva*[1], and *A. V. Morshinin*[2,b]

[1] Dostoevsky Omsk State University,
55a Mira Avenue, 644077 Omsk, Russia

[2] Omsk Branch of Sobolev Institute of Mathematics,
13 Pevtsov Street, 644043 Omsk, Russia

E-mail: [a] iljev@mail.ru, [b] morshinin.alexander@gmail.com

Abstract. We study a version of the graph 2-clustering problem and the related semi-supervised problem. In these problems, given an undirected graph, we have to find a nearest 2-cluster graph, i. e. a graph on the same vertex set with exactly two nonempty connected components each of which is a complete graph. The distance between two graphs is the number of noncoinciding edges. The problems under consideration are NP-hard. In 2008, Coleman, Saunderson, and Wirth presented a polynomial time 2-approximation algorithm for the analogous problem in which the number of clusters does not exceed 2. Unfortunately, the method of proving the performance guarantee of the Coleman, Saunderson, and Wirth algorithm is inappropriate for the graph 2-clustering problem in which the number of clusters equals 2. We propose a polynomial time 2-approximation algorithm for the 2-clustering problem on an arbitrary graph. In contrast to the proof by Coleman, Saunderson, and Wirth, our proof of the performance guarantee of this algorithm does not use switchings. Moreover, we propose an analogous 2-approximation algorithm for the related semi-supervised problem. Bibliogr. 9.

Keywords: graph, clustering, NP-hard problem, approximation algorithm, guaranteed approximation ratio.

REFERENCES

1. N. Bansal, A. Blum, and **S. Chawla,** Correlation clustering, *Mach. Learn.* **56** (1–3), 89–113 (2004).

English ver.: Journal of Applied and Industrial Mathematics **14** (3), 490–502 (2020), DOI 10.1134/S1990478920030084.

2. T. Coleman, J. Saunderson, and **A. Wirth,** A local-search 2-approximation for 2-correlation-clustering, in *Algorithms – ESA 2008* (Proc. 16th Annu. Eur. Symp., Karlsruhe, Germany, Sept. 15–17, 2008) (Springer, Heidelberg, 2008), pp. 308–319 (Lect. Notes Comput. Sci., Vol. 5193).

3. R. Shamir, R. Sharan, and **D. Tsur,** Cluster graph modification problems, *Discrete Appl. Math.* **144** (1–2), 173–182 (2004).

4. C. T. Zahn, Approximating symmetric relations by equivalence relations, *J. Soc. Ind. Appl. Math.* **12** (4), 840–847 (1964).

5. V. P. Il'ev and **G. Š. Fridman,** On the problem of approximation by graphs with a fixed number of components, *Dokl. AN SSSR* **264** (3), 533–538 (1982) [Russian] [*Soviet Math. Dokl.* **25** (3), 666–670 (1982)].

6. A. A. Ageev, V. P. Il'ev, A. V. Kononov, and **A. S. Talevnin,** Computational complexity of the graph approximation problem, *Diskretn. Anal. Issled. Oper., Ser. 1,* **13** (1), 3–11 (2006) [Russian] [*J. Appl. Ind. Math.* **1** (1), 1–8 (2007)].

7. I. Giotis and **V. Guruswami,** Correlation clustering with a fixed number of clusters, *Theory Comput.* **2** (1), 249–266 (2006).

8. O. Chapelle, B. Schölkopf, and **A. Zein,** Semi-Supervised Learning (MIT Press, Cambridge, MA, 2006).

9. E. Bair, Semi-supervised clustering methods, *Wiley Interdiscip. Rev., Comput. Stat.* **5** (5), 349–361 (2013).

Viktor P. Il'ev
Svetlana D. Il'eva
Aleksandr V. Morshinin

Received January 10, 2020
Revised May 6, 2020
Accepted May 25, 2020