ПРИКЛАДНАЯ ТЕОРИЯ ГРАФОВ

УДК 519.1

ОБ ОДНОЙ ЗАДАЧЕ КЛАСТЕРИЗАЦИИ ГРАФА С ЧАСТИЧНЫМ ОБУЧЕНИЕМ[1]

А. В. Ильев*,**, В. П. Ильев*,***

**Омский государственный технический университет, г. Омск, Россия,*
***Институт математики им. С. Л. Соболева СО РАН, г. Омск, Россия,*
****Омский государственный университет им. Ф. М. Достоевского, г. Омск, Россия*

В задачах кластеризации требуется разбить данное множество объектов на несколько подмножеств (кластеров) только на основе сходства объектов друг с другом. Рассматривается вариант задачи кластеризации графа, являющийся одной из формализаций задачи кластеризации с частичным обучением. Доказано, что эта задача является NP-трудной. Для одного варианта задачи предложен полиномиальный 3-приближённый алгоритм.

Ключевые слова: *граф, кластер, кластеризация с частичным обучением.*

DOI 10.17223/20710410/42/5

ON A SEMI-SUPERWIZED GRAPH CLUSTERING PROBLEM

A. V. Ilev*,**, V. P. Il'ev*,***

**Omsk State Technical University, Omsk, Russia*
***Sobolev Institute of Mathematics SB RAS, Omsk, Russia*
****Dostoevsky Omsk State University, Omsk, Russia*

E-mail: artyom_iljev@mail.ru, iljev@mail.ru

In the clustering problems one has to partition a given set of objects into some subsets (called clusters) taking into consideration only similarity of the objects. In this paper we study a version of the graph correlation clustering that can be considered as a formalization of the semi-supervised clustering. We prove that the problem under consideration is NP-hard and propose a polynomial-time 3-approximation algorithm for a version of the problem.

Keywords: *graph, cluster, semi-supervised clustering.*

Введение

Важный раздел теории распознавания образов и машинного обучения составляют методы решения задач кластеризации [1]. В *задаче кластеризации* требуется разбить данное множество объектов на несколько подмножеств (*кластеров*) только на основе сходства объектов друг с другом. Одной из наиболее наглядных формализаций

[1]Работа поддержана грантом РНФ № 17-11-01117.

задач кластеризации является *задача кластеризации графа* [2, 3]. В ней отношение сходства объектов задаётся посредством неориентированного графа, вершины которого взаимно однозначно соответствуют объектам, а рёбра соединяют похожие объекты, обладающие достаточным количеством одинаковых признаков. Требуется разбить множество исходных объектов на попарно непересекающиеся группы (кластеры) так, чтобы минимизировать число рёбер между кластерами и число недостающих рёбер внутри кластеров. Количество кластеров может быть задано, ограничено или заранее не определено.

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 [4, 5]. Задачи кластеризации графов имеют многочисленные приложения. Задачи кластеризации графов с частичным обучением тесно связаны с исследованием систем уравнений над графами [6].

В п. 1 настоящей работы рассматриваются три варианта задачи кластеризации графа, приводится краткий обзор известных результатов. В п. 2 рассматривается постановка задачи кластеризации графа, которая является одной из формализаций задачи кластеризации с частичным обучением. В этой задаче дано множество, состоящее из n объектов, которые необходимо распределить по k кластерам, $k < n$. Среди заданных объектов выделены k объектов, никакие два из которых не должны принадлежать одному и тому же кластеру. Отношение сходства объектов задано с помощью неориентированного графа. В работе доказано, что рассматриваемая задача является NP-трудной. Для случая $k = 2$ предложен полиномиальный 3-приближённый алгоритм.

1. Задачи кластеризации графов

1.1. Постановки задач

Будут рассматриваться только *обыкновенные* графы, т. е. графы без петель и кратных рёбер. Обыкновенный граф называется *кластерным*, если каждая его компонента связности является полным графом [7].

Обозначим через $\mathcal{M}(V)$ множество всех кластерных графов на множестве вершин V, $\mathcal{M}_k(V)$ — множество всех кластерных графов на множестве вершин V, имеющих ровно k непустых компонент связности, $\mathcal{M}_{1,k}(V)$ — 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 на одном и том же множестве вершин V, то *расстояние* $\rho(G_1, G_2)$ между ними определяется как

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

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

В 60–80-е годы XX века в литературе изучались следующие три варианта задачи кластеризации графа, которые в то время принято было называть *задачами аппроксимации графов*. Постановки и различные интерпретации задач аппроксимации графов можно найти в [8 – 11]. В дальнейшем задачи аппроксимации графов неоднократно переоткрывались и независимо изучались под разными названиями (Correlation Clustering [12], Cluster Editing [7, 13] и др.).

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

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

Рассматривались также варианты задач кластеризации графов, в которых ограничения накладываются на размеры кластеров [14].

1.2. Вычислительная сложность и приближённые алгоритмы

В 1986 г. М. Крживанек и Дж. Моравек показали [15], что задача **GC** является NP-трудной, однако их работа осталась незамеченной. В 2004 г. в [7, 12] доказана NP-трудность задачи **GC**. В [7] доказано также, что задача $\mathbf{GC_k}$ NP-трудна при любом фиксированном $k \geqslant 2$; в [16] приведено более простое доказательство этого же результата. В [17] доказано, что задачи $\mathbf{GC_2}$ и $\mathbf{GC_{1,2}}$ NP-трудны уже на кубических графах, откуда следует, что все упомянутые ранее варианты задачи кластеризации графа являются NP-трудными, включая и задачу $\mathbf{GC_{1,k}}$.

В 2004 г. в [12] предложен полиномиальный 3-приближённый алгоритм для задачи $\mathbf{GC_{1,2}}$. В 2006 г. в [17] доказано существование рандомизированной полиномиальной приближённой схемы для задачи $\mathbf{GC_{1,2}}$, а в [16] предложена рандомизированная полиномиальная приближённую схема для задачи $\mathbf{GC_k}$ (для любого фиксированного $k \geqslant 2$). Указав, что сложность полиномиальной приближенной схемы из [16] лишает её перспективы практического использования, авторы [18] предложили 2-приближённый алгоритм для задачи $\mathbf{GC_{1,2}}$, применив процедуру локального поиска к допустимому решению, полученному с помощью 3-приближённого алгоритма из [12]. Для задачи $\mathbf{GC_2}$ в работе [19] предложен $(3 - 6/n)$-приближённый алгоритм с достижимой гарантированной оценкой точности. В 2005 г. авторы [20] показали, что задача **GC** является APX-трудной и разработали для неё 4-приближённый алгоритм. В 2008 г. в [21] предложен 2,5-приближённый алгоритм для задачи **GC**.

Более подробный обзор результатов, относящихся к задачам **GC**, $\mathbf{GC_k}$, $\mathbf{GC_{1,k}}$ и их взвешенным вариантам, можно найти в [22].

2. Задача кластеризации с частичным обучением

2.1. Вычислительная сложность

Рассмотрим задачу кластеризации графа с частичным обучением, самая общая постановка которой такова.

Дан обыкновенный n-вершинный граф $G = (V, E)$ и целое число k, $2 \leqslant k \leqslant n$. Выделено множество попарно различных вершин $X = \{x_1, \ldots, x_k\} \subset V$, на котором задано бинарное симметричное иррефлексивное отношение P. Требуется найти ближайший к G граф $M^* \in \mathcal{M}(V)$, такой, что никакие две вершины множества X, связанные отношением P, не принадлежат одной и той же компоненте связности (т. е. одному кластеру) графа M^*.

Более подробно изучим следующий вариант задачи кластеризации графа с частичным обучением.

Задача $\mathbf{SGC_k}$. Дан обыкновенный n-вершинный граф $G = (V, E)$ и целое число k, $2 \leqslant k \leqslant n$. Выделено множество попарно различных вершин $X = \{x_1, \ldots, x_k\} \subset V$. Требуется найти такой граф $M^* \in \mathcal{M}_k(V)$, 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 $x_i, x_j \notin E_M$ для любых $i, j \in \{1, \ldots, k\}$; другими словами, никакие две вершины множества $X = \{x_1, \ldots, x_k\}$ не принадлежат одной и той же компоненте связности (т. е. одному кластеру) графа M.

Несложно свести по Тьюрингу задачу $\mathbf{GC_k}$ к $\mathbf{SGC_k}$ и тем самым показать, что задача $\mathbf{SGC_k}$ NP-трудна. Действительно, рассмотрим произвольный граф $G = (V, E)$ — вход задачи $\mathbf{GC_k}$ — и фиксируем целое число k и произвольный набор $\{x_1, \ldots, x_k\}$, состоящий из k попарно различных вершин графа G. Имея оптимальное решение $M(x_1, \ldots, x_k)$ задачи $\mathbf{SGC_k}$ для любого такого набора $\{x_1, \ldots, x_k\} \subset V$ и выбрав среди них ближайший к графу G кластерный граф

$$M^*(x_1, \ldots, x_k) = \operatorname*{arg\,min}_{\{x_1, \ldots, x_k\} \subset V} \rho(G, M(x_1, \ldots, x_k)),$$

мы, очевидно, получим оптимальное решение исходной задачи $\mathbf{GC_k}$. Легко видеть, что при фиксированном k построение всех C_n^k входов задачи $\mathbf{SGC_k}$ и получение оптимального решения исходной задачи $\mathbf{GC_k}$ можно выполнить за время $\mathrm{O}(n^k)$, где $n = |V|$. Таким образом, справедлива следующая теорема.

Теорема 1. Задача $\mathbf{SGC_k}$ NP-трудна при любом фиксированном $k \geqslant 2$.

2.2. Задача $\mathbf{SGC_2}$

Рассмотрим частный случай задачи $\mathbf{SGC_k}$, когда $k = 2$.

Пусть $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — n-вершинные графы на одном и том же множестве вершин V. Обозначим через $D(G_1, G_2)$ граф на множестве вершин V с множеством рёбер $E_1 \Delta E_2$. Через $d_D(u)$ обозначим степень вершины u в графе $D = D(G_1, G_2)$.

Нетрудно заметить, что $\rho(G_1, G_2)$ равно числу рёбер в графе D, которое по лемме о рукопожатиях равно половине суммы степеней вершин графа D. Отсюда получаем следующее утверждение.

Лемма 1 [19]. Пусть $d_{\min} = \min_{u \in V} d_D(u)$ — минимум степеней вершин в графе $D = D(G_1, G_2)$. Тогда

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

Введём следующие обозначения. Для произвольного графа $G = (V, E)$ и $u \in V$ обозначим $N_G(u) = \{v \in V : uv \in E\}$, $\overline{N_G}(u) = V \setminus (N_G(u) \cup \{u\})$. Таким образом, $|N_G(u)| = d_G(u)$ — степень вершины u в графе G.

Для множеств $V_1, V_2 \subseteq V$, таких, что $V_1 \cap V_2 = \varnothing$ и $V_1 \cup V_2 = V$, обозначим через $M(V_1, V_2)$ кластерный граф из класса $\mathcal{M}_2(V)$ с компонентами связности, порождёнными множествами V_1, V_2. Нам понадобится следующая простая лемма.

Лемма 2. Пусть $G = (V, E)$ — n-вершинный граф, $u \in V$ — произвольная вершина, $M, M' \in \mathcal{M}_2(V)$ — кластерные графы на множестве V, такие, что M' получен из M путём переноса вершины u в другую компоненту связности. Тогда

$$\rho(G, M') - \rho(G, M) \leqslant n - 1.$$

Доказательство. Пусть $M = M(V_1, V_2)$, где V_1 — множество, содержащее вершину u. Тогда $M' = M(V_1 \setminus \{u\}, V_2 \cup \{u\})$. Заметим, что графы $D = D(G, M)$ и $D' = D(G, M')$ отличаются только рёбрами вида uv, $v \in V \setminus \{u\}$, а именно: u, v смежны в D тогда и только тогда, когда они не смежны в D'. Остальные рёбра графов D и D' совпадают. Следовательно,

$$\rho(G, M') - \rho(G, M) = |N_{D'}(u)| - |N_D(u)| = d_{D'}(u) - d_D(u).$$

Очевидно, $d_D(u) \geqslant 0$, $d_{D'}(u) \leqslant n - 1$, поэтому

$$\rho(G, M') - \rho(G, M) = d_{D'}(u) - d_D(u) \leqslant n - 1,$$

что и требовалось. ■

2.3. Приближённый алгоритм для задачи $\mathbf{SGC_2}$

Пусть $x_1, x_2 \in V$ — фиксированные вершины графа $G = (V, E)$, $x_1 \neq x_2$, $v \in V$ — произвольная вершина в G. Построим граф $M_v \in \mathcal{M}_2(V)$ по следующим правилам:

(а) Если в графе G вершина v смежна ровно с одной из вершин x_1, x_2 и не совпадает с другой, то полагаем $M_v = M(V_1, V_2)$, где $V_1 = \{v\} \cup N_G(v)$, $V_2 = V \setminus V_1$.

(б) Если в графе G вершина v смежна с обеими вершинами x_1, x_2, то полагаем $M'_v = M(V'_1, V'_2)$. Здесь $V'_1 = (\{v\} \cup N_G(v)) \setminus \{x_2\}$, $V'_2 = V \setminus V'_1$, $M''_v = M(V''_1, V''_2)$, где $V''_1 = (\{v\} \cup N_G(v)) \setminus \{x_1\}$, $V''_2 = V \setminus V''_1$. Если при этом $\rho(G, M'_v) \leqslant \rho(G, M''_v)$, то полагаем $M_v = M'_v$, в противном случае $M_v = M''_v$.

(в) Если в графе G вершина v не смежна и не совпадает ни с одной из вершин x_1, x_2, то полагаем $M'_v = M(V'_1, V'_2)$. Здесь $V'_1 = \{v\} \cup N_G(v) \cup \{x_1\}$, $V'_2 = V \setminus V'_1$, $M''_v = M(V''_1, V''_2)$, где $V''_1 = \{v\} \cup N_G(v) \cup \{x_2\}$, $V''_2 = V \setminus V''_1$. Если при этом $\rho(G, M'_v) \leqslant \rho(G, M''_v)$, то полагаем $M_v = M'_v$, в противном случае $M_v = M''_v$.

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

Очевидно, что построенный по этим правилам кластерный граф M_v является допустимым решением задачи $\mathbf{SGC_2}$ для графа G.

Лемма 3. Пусть $M^* \in \mathcal{M}_2(V)$ — оптимальное решение задачи $\mathbf{SGC_2}$ на графе $G = (V, E)$, $D = D(G, M^*)$, $v \in V$ — вершина минимальной степени в графе D: $d_D(v) = \min\limits_{u \in V} d_D(u) = d_{\min}$. Тогда

$$\rho(G, M_v) \leqslant \rho(G, M^*) + d_{\min}(n - 1), \tag{1}$$

где $M_v = M(V_1, V_2)$ — кластерный граф, построенный по правилам (а)–(г).

Доказательство. Пусть $M^* = M(V_1^*, V_2^*)$, где V_1^* — множество, содержащее вершину v. Тогда из определения графа D вытекает, что

$$V_1^* = \{v\} \cup \Big(N_G(v) \setminus N_D(v)\Big) \cup \Big(\overline{N_G}(v) \cap N_D(v)\Big). \tag{2}$$

С л у ч а й 1. Пусть в графе $G = (V, E)$ вершина v смежна ровно с одной из вершин x_1, x_2 и не совпадает с другой. Покажем, что граф M_v, построенный по правилу (а), может быть получен из графа M^* путем переноса $d_{\min}$ вершин в другую компоненту. Для этого достаточно оценить, как сильно отличаются множества V_1 и V_1^*. По построению графа M_v имеем

$$V_1 = \{v\} \cup N_G(v) = \{v\} \cup \Big(N_G(v) \setminus N_D(v)\Big) \cup \Big(N_G(v) \cap N_D(v)\Big).$$

Учитывая (2), получаем

$$V_1^* \Delta V_1 = (V_1^* \setminus V_1) \cup (V_1 \setminus V_1^*) = \Big(\overline{N_G}(v) \cap N_D(v)\Big) \cup \Big(N_G(v) \cap N_D(v)\Big) = N_D(v).$$

Таким образом, $|V_1^* \Delta V_1| = |N_D(v)| = d_D(v) = d_{\min}$, поэтому граф M_v может быть получен из графа M^* путём переноса $d_{\min}$ вершин множества $N_D(v)$ в другую компоненту связности.

Нетрудно заметить, что если в графе M^* последовательно переносить все вершины множества $N_D(v)$ в другую компоненту, то после каждого переноса для вновь получаемых кластерных графов будут выполнены условия леммы 2. Так как граф M_v получен из графа M^* путём переноса $d_{\min}$ вершин множества $N_D(v)$ в другую компоненту, то, применяя $d_{\min}$ раз лемму 2, получаем $\rho(G, M_v) - \rho(G, M^*) \leqslant d_{\min}(n-1)$, т. е. $\rho(G, M_v) \leqslant \rho(G, M^*) + d_{\min}(n-1)$. Итак, неравенство (1) выполнено.

С л у ч а й 2. Пусть в графе $G = (V, E)$ вершина v смежна с обеими вершинами x_1, x_2, т. е. $x_1, x_2 \in N_G(v)$. В этом случае граф M_v строится по правилу (б). Так как вершины x_1 и x_2 находятся в разных компонентах связности графа M^* и $x_1, x_2 \in N_G(v)$, то $d_{\min} \geqslant 1$. Покажем, что один из графов M'_v, M''_v может быть получен из графа M^* путём переноса $d_{\min} - 1$ вершин в другую компоненту связности.

Рассмотрим случай, когда $x_1 \in V_1^*$ (следовательно, $x_2 \in V_2^*$). Докажем, что в этом случае граф M'_v может быть получен из графа M^* путём переноса $d_{\min} - 1$ вершин в другую компоненту связности. По построению графа M'_v справедливо

$$V'_1 = \Big(\{v\} \cup N_G(v)\Big) \setminus \{x_2\} = \Big(\{v\} \cup \Big(N_G(v) \setminus N_D(v)\Big) \cup \Big(N_G(v) \cap N_D(v)\Big)\Big) \setminus \{x_2\}.$$

В силу (2), учитывая, что $x_2 \in N_G(v) \cap N_D(v)$, получаем

$$V_1^* \Delta V'_1 = \Big(\Big(\overline{N_G}(v) \cap N_D(v)\Big) \cup \Big(N_G(v) \cap N_D(v)\Big)\Big) \setminus \{x_2\} = N_D(v) \setminus \{x_2\},$$

т. е. $|V_1^* \Delta V'_1| = |N_D(v)| - 1 = d_{\min} - 1$, поэтому граф M'_v может быть получен из графа M^* путём переноса $d_{\min} - 1$ вершин множества $N_D(v) \setminus \{x_2\}$ в другую компоненту связности.

Применяя $d_{\min} - 1$ раз лемму 2, получаем $\rho(G, M'_v) - \rho(G, M^*) \leqslant (d_{\min} - 1)(n-1)$, т. е. $\rho(G, M'_v) \leqslant \rho(G, M^*) + (d_{\min} - 1)(n-1)$.

В случае $x_2 \in V_1^*$ (значит, $x_1 \in V_2^*$) с помощью аналогичных рассуждений доказывается неравенство $\rho(G, M''_v) \leqslant \rho(G, M^*) + (d_{\min} - 1)(n-1)$. Таким образом,

$$\rho(G, M_v) = \min\Big(\rho(G, M'_v), \rho(G, M''_v)\Big) \leqslant \rho(G, M^*) + (d_{\min} - 1)(n-1) <$$
$$< \rho(G, M^*) + d_{\min}(n-1),$$

т. е. неравенство (1) выполнено.

С л у ч а й 3. Пусть в графе $G = (V, E)$ вершина v не смежна и не совпадает ни с одной из вершин x_1, x_2, т. е. $x_1, x_2 \in \overline{N_G}(v)$. В этом случае граф M_v строится по правилу (в). Так как вершины x_1 и x_2 находятся в разных компонентах связности графа M^* и $x_1, x_2 \in \overline{N_G}(v)$, то $d_{\min} \geqslant 1$. Покажем, что один из графов M'_v, M''_v может быть получен из графа M^* путём переноса $d_{\min} - 1$ вершин в другую компоненту связности.

Рассмотрим случай, когда $x_1 \in V_1^*$ (следовательно, $x_2 \in V_2^*$). Докажем, что в этом случае граф M'_v может быть получен из графа M^* путем переноса $d_{\min} - 1$ вершин в другую компоненту связности. По построению графа M'_v,

$$V'_1 = \{v\} \cup N_G(v) \cup \{x_1\} = \{v\} \cup \Big(N_G(v) \setminus N_D(v)\Big) \cup \Big(N_G(v) \cap N_D(v)\Big) \cup \{x_1\}.$$

В силу (2), учитывая, что $x_1 \in \overline{N_G}(v) \cap N_D(v)$, получаем

$$V_1^* \Delta V'_1 = \Big(\Big(\overline{N_G}(v) \cap N_D(v)\Big) \setminus \{x_1\}\Big) \cup \Big(N_G(v) \cap N_D(v)\Big) = N_D(v) \setminus \{x_1\},$$

т. е. $|V_1^* \Delta V'_1| = |N_D(v)| - 1 = d_{\min} - 1$, поэтому граф M'_v может быть получен из графа M^* путём переноса $d_{\min}-1$ вершин множества $N_D(v) \setminus \{x_1\}$ в другую компоненту. Рассуждая, как в случае 2, получаем $\rho(G, M'_v) \leqslant \rho(G, M^*) + (d_{\min} - 1)(n - 1)$.

В случае $x_2 \in V_1^*$ (значит, $x_1 \in V_2^*$) аналогично доказывается неравенство $\rho(G, M''_v) \leqslant \rho(G, M^*) + (d_{\min} - 1)(n - 1)$. Таким образом,

$$\rho(G, M_v) = \min\Big(\rho(G, M'_v), \rho(G, M''_v)\Big) < \rho(G, M^*) + d_{\min}(n - 1),$$

что и требовалось доказать.

С л у ч а й 4. Пусть в графе $G = (V, E)$ вершина v совпадает с одной из вершин x_1, x_2. В этом случае граф M_v строится по правилу (г). Без ограничения общности будем считать, что $v = x_1$. Тогда $V_1 = \Big(\{v\} \cup N_G(v)\Big) \setminus \{x_2\}$. Если $x_2 \notin N_G(v)$, то $V_1 = \{v\} \cup N_G(v)$ и неравенство (1) доказывается, как в случае 1, а если $x_2 \in N_G(v)$, то как в случае 2.

Итак, в каждом из четырёх случаев справедливо неравенство (1). ■

Рассмотрим следующий алгоритм приближённого решения задачи $\mathbf{SGC_2}$.

Алгоритм 1. Приближённое решение задачи $\mathbf{SGC_2}$

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

2: Среди всех графов M_v выбрать такой граф M, что $\rho(G, M) = \min\limits_{v \in V} \rho(G, M_v)$.

Вычислительная сложность алгоритма 1 может быть оценена как $\mathrm{O}(n^3)$, где n — число вершин графа G.

Справедлива следующая гарантированная оценка точности алгоритма 1.

Теорема 2. Для любого n-вершинного графа $G = (V, E)$ и любого множества $X = \{x_1, x_2\} \subset V$ $(x_1 \neq x_2)$ алгоритм 1 находит такой кластерный граф $M \in \mathcal{M}_2(V)$ — допустимое решение задачи $\mathbf{SGC_2}$, что

$$\rho(G, M) \leqslant \left(3 - \frac{2}{n}\right) \rho(G, M^*),$$

где $M^* \in \mathcal{M}_2(V)$ — оптимальное решение задачи $\mathbf{SGC_2}$ на графе G.

Доказательство. Очевидно, среди всех вершин на шаге 1 будет выбрана такая вершина $v \in V$, что $d_D(v) = d_{\min}$, поэтому $\rho(G, M) \leqslant \rho(G, M_v)$. Оценим величину $\rho(G, M_v)$ с учётом неравенства (1) и леммы 1:

$$\rho(G, M_v) \leqslant \rho(G, M^*) + d_{\min}(n-1) \leqslant \rho(G, M^*) + 2\rho(G, M^*)\left(1 - \frac{1}{n}\right) = (3 - \frac{2}{n})\rho(G, M^*).$$

Отсюда получаем гарантированную оценку точности алгоритма 1:

$$\rho(G, M) \leqslant \rho(G, M_v) \leqslant \left(3 - \frac{2}{n}\right)\rho(G, M^*).$$

Теорема 2 доказана. ■

ЛИТЕРАТУРА

1. *Журавлев Ю. И., Рязанов В. В., Сенько О. В.* Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2005.
2. *Kulis B., Basu S., Dhillon I., and Mooney R.* Semi-supervised graph clustering: a kernel approach // Mach. Learn. 2009. V. 74. No. 1. P. 1–22.
3. *Schaeffer S. E.* Graph clustering // Comput. Sci. Rev. 2005. V. 1. No. 1. P. 27–64.
4. *Bair E.* Semi-supervised clustering methods // Wiley Interdisciplinary Reviews: Computational Statistics. 2013. V. 5. No. 5. P. 349–361.
5. *Chapelle O., Schölkopf B., and Zein A.* Semi-supervised Learning. Cambridge, Massachusets: MIT Press, 2006.
6. *Ильев А. В., Ремесленников В. Н.* Исследование совместности систем уравнений над графами и нахождение их общих решений // Вестник Омского университета. 2017. № 4(86). С. 26–32.
7. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems // Discrete Appl. Math. 2004. V. 144. No. 1–2. P. 173–182.
8. *Zahn C. T.* Approximating symmetric relations by equivalence relations // J. Soc. Indust. Appl. Math. 1964. V. 12. No. 4. P. 840–847.
9. *Tomescu I.* La reduction minimale d'un graphe à une reunion de cliques // Discrete Math. 1974. V. 10. No. 1–2. P. 173–179.
10. *Фридман Г. Ш.* Исследование одной задачи классификации на графах // Методы моделирования и обработка информации. Новосибирск: Наука, 1976. С. 147–177.
11. *Ильев В. П., Фридман Г. Ш.* К задаче аппроксимации графами с фиксированным числом компонент // Докл. АН СССР. 1982. Т. 264. № 3. С. 533–538.
12. *Bansal N., Blum A., and Chawla S.* Correlation clustering // Mach. Learn. 2004. V. 56. No. 1–3. P. 89–113.
13. *Ben-Dor A., Shamir R., and Yakhimi Z.* Clustering gene expression patterns // J. Comput. Biol. 1999. V. 6. No. 3–4. P. 281–297.
14. *Ильев В. П., Навроцкая А. А.* Вычислительная сложность задачи аппроксимации графами с компонентами связности ограниченного размера // Прикладная дискретная математика. 2011. № 3(13). С. 80–84.
15. *Křivánek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering // Acta Inform. 1986. V. 23. P. 311–323.
16. *Giotis I. and Guruswami V.* Correlation clustering with a fixed number of clusters // Theory Comput. 2006. V. 2. No. 1. P. 249–266.
17. *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.

18. *Coleman T., Saunderson J., and Wirth A.* A local-search 2-approximation for 2-correlation clustering // ESA 2008. LNCS. 2008. V. 5193. P. 308–319.
19. *Ильев В. П., Ильева С. Д., Навроцкая А. А.* Приближенные алгоритмы для задач аппроксимации графов // Дискрет. анализ и исслед. операций. 2011. Т. 18. № 1. С. 41–60.
20. *Charikar M., Guruswami V., and Wirth A.* Clustering with qualitative information // J. Comput. Syst. Sci. 2005. V. 71. No. 3. P. 360–383.
21. *Ailon N., Charikar M., and Newman A.* Aggregating inconsistent information: Ranking and clustering // J. ACM. 2008. V. 55. No. 5. P. 1–27.
22. *Il'ev V., Il'eva S., Kononov A.* Short survey on graph correlation clustering with minimization criteria // DOOR 2016. LNCS. 2016. V. 9869. P. 25–36.

REFERENCES

1. *Zhuravlev Yu. I., Ryazanov V. V., and Sen'ko O. V.* Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya [Recognition. Mathematical methods. Program system. Practical applications]. Moscow, FAZIS Publ., 2005. (in Russian)
2. *Kulis B., Basu S., Dhillon I., and Mooney R.* Semi-supervised graph clustering: a kernel approach. Mach. Learn., 2009, vol. 74, no. 1, pp. 1–2.
3. *Schaeffer S. E.* Graph clustering. Comput. Sci. Rev., 2005, vol. 1, no. 1, pp. 27–64.
4. *Bair E.* Semi-supervised clustering methods. Wiley Interdisciplinary Reviews: Computational Statistics. 2013, vol. 5, no. 5, pp. 349–361.
5. *Chapelle O., Schölkopf B., and Zein A.* Semi-supervised Learning. Cambridge, Massachusets, MIT Press, 2006.
6. *Ilev A. V. and Remeslennikov V. N.* Issledovanie sovmestnosti sistem uravnenij nad grafami i nahozhdenie ih obshchih reshenij [Study of the compatibility of systems of equations over graphs and finding their general solutions]. Vestnik Omskogo Universiteta, 2017, no. 4(86), pp. 26–32. (in Russian)
7. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems. Discrete Appl. Math., 2004, vol. 144, no. 1–2, pp. 173–182.
8. *Zahn C. T.* Approximating symmetric relations by equivalence relations. J. Soc. Indust. Appl. Math., 1964, vol. 12, no. 4, pp. 840–847.
9. *Tomescu I.* La reduction minimale d'un graphe à une reunion de cliques. Discrete Math., 1974, vol. 10, no. 1–2, pp. 173–179.
10. *Fridman G. Sh.* Issledovanie odnoj zadachi klassifikacii na grafah [Investigation of a classifying problem on graphs]. In: Metody Modelirovaniya i Obrabotka Informatsii, Novosibirsk, Nauka Publ., 1976, pp. 147–177. (in Russian)
11. *Il'ev V. P. and Fridman G. Sh.* On the problem of approximation by graphs with a fixed number of components. Sov. Math. Dokl., 1982, vol. 25, pp. 666–670.
12. *Bansal N., Blum A., and Chawla S.* Correlation clustering. Mach. Learn., 2004, vol. 56, no. 1–3, pp. 89–113.
13. *Ben-Dor A., Shamir R., and Yakhimi Z.* Clustering gene expression patterns. J. Comput. Biol., 1999, vol. 6, no. 3–4, pp. 281–297.
14. *Il'ev V. P. and Navrotskaya A. A.* Vychislitel'naya slozhnost' zadachi approksimacii grafami s komponentami svyaznosti ogranichennogo razmera [Computational complexity of the problem of approximation by graphs with connected components of bounded size]. Prikladnaya Diskretnaya Matematika, 2011, no. 3(13), pp. 80–84. (in Russian)
15. *Křivánek M. and Morávek J.,* NP-hard problems in hierarchical-tree clustering. Acta inform., 1986, vol. 23, pp. 311–323.

16. *Giotis I. and Guruswami V.* Correlation clustering with a fixed number of clusters. Theory Comput., 2006, vol. 2, no. 1, pp. 249–266.
17. *Ageev A. A., Il'ev V. P., Kononov A. V., and Talevnin A. S.* Computational complexity of the graph approximation problem. J. Appl. Indust. Math., 2007, vol. 1, no. 1, pp. 1–8.
18. *Coleman T., Saunderson J., and Wirth A.* A local-search 2-approximation for 2-correlation clustering. ESA 2008, LNCS, 2008, vol. 5193, pp. 308–319.
19. *Il'ev V. P., Il'eva S. D., and Navrotskaya A. A.* Approximation algorithms for graph approximation problems. J. Appl. Indust. Math., 2011, vol. 5, no. 4, pp. 569–581.
20. *Charikar M., Guruswami V., and Wirth A.* Clustering with qualitative information. J. Comput. Syst. Sci., 2005, vol. 71, no. 3, pp. 360–383.
21. *Ailon N., Charikar M., and Newman A.* Aggregating inconsistent information: Ranking and clustering. J. ACM, 2008, vol. 55, no. 5, pp. 1–27.
22. *Il'ev V., Il'eva S., and Kononov A.* Short survey on graph correlation clustering with minimization criteria. DOOR 2016, LNCS, 2016, vol. 9869, pp. 25–36.