МАТЕМАТИЧЕСКИЕ ОСНОВЫ ИНФОРМАТИКИ И ПРОГРАММИРОВАНИЯ

УДК 510.52 DOI 10.17223/20710410/57/6

О ГЕНЕРИЧЕСКОЙ СЛОЖНОСТИ ОГРАНИЧЕННОЙ ПРОБЛЕМЫ КЛАСТЕРИЗАЦИИ ГРАФОВ[1]

А. Н. Рыбалов

Институт математики им. С. Л. Соболева СО РАН, г. Омск, Россия

E-mail: alexander.rybalov@gmail.com

Изучается генерическая сложность проблемы кластеризации графов с ограничениями на число кластеров. В этой проблеме структура взаимосвязей объектов задаётся с помощью графа, вершины которого соответствуют объектам, а рёбра соединяют похожие объекты. Требуется разбить множество объектов на ограниченное число попарно непересекающихся групп (кластеров) так, чтобы минимизировать число связей между кластерами и число недостающих связей внутри кластеров. Строится подпроблема этой проблемы, для которой, при условии $\mathrm{P} \neq \mathrm{NP}$ и $\mathrm{P} = \mathrm{BPP}$, не существует полиномиального генерического алгоритма.

Ключевые слова: *генерическая сложность, кластеризация графа.*

THE GENERIC COMPLEXITY OF THE BOUNDED PROBLEM OF GRAPHS CLUSTERING

A. N. Rybalov

Sobolev Institute of Mathematics, Omsk, Russia

Generic-case approach to algorithmic problems studies behavior of an algorithm on typical (almost all) inputs and ignores the rest of inputs. In this paper, we study the generic complexity of the bounded problem of graphs clustering. In this problem the structure of objects relations is presented as a graph: vertices correspond to objects, and edges connect similar objects. It is required to divide the set of objects into bounded disjoint groups (clusters) to minimize the number of connections between clusters and the number of missing links within clusters. We have constructed a subproblem of this problem, for which there is no polynomial generic algorithm provided $\mathrm{P} \neq \mathrm{NP}$ and $\mathrm{P} = \mathrm{BPP}$. To prove the theorem, we use the method of generic amplification, which allows to construct generically hard problems from the problems hard in the classical sense. The main component of this method is the cloning technique, which merges the inputs of a problem together into sufficiently large sets of equivalent inputs. Equivalence is understood in the sense that the problem for them is solved in a similar way.

Keywords: *generic complexity, graph clustering.*

[1]Работа поддержана грантом РНФ № 22-11-20019.

Введение

Одной из важных проблем машинного обучения является проблема кластеризации графов. В этой задаче структура взаимосвязей объектов задаётся с помощью графа, вершины которого соответствуют объектам, а рёбра соединяют похожие объекты. Требуется разбить множество объектов на попарно непересекающиеся группы (кластеры) так, чтобы минимизировать число связей между кластерами и число недостающих связей внутри кластеров. В работах [1 – 7] доказана NP-трудность проблемы кластеризации графа для различных её постановок. Таким образом, при условии $\mathrm{P} \neq \mathrm{NP}$ не существует полиномиального алгоритма для решения этой задачи, а при условии совпадения классов P и BPP (класс проблем, решаемых за полиномиальное время вероятностными алгоритмами) для неё не существует и полиномиальных вероятностных алгоритмов. Имеются серьёзные доводы в пользу равенства $\mathrm{P} = \mathrm{BPP}$. В частности, доказано [8], что это равенство следует из весьма правдоподобных гипотез о вычислительной сложности некоторых трудных проблем.

Генерический подход [9] — это один из подходов к изучению алгоритмических проблем для «почти всех» входов. В рамках этого подхода алгоритмическая проблема рассматривается не на всём множестве входов, а на некотором подмножестве «почти всех» входов. Такие входы образуют так называемое генерическое множество. Понятие «почти все» формализуется введением естественной меры на множестве входных данных. С точки зрения практики алгоритмы, решающие быстро проблему на генерическом множестве, так же хороши, как и быстрые алгоритмы для всех входов. Отметим, что похожий подход для изучения проблем оптимизации был предложен ранее в [10].

Большой интерес как с теоретической точки зрения, так и с точки зрения практических приложений представляют алгоритмические проблемы, которые остаются неразрешимыми или трудноразрешимыми и в генерическом случае. Например, в современной криптографии интересны такие проблемы, которые, являясь (гипотетически) трудными в классическом смысле, остаются трудными и в генерическом смысле, т.е. для почти всех входов. Это объясняется тем, что при случайной генерации ключей в криптографическом алгоритме происходит генерация входа некоторой трудной алгоритмической проблемы, лежащей в основе алгоритма. Если проблема будет генерически легкоразрешимой, то для почти всех таких входов ее можно будет быстро решить и ключи почти всегда будут нестойкими. Поэтому проблема должна быть трудной для почти всех входов. Например, таким поведением обладают классические алгоритмические проблемы криптографии: проблема распознавания квадратичных вычетов, проблема дискретного логарифма, проблема извлечения корня в группах вычетов (проблема обращения функции RSA).

Метод генерической амплификации [11] позволяет строить проблемы, неразрешимые или трудноразрешимые для почти всех входов. С его помощью была доказана генерическая неразрешимость и трудноразрешимость многих алгоритмических проблем: проблема остановки для машин Тьюринга [12], проблема равенства в некоторых конечно определенных полугруппах [11], проблема разрешимости элементраных теорий, неразрешимых в классическом случае [11], арифметика Пресбургера [13], десятая проблема Гильберта [14], проблема выполнимости булевых формул [15], проблема кластеризации графов [16], проблема распознавания гамильтоновых графов [17].

Данная работа посвящена изучению генерической сложности ограниченной проблемы кластеризации графов. Эта задача является проблемой кластеризации графов, в которой число кластеров ограничено заранее заданной константой $k \geqslant 2$. В [3] была

доказана NP-трудность такой задачи для любого $k \geqslant 2$. В данной работе строится подпроблема этой проблемы, для которой, при условии $\mathrm{P} \neq \mathrm{NP}$ и $\mathrm{P} = \mathrm{BPP}$, не существует полиномиального генерического алгоритма. Отметим, что в [16] была изучена генерическая сложность проблемы кластеризации графов без ограничения на число кластеров.

1. Предварительные сведения

Пусть I — некоторое множество входов, а I_n — подмножество входов размера n. Для подмножества $S \subseteq I$ определим последовательность

$$\rho_n(S) = \frac{|S_n|}{|I_n|},\ n = 1, 2, 3, \ldots,$$

где $S_n = S \cap I_n$ — множество входов из S размера n. Заметим, что $\rho_n(S)$ — это вероятность попасть в S при случайной и равновероятной генерации входов из I_n. *Асимптотической плотностью* S назовём предел

$$\rho(S) = \lim_{n\to\infty} \rho_n(S).$$

Множество S называется *пренебрежимым*, если $\rho(S) = 0$.

Алгоритм $\mathcal{A}$ с множеством входов I и множеством выходов $J \cup \{?\}$ $(? \notin J)$ называется *генерическим*, если

1) $\mathcal{A}$ останавливается на всех входах из I;
2) множество $\{x \in I : \mathcal{A}(x) = ?\}$ является пренебрежимым.

Генерический алгоритм $\mathcal{A}$ вычисляет функцию $f : I \to J$, если для всех $x \in I$ выполнено

$$(\mathcal{A}(x) = y \in J) \Rightarrow (f(x) = y).$$

Ситуация $\mathcal{A}(x) = ?$ означает, что $\mathcal{A}$ не может вычислить функцию f на аргументе x. Но условие 2 гарантирует, что $\mathcal{A}$ корректно вычисляет f на почти всех входах (входах из генерического множества).

Напомним также некоторые понятия классической теории сложности вычислений. *Время работы* $t_M(x)$ машины Тьюринга M на входе $x \in I$ – это число шагов машины от начала работы до остановки. Если M на x не останавливается, полагаем $t_M(x) = \infty$. Машина Тьюринга M *полиномиальна*, если существует полином $p(n)$, такой, что для любого $x \in I$ имеет место $t_M(x) < p(|x|)$. Класс P состоит из подмножеств I, распознаваемых полиномиальными машинами Тьюринга.

Вероятностная машина Тьюринга — это машина Тьюринга, в программе которой допускаются пары правил вида

$$(q_i, a) \to (q_j, b, S_1),$$
$$(q_i, a) \to (q_k, c, S_2).$$

В процессе работы такой машины с вероятностью $1/2$ выбирается первое правило и с вероятностью $1/2$ — второе. Обозначим через $P(M(x) = y)$ вероятность того, что машина M на входе x выдаёт ответ y. Время работы $t_M(x, \tau)$ вероятностной машины Тьюринга на входе x зависит от вычислительного пути (последовательности выполненных команд) τ. Проблема $S \subseteq I$ принадлежит *классу* BPP, если существуют вероятностная машина Тьюринга M и полином $p(n)$, такие, что

1) Для любого x и для любого вычислительного пути τ машины M на x имеет место $t_M(x,\tau) < p(|x|)$.
2) Если $x \in S$, то $P(M(x) = 1) > 2/3$.
3) Если $x \notin S$, то $P(M(x) = 0) > 2/3$.

Вероятностные машины Тьюринга формализуют понятие алгоритма, использующего генератор случайных чисел. Класс BPP — это класс проблем, эффективно решаемых такими вероятностными алгоритмами. Большинство специалистов по теоретической информатике сейчас считает, что имеет место равенство P = BPP. Это равенство означает, что любой полиномиальный вероятностный алгоритм можно эффективно дерандомизировать, т. е. построить полиномиальный детерминированный алгоритм, решающий ту же задачу. Хотя равенство пока не доказано, имеются серьёзные результаты в пользу него [8].

2. Ограниченная проблема кластеризации графов

Будем рассматривать неориентированные графы без петель и кратных рёбер. Граф называется *кластерным*, если каждая его компонента связности является полным графом. Обозначим через $\mathcal{M}_{\leqslant k}(V)$ множество всех кластерных графов на множестве вершин V, имеющих не более k компонент связности. Если $G_1 = (V, E_1)$ и $G_2 = (V, E_2)$ — графы на одном и том же множестве вершин V, то *расстояние* $\rho(G_1, G_2)$ между ними есть число несовпадающих рёбер в графах G_1 и G_2, то есть

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

Ограниченная проблема кластеризации графов $\mathcal{GC}_{\leqslant k}$ состоит в следующем. Задан граф $G = (V, E)$ и целое число $k \geqslant 2$. Найти такой граф $M^* \in \mathcal{M}_{\leqslant k}(V)$, что

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

В [3] доказана NP-трудность этой проблемы для любого $k \geqslant 2$. В частности, это означает, что существует полиномиально эквивалентная ей проблема распознавания из класса NP, которая является NP-полной.

Рассмотрим бесконечную последовательность графов $\gamma = \{G_1, G_2, \ldots, G_n, \ldots\}$, такую, что G_n имеет n вершин для любого n. Напомним, что два графа G_1 и G_2 называются изоморфными (обозначается $G_1 \cong G_2$), если существует биекция π между множествами вершин G_1 и G_2, такая, что для любых вершин v, u графа G_1 v и u соединены ребром в G_1 тогда и только тогда, когда $\pi(v)$ и $\pi(u)$ соединены ребром в G_2. Биекция π, осуществляющая изоморфизм, является перестановкой множества вершин графов G_1, G_2, если вершины обоих графов занумерованы числами $\{1, 2, \ldots, n\}$. Будем обозначать также $G_2 = \pi(G_1)$.

Для каждой последовательности графов γ определим проблему $\mathcal{GC}_{\leqslant k}(\gamma)$ как ограничение проблемы $\mathcal{GC}_{\leqslant k}$ на множество входов $\{G : G \cong G_n, G_n \in \gamma\}$. Заметим, что множество всех входов размера n в проблеме $\mathcal{GC}_{\leqslant k}(\gamma)$ состоит из всевозможных графов G с произвольным графом G, изоморфным фиксированному графу G_n из последовательности γ. Очевидно, что проблема $\mathcal{GC}_{\leqslant k}(\gamma)$ является подпроблемой ограниченной проблемы кластеризации графов. Следующее утверждение говорит о том, что эта проблема для некоторых последовательностей γ может быть так же трудна, как и ограниченная проблема кластеризации графов.

Лемма 1. Если не существует полиномиального вероятностного алгоритма для решения проблемы $\mathcal{GC}_{\leqslant k}$, то найдется последовательность графов γ, такая, что не существует полиномиального вероятностного алгоритма для решения проблемы $\mathcal{GC}_{\leqslant k}(\gamma)$.

Доказательство. Пусть $P_1, P_2, \ldots$ — все полиномиальные вероятностные алгоритмы. Если не существует полиномиального вероятностного алгоритма для проблемы $\mathcal{GC}_{\leqslant k}$, то для любого вероятностного полиномиального алгоритма P_n найдётся бесконечно много графов, для которых P_n не может решить $\mathcal{GC}_{\leqslant k}$. Из этого следует, что можно выбрать такую последовательность $\gamma' = \{G_1, G_2, \ldots, G_n, \ldots\}$, что алгоритм P_n не может решить $\mathcal{GC}_{\leqslant k}$ для G_n для всех n. Более того, γ' упорядочена по возрастанию числа вершин в графах. Теперь можно расширить последовательность γ' до последовательности графов γ с графами G_n для всех размеров n. Из построения γ следует, что не существует полиномиального вероятностного алгоритма для решения проблемы $\mathcal{GC}_{\leqslant k}(\gamma)$. ■

3. Основной результат

Для изучения генерической сложности проблемы кластеризации графов будем использовать представление графов с помощью матриц смежности. Под размером графа будем понимать число вершин.

Теорема 1. Пусть γ — произвольная последовательность графов. Если существует генерический полиномиальный алгоритм, решающий проблему $\mathcal{GC}_{\leqslant k}(\gamma)$, то существует вероятностный полиномиальный алгоритм, решающий эту проблему на всём множестве входов.

Доказательство. Допустим, что существует генерический полиномиальный алгоритм $\mathcal{A}$, решающий проблему ограниченной кластеризации графов $\mathcal{GC}_{\leqslant k}(\gamma)$. Построим вероятностный полиномиальный алгоритм $\mathcal{B}$, решающий эту проблему на всём множестве входов. На графе G с n вершинами алгоритм $\mathcal{B}$ работает следующим образом:

1) Запускает алгоритм $\mathcal{A}$ на G.
2) Если $\mathcal{A}(G) \neq ?$, то $\mathcal{B}$ выдаёт ответ $\mathcal{A}(G)$ и останавливается, иначе идёт на шаг 3.
3) Генерирует случайно и равномерно перестановку π на вершинах $\{1, \ldots, n\}$ и вычисляет граф $G' = \pi(G)$.
4) Запускает алгоритм $\mathcal{A}$ на (G').
5) Если $\mathcal{A}(G') = ?$, выдаёт полный граф на всех вершинах (возможно, неправильный ответ).
6) Пусть $\mathcal{A}(G') = M^*$ — решение задачи ограниченной кластеризации для графа G' и $M^* \in \mathcal{M}_{\leqslant k}(\{1, \ldots, n\})$. Тогда если $M^* = C_1 \cup C_2 \cup \ldots C_m$, где C_i — непересекающиеся кластеры на множестве вершин $\{1, \ldots, n\}$, то легко видеть, что

$$\pi^{-1}(M^*) = \pi^{-1}(C_1) \cup \pi^{-1}(C_2) \cup \ldots \pi^{-1}(C_m)$$

является решением задачи ограниченной кластеризации для исходного графа $G = \pi^{-1}(G')$.

Для доказательства корректности работы вероятностного алгоритма надо показать, что вероятность того, что $\mathcal{A}(G') = ?$, меньше $1/3$. Заметим, что $\pi(G)$ при варьировании перестановки π пробегает всё множество входов размера n. Множество $\{G : \mathcal{A}(G) = ?\}$ пренебрежимо, поэтому вероятность того, что $\mathcal{A}(G') = ?$, стремится к 0 при увеличении n. ■

Непосредственным следствием теоремы 1 является следующее утверждение:

Теорема 2. Если P $\neq$ NP и P = BPP, то существует последовательность графов γ, такая, что для решения ограниченной проблемы кластеризации $\mathcal{GC}_{\leqslant k}(\gamma)$ не существует генерического полиномиального алгоритма.

Доказательство. Покажем сначала, что, при условиях $\mathrm{P} \neq \mathrm{NP}$ и $\mathrm{P} = \mathrm{BPP}$, не существует полиномиального вероятностного алгоритма для решения проблемы $\mathcal{GC}_{\leqslant k}$. Действительно, пусть такой алгоритм существует. Так как проблема $\mathcal{GC}_{\leqslant k}$ является NP-трудной, то существует полиномиально эквивалентная ей NP-полная проблема распознавания A. Из полиномиального вероятностного алгоритма для $\mathcal{GC}_{\leqslant k}$ легко получается полиномиальный вероятностный алгоритм для решения проблемы A. А так как $\mathrm{P} = \mathrm{BPP}$, то существует и детерминированный полиномиальный алгоритм для A, откуда $\mathrm{P} = \mathrm{NP}$. Противоречие.

Теперь нужное утверждение следует из леммы 1 и теоремы 1. ■

Автор выражает благодарность рецензенту за полезные замечания и предложения по улучшению текста статьи.

ЛИТЕРАТУРА

1. *Křivanek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering // Acta Informatica. 1986. V. 23. P. 311–323.
2. *Bansal N., Blum A., and Chawla S.* Correlation clustering // Machine Learning. 2004. V. 56. P. 89–113.
3. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems // Discrete Appl. Math. 2004. V. 144. No. 1–2. P. 173–182.
4. *Агеев А. А., Ильев В. П., Кононов А. В., Талевнин А. С.* Вычислительная сложность задачи аппроксимации графов // Дискретный анализ и исследование операций. Сер. 1. 2006. Т. 13. № 1. С. 3–11.
5. *Ильев В. П., Ильева С. Д.* О задачах кластеризации графов // Вестник Омского университета. 2016. № 2. С. 16–18.
6. *Ильев А. В., Ильев В. П.* Об одной задаче кластеризации графа с частичным обучением // Прикладная дискретная математика. 2018. № 42. С. 66–75.
7. *Талевнин А. С.* О сложности задачи аппроксимации графов // Вестник Омского университета. 2004. № 4. С. 22–24.
8. *Impagliazzo R. and Wigderson A.* $\mathrm{P} = \mathrm{BPP}$ unless E has subexponential circuits: Derandomizing the XOR Lemma // Proc. 29th STOC. El Paso: ACM, 1997. P. 220–229.
9. *Kapovich I., Miasnikov A., Schupp P., and Shpilrain V.* Generic-case complexity, decision problems in group theory and random walks // J. Algebra. 2003. V. 264. No. 2. P. 665–694.
10. *Гимади Э. Х., Глебов Н. И., Перепелица В. А.* Алгоритмы с оценками для задач дискретной оптимизации // Проблемы кибернетики. 1975. Т. 31. С. 35–42.
11. *Myasnikov A. G. and Rybalov A.N.* Generic complexity of undecidable problems // J. Symbolic Logic. 2008. V. 73. No. 2. P. 656–673.
12. *Rybalov A. N.* On the strongly generic undecidability of the Halting Problem // Theor. Comput. Sci. 2007. V. 377. P. 268–270.
13. *Rybalov A. N.* Generic complexity of Presburger arithmetic // Theory Comput. Systems. 2010. V. 46. No. 1. P. 2–8.
14. *Rybalov A. N.* Generic complexity of the Diophantine problem // Groups Complexity Cryptology. 2013. V. 5. No. 1. P. 25–30.
15. *Rybalov A. N.* Generic hardness of the Boolean satisfiability problem // Groups Complexity Cryptology. 2017. V. 9. No. 2. P. 151–154.
16. *Рыбалов А. Н.* О генерической сложности проблемы кластеризации графов // Прикладная дискретная математика. 2019. № 46. С. 72–77.

17. *Рыбалов А. Н.* О генерической сложности проблемы распознавания гамильтоновых путей // Прикладная дискретная математика. 2021. № 53. С. 120–126.

REFERENCES

1. *Křivanek M. and Morávek J.* NP-hard problems in hierarchical-tree clustering. Acta Informatica, 1986, vol. 23, pp. 311–323.
2. *Bansal N., Blum A., and Chawla S.* Correlation clustering. Machine Learning, 2004, vol. 56, pp. 89–113.
3. *Shamir R., Sharan R., and Tsur D.* Cluster graph modification problems. Discrete Appl. Math., 2004, vol. 144, no. 1–2, pp. 173–182.
4. *Ageev A. A., Il'ev V. P., Kononov A. V., and Talevnin A. S.* Computational complexity of the graph approximation problem. J. Appl. Ind. Math., 2007, vol. 1, no. 1, pp. 1–8.
5. *Il'ev V. P. and Il'eva S. D.* O zadachah klasterizacii grafov [On problems of graph clustering]. Vestnik Omskogo Universiteta, 2016, no. 2, pp. 16–18. (in Russian)
6. *Il'ev A. V. and Il'ev V. P.* Ob odnoy zadache klasterizatsii grafa s chastichnym obucheniem [On a semi-superwized graph clustering problem]. Prikladnaya Diskretnaya Matematika, 2018, no. 42, pp. 66–75. (in Russian)
7. *Talevnin A. S.* O slozhnosti zadachi approksimacii grafov [On the complexity of the graph approximation problem]. Vestnik Omskogo Universiteta, 2004, no. 4, pp. 22–24. (in Russian)
8. *Impagliazzo R. and Wigderson A.* P = BPP unless E has subexponential circuits: Derandomizing the XOR Lemma. Proc. 29th STOC, El Paso, ACM, 1997, pp. 220–229.
9. *Kapovich I., Miasnikov A., Schupp P., and Shpilrain V.* Generic-case complexity, decision problems in group theory and random walks. J. Algebra, 2003, vol. 264, no. 2, pp. 665–694.
10. *Gimadi E. H., Glebov N. I., and Perepelitsa V. A.* Algoritmy s ocenkami dlya zadach diskretnoi optimizacii [Algorithms with bounds for problems of discrete optimization]. Problemy Kibernetiki, 1975, V. 31, pp. 35–42. (in Russian)
11. *Myasnikov A. G. and Rybalov A. N.* Generic complexity of undecidable problems. J. Symbolic Logic, 2008, vol. 73, no. 2, pp 656–673.
12. *Rybalov A. N.* On the strongly generic undecidability of the Halting Problem. Theor. Comput. Sci., 2007, vol. 377, pp. 268–270.
13. *Rybalov A. N.* Generic complexity of Presburger arithmetic. Theory Comput. Systems, 2010, vol. 46, no. 1, pp. 2–8.
14. *Rybalov A. N.* Generic complexity of the Diophantine problem. Groups Complexity Cryptology, 2013, vol. 5, no. 1, pp. 25–30.
15. *Rybalov A. N.* Generic hardness of the Boolean satisfiability problem. // Groups Complexity Cryptology, 2017, vol. 9, no. 2, pp. 151–154.
16. *Rybalov A. N.* O genericheskoy slozhnosti problemy klasterizatsii grafov [On generic complexity of the graph clustering problem]. Prikladnaya Diskretnaya Matematika, 2019, no. 46, pp. 72–77. (in Russian)
17. *Rybalov A. N.* O genericheskoy slozhnosti problemy raspoznavaniya gamil'tonovykh putey [The general complexity of the problem to recognize Hamiltonian paths]. Prikladnaya Diskretnaya Matematika, 2021, no. 53, pp. 120–126. (in Russian)