Вестн. Ом. ун-та. 2016. № 2. С. 16–18.

УДК 519.8

В.П. Ильев, С.Д. Ильева

О ЗАДАЧАХ КЛАСТЕРИЗАЦИИ ГРАФОВ*

В работе представлен краткий обзор результатов по вычислительной сложности и аппроксимируемости различных вариантов задачи кластеризации графа, известной так же как задача аппроксимации графа. Кроме того, исследуется новый вариант задачи кластеризации с частичным обучением. Доказано, что рассматриваемая задача является NP-трудной. В случае, когда число кластеров равно 2, доказана NP-трудность задачи на кубических графах.

Ключевые слова: граф, кластеризация, аппроксимация, NP-трудная задача.

Введение

В *задаче кластеризации* требуется разбить заданное множество объектов на несколько подмножеств (*кластеров*) только на основе сходства объектов друг с другом. Мера сходства оценивается по-разному в разных задачах. 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 [1; 2].

Одной из наиболее наглядных формализаций задач кластеризации взаимосвязанных объектов является *задача аппроксимации графа*, которая представляет собой один из вариантов *задачи кластеризации графа* [3; 4]. В этой задаче структура взаимосвязей объектов задается посредством неориентированного графа, вершины которого взаимно однозначно соответствуют объектам, а ребра соединяют похожие объекты, обладающие достаточным количеством одинаковых признаков. Требуется разбить множество исходных объектов на попарно непересекающиеся группы (кластеры) так, чтобы минимизировать число связей между кластерами и число недостающих связей внутри кластеров. Количество кластеров может быть задано, ограничено или заранее не определено. Постановки и различные интерпретации задачи аппроксимации графа можно найти в [5–8].

В первой части настоящей работы рассматриваются три варианта задачи аппроксимации графа, являющейся формализацией задач кластеризации взаимосвязанных объектов. Приводится краткий обзор известных результатов по этим задачам. Во второй части рассматривается новая постановка задачи аппроксимации 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. Структура взаимосвязей задана с помощью неориентированного графа. Задана также выборка из k объектов, каждый из которых принадлежит одному из кластеров. В работе доказано, что рассматриваемая задача является NP-трудной. Для случая $k = 2$ доказана NP-трудность задачи на кубических графах.

Будем рассматривать только *обыкновенные* графы, т. е. графы без петель и кратных ребер. Обыкновенный граф называется *кластерным графом*, если каждая его компонента связности является полным графом [9]. Обозначим через $\mathbf{M}(V)$ множество всех кластерных графов на множестве вершин V, $\mathbf{M}_k(V)$ – множество всех кластерных графов на множестве вершин V, имеющих ровно k непустых компонент связности, $\mathbf{M}_{1,k}(V)$ – множество всех кластерных графов на множестве V, имеющих не более k компонент связности, $2 \le k \le |V|$.

* Работа первого автора поддержана грантом РНФ (проект 15-11-10009).

1. Задачи кластеризации графов

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 ребер в графах G_1 и G_2.

В 60–80-е гг. XX века в литературе изучались следующие три варианта задачи аппроксимации графа, которые можно рассматривать как различные формализации задачи кластеризации взаимосвязанных объектов [6–8; 10; 11]. В дальнейшем задачи аппроксимации графов неоднократно переоткрывались и независимо изучались под разными названиями (**Correlation Clustering** [12], **Cluster Editing** [9; 13]).

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

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

Задача $\mathbf{A_k}$. Дан обыкновенный граф $G = (V, E)$ и целое число k, $2 \le k \le |V|$. Найти такой граф $M^* \in \mathrm{M}_k(V)$, что

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

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

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

Первые теоретические результаты, относящиеся к задачам аппроксимации графов, были получены в 60–70-е гг. XX в. В 1964 г. Заном [8] была решена задача **А** для графов, представляющих 2- и 3-иерархические структуры. В 1971 г. Фридман [6] выделил первый полиномиально разрешимый случай задачи аппроксимации графа **А**. Он показал, что задача **А** для любого графа без треугольников сводится к построению в нем наибольшего паросочетания.

В 1986 г. Крживанек и Моравек [14] доказали, что задача **А** является NP-трудной, однако их работа осталась незамеченной. В 2004 г. Бансал, Блюм и Чаула [12] и независимо Шамир, Шаран и Цур [9] доказали NP-трудность задачи **А**, а Ильев и Талевнин (см. [15]) установили, что взвешенная задача $\mathbf{A_k}$ NP-трудна при любом фиксированном $k \ge 2$. В [9] доказано также, что задача $\mathbf{A_k}$ NP-трудна при любом фиксированном $k \ge 2$; в 2006 г. Гиотис и Гурусвами [16] опубликовали более простое доказательство этого же результата. В том же году независимо Агеев, Ильев, Кононов и Талевнин [17] доказали, что задачи $\mathbf{A_2}$ и $\mathbf{A_{1,2}}$ NP-трудны уже на кубических графах, откуда вывели, что все упомянутые ранее варианты задачи аппроксимации графа являются NP-трудными, включая и задачу $\mathbf{A_{1,k}}$.

В 2004 г. Бансал, Блюм и Чаула [12] предложили 3-приближенный алгоритм для задачи $\mathbf{A_{1,2}}$. В 2006 г. Агеев, Ильев, Кононов и Талевнин [17] доказали существование рандомизированной полиномиальной приближенной схемы для задачи $\mathbf{A_{1,2}}$, а Гиотис и Гурусвами [16] предложили рандомизированную полиномиальную приближенную схему для задачи $\mathbf{A_k}$ (для любого фиксированного $k \ge 2$). В том же году Ильев, Навроцкая и Талевнин [18] показали, что алгоритм локального поиска является гарантированно асимптотически точным для задачи $\mathbf{A_{1,2}}$ на неплотных графах. Указав, что сложность полиномиальной приближенной схемы из [16] лишает ее перспективы практического использования, Коулман, Саундерсон и Вирт [19] в 2008 г. предложили 2-приближенный алгоритм для задачи $\mathbf{A_{1,2}}$, применив процедуру локального поиска к допустимому решению, полученному с помощью 3-приближенного алгоритма из статьи [12]. Для задачи $\mathbf{A_2}$ в работе [20] Ильевым, Ильевой и Навроцкой предложен $(3 - 6/|V|)$-приближенный алгоритм с достижимой гарантированной оценкой точности.

Что касается задачи **А**, то в 2005 г. Чарикар, Гурусвами и Вирт [21] показали, что задача **А** является *APX*-трудной и разработали для нее 4-приближенный алгоритм. В 2008 г. Айлон, Чарикар и Ньюман [22] предложили 2,5-приближенный алгоритм для задачи **А**.

2. Задача кластеризации с частичным обучением

Рассмотрим следующую формализацию задачи кластеризации с частичным обучением.

Задача A_k^+. Дан обыкновенный граф $G = (V, E)$ и целое число k, $2 \le k \le |V|$. Выделено множество попарно различных вершин $X = \{x_1, ..., x_k\} \subseteq V$. Требуется найти такой граф $M^* \in \mathrm{M}_k(V)$, что

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

где минимум берется по всем кластерным графам $M = (V, E_M) \in \mathrm{M}_k(V)$, в которых $x_i x_j \notin E_M$ для любых $i, j \in \{1, ..., k\}$; другими словами, никакие две вершины множества $X = \{x_1, ..., x_k\}$ не принадлежат одной и той же компоненте связности (т. е. одному кластеру) графа M.

Исследуем вычислительную сложность задачи A_k^+. Рассмотрим сначала частный случай задачи, когда $k = 2$. Напомним, что граф называется *кубическим*, если степени всех его вершин равны 3.

Теорема 1. *Задача A_2^+ на кубических графах NP-трудна.*

Доказательство. Как было показано в работе [1], задача $\mathbf{A_2}$ на кубических графах

NP-трудна. Следовательно, для доказательства *NP*-трудности задачи A_2^+ на кубических графах достаточно свести к ней по Тьюрингу задачу $\mathbf{A_2}$ на кубических графах.

Рассмотрим произвольный кубический граф $G = (V, E)$ – вход задачи $\mathbf{A_2}$ – и фиксируем две несовпадающие вершины x_1, x_2 графа G.

Имея оптимальное решение $M^*(x_1, x_2)$ задачи A_2^+ для любой такой пары вершин $\{x_1, x_2\} \in V$ и выбрав среди них ближайший к графу G кластерный граф

$$M^* = \arg\min_{\{x_1,\, x_2\} \in V} \rho(G, M^*(x_1, x_2)),$$

мы, очевидно, получим оптимальное решение исходной задачи $\mathbf{A_2}$. Легко видеть, что построение всех $n(n-1)/2$ входов задачи A_2^+ и получение оптимального решения исходной задачи $\mathbf{A_2}$ можно выполнить за время $O(n^2)$, где $n = |V|$.

Теорема доказана.

Теперь рассмотрим задачу A_k^+ для произвольного фиксированного k. Докажем, что задача A_k^+ является NP-трудной.

Теорема 2. *Задача* A_k^+ *на кубических графах NP-трудна при любом фиксированном* $k \geq 2$.

Доказательство. Для доказательства сведем по Тьюрингу NP-трудную задачу $\mathbf{A_k}$ к A_k^+. Рассмотрим произвольный граф $G = (V, E)$ – вход задачи $\mathbf{A_k}$ – и фиксируем целое число k и произвольный набор $\{x_1, \ldots, x_k\}$, состоящий из k попарно различных вершин графа G.

Имея оптимальное решение $M^*(x_1, ..., x_k)$ задачи A_k^+ для любого такого набора $\{x_1, ..., x_k\} \subseteq V$ и выбрав среди них ближайший к графу G кластерный граф

$$M^* = \arg\min_{\{x_1, \ldots, x_k\} \in V} \rho(G, M^*(x_1, ..., x_k))$$

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

Теорема доказана.

ЛИТЕРАТУРА

[1] *Bair E.* Semi-supervised clustering methods // Wiley Interdisciplinary Reviews: Computational Statistics. 2013. Vol. 5. № 5. P. 349–361.

[2] *Chapelle O., Scholkopf B., Zein A.* Semi-Supervised Learning. MIT Press: Cambridge, Massachusets, 2006.

[3] *Kulis B., Basu S., Dhillon I., Mooney R.* Semi-supervised graph clustering: a kernel approach // Machine Learning. 2009. Vol. 74. № 1. P. 1–22.

[4] *Schaeffer S.E.* Graph clustering // Computer Science Review. 2005. Vol. 1. № 1. P. 27–64.

[5] *Ляпунов А. А.* О строении и эволюции управляющих систем в связи с теорией классификации // Проблемы кибернетики. М.: Наука, 1973. Вып. 27. С. 7–18.

[6] *Фридман Г.Ш.* Одна задача аппроксимации графов // Управляемые системы. 1971. Вып. 8. С. 73–75.

[7] *Tomescu I.* La reduction minimale d'un graphe `a une reunion de cliques // Discrete Math. 1974. Vol. 10. № 1–2. P. 173–179.

[8] *Zahn C.T.* Approximating symmetric relations by equivalence relations // J. Soc. Indust. Appl. Math. 1964. V. 12. № 4. P. 840–847.

[9] *Shamir R., Sharan R., Tsur D.* Cluster graph modification problems // Discrete Appl. Math. 2004. Vol. 144. № 1–2. P. 173–182.

[10] *Ильев В. П., Фридман Г. Ш.* К задаче аппроксимации графами с фиксированным числом компонент // Доклады АН СССР. 1982. Т. 264. № 3. С. 533– 538.

[11] *Фридман Г.Ш.* Исследование одной задачи классификации на графах // Методы моделирования и обработка информации. Новосибирск : Наука, 1976. С. 147–177.

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

[13] *Ben-Dor A., Shamir R., Yakhimi Z.* Clustering gene expression patterns // J. Comput. Biol. 1999. Vol. 6. № 3–4. P. 281–297.

[14] *K˘rivanek M., Mor´avek J.* NP-hard problems in hierarchical-tree clustering // Acta informatica. 1986. Vol. 23. P. 311–323.

[15] *Талевнин А.С.* О сложности задачи аппроксимации графов // Вестник Омского университета. 2004. № 4. С. 22–24.

[16] *Giotis I., Guruswami V.* Correlation clustering with a fixed number of clusters // Theory of Computing. 2006. Vol. 2. № 1. P. 249–266.

[17] *Агеев А. А., Ильев В. П., Кононов А. В., Талевнин А. С.* Вычислительная сложность задачи аппроксимации графов // Дискретный анализ и исследование операций. Серия 1. 2006. Т. 13. № 1. С. 3–11.

[18] *Ильев В. П., Навроцкая А. А., Талевнин А. С.* Полиномиальная приближенная схема для задачи аппроксимации неплотных графов // Вестник Омского университета. 2007. Вып. 4. С. 24–27.

[19] *Coleman T., Saunderson J., Wirth A.* A local-search 2-approximation for 2-correlation clustering // Algorithms – ESA 2008: Lecture Notes in Comput. Sci. 2008. Vol. 5193. P. 308–319.

[20] *Ильев В. П., Ильева С. Д., Навроцкая А. А.* Приближенные алгоритмы для задач аппроксимации графов // Дискретный анализ и исследование операций. 2011. Т. 18. № 1. С. 41–60.

[21] *Charikar M., Guruswami V., Wirth A.* Clustering with qualitative information // J. Comput. Syst. Sci. 2005. Vol. 71. № 3. P. 360–383.

[22] *Ailon N., Charikar M., Newman A.* Aggregating inconsistent information: Ranking and clustering // J. ACM. 2008. V. 55. № 5. P. 1–27.