МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

МАТЕРИАЛЫ
VIII Международной молодежной научной
конференции
«МАТЕМАТИЧЕСКОЕ
И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ
ИНФОРМАЦИОННЫХ,
ТЕХНИЧЕСКИХ
И ЭКОНОМИЧЕСКИХ СИСТЕМ»

Томск, 26–30 мая 2021 г.

Под общей редакцией И.С. Шмырина

Томск
Издательство Томского государственного университета
2021

О ДВУХ ЗАДАЧАХ КЛАСТЕРИЗАЦИИ ЗНАКОВОГО ГРАФА

Ибрагимов Э.И.

Сибирский федеральный университет
ibragimovaei@mail.ru

Введение

В настоящее время широкое применение знаковых графов в социологии, психологии, биоинформатике и кластерном анализе данных вызывает растущий интерес к исследованиям математических моделей проблем знакового баланса и поиска эффективных методов решения этих проблем. В связи с широким применением знаковых графов в социологии, психологии, биоинформатике и кластерном анализе актуальны исследования, направленные на разработку алгоритмов решения задач, связанных с ними. Центральное место в теории знаковых графов занимает поиск эффективных методов решения проблемы знакового баланса и связанных с ней задач [1,2].

Знаковым графом называется пара $\Sigma = (G, \sigma)$, где $G = (V, E)$ – неориентированный граф ($n = |V| \geq 2$, $m = |E| \geq 1$), на ребрах которого задана функция знака $\sigma: E \rightarrow \{+, -\}$. Знаковый граф называется k -кластеризуемым (k -сбалансированным), если множество его вершин можно разбить на k попарно непересекающихся непустых кластеров так, что все положительные рёбра находятся внутри кластеров, а отрицательные между кластерами [3,4]. Задача распознавания k -кластеризуемости знакового графа формулируется следующим образом [5].

Условие. Задан знаковый граф $\Sigma = (G, \sigma)$, где $G = (V, E)$ – простой неориентированный граф, $n = |V| \geq 2$, $m = |E| \geq 1$.

Вопрос. При каких k граф $\Sigma = (G, \sigma)$ является k -кластеризуемым?

Т.к. знаковые графы не всегда являются k -кластеризуемыми, возникает задача поиска такого разбиения на k попарно непересекающихся непустых кластеров, чтобы некоторый функционал ошибки кластеризации был минимальным.

1. Задача k -кластеризации знакового графа

Будем говорить, что система множеств $C_k = \{V_1, V_2, \dots, V_k\}$ образует разбиение множества вершин V на k подмножеств, если подмножества непустые ($V_i \neq \emptyset$, $i = \overline{1, k}$), попарно не пересекаются ($V_i \cap V_j = \emptyset$, $i \neq j$), и их объединение равно

$V: V = V_1 \cup V_2 \cup \dots \cup V_k$. Обозначим $\Phi = \bigcup_{k=1}^n \Phi_k$ – множество всех возможных разбиений

V , где Φ_k – множество возможных разбиений V ровно на k подмножеств. Пусть C_k – какое-либо разбиение графа. Под отрицательной ошибкой $N(C_k)$ будем понимать число отрицательных ребер внутри подмножеств, под положительной $P(C_k)$ – число положительных ребер между подмножествами. Критерием качества разбиения будет являться функция суммарной ошибки – выпуклая комбинация положительной и отрицательной ошибок:

$$E_\alpha(C_k) = \alpha N(C_k) + (1 - \alpha) P(C_k), \quad \alpha \in [0, 1]. \quad (1)$$

Тогда задача k -кластеризации знакового графа может быть сформулирована следующим образом [5].

Условие. Задан знаковый граф $\Sigma = (G, \sigma)$, где $G = (V, E)$ – простой неориентированный граф, $n = |V| \geq 2$, $m = |E| \geq 1$.

Вопрос. Для заданного $\alpha \in [0, 1]$ требуется найти такое разбиение множества вершин V знакового графа Σ , для которого суммарная ошибка минимальна.

Решением задачи будет множество кластеров $C^* = \arg \min_{C \in \Phi} E_\alpha(C)$ и $k = |C^*|$ – количество кластеров. Известно, что задача минимизации суммарной ошибки является NP-полной [2].

В работе исследуются два частных случая этой задачи при $\alpha = 0$ и при $\alpha = 1$.

Случай 1. При $\alpha = 0$ ищется такое разбиение C^* , для которого минимизируется положительная ошибка, т.е.

$$C^* = \arg \min_{C \in \Phi} P(C). \quad (2)$$

Очевидно, что положительная ошибка достигает минимума на разбиениях, для которых $P(C^*) = 0$.

Случай 2. При $\alpha = 1$ ищется такое C^* , для которого минимизируется отрицательная ошибка, т.е.:

$$C^* = \arg \min_{C \in \Phi} N(C). \quad (3)$$

Аналогично, предыдущему случаю отрицательная ошибка достигает минимума на разбиении, для которого $N(C^*) = 0$.

2. Алгоритмы k -кластеризации знакового графа

Для решения задач (2) и (3) разработаны алгоритмы *SGClust0* и *SGClust1* соответственно.

Алгоритм *SGClust0* решает задачу минимизации положительной ошибки (случай 1). В ходе работы данного алгоритма для заданного графа Σ с помощью удаления отрицательных ребер строится граф Σ^+ . Искомое разбиение C^* будет состоять из компонент связности Σ^+ . Доказана теорема 1, устанавливающая вычислительную сложность алгоритма *SGClust0*.

Теорема 1. Алгоритм *SGClust0* находит решение задачи k -кластеризации знакового графа при $\alpha = 0$ за время, не превышающее $O(n + m)$.

Схема работы алгоритма *SGClust1*, решающего задачу минимизации отрицательной ошибки ($\alpha = 1$), приведена на рис. 1.

Рис. 1. Схема работы алгоритма *SGClust1*

В схеме используются следующие обозначения: Σ – исходный граф, Σ^+ – граф, полученный из Σ удалением отрицательных ребер, C – множество компонент связности в Σ^+ . Алгоритм на вход получает граф Σ и находит решение в виде множества кластеров C^* . Алгоритм *SGClust1* состоит из двух основных блоков. В блоке *Initialization* выделяются компоненты связности C в Σ^+ . В блоке *Split* формируется множество кластеров C^* . Функция $\Gamma^-(v)$ вычисляется для каждой вершины v из C и определяет количество инцидентных отрицательных ребер вершине v . Функция *Find*(C) находит вершину $v_{i\max}$ с наибольшим количеством инцидентных отрицательных ребер в компоненте c согласно формуле $v_{i\max} = \arg \max_{v \in C} \Gamma^-(v)$. Функция *Del*($c_{i\max}$) выделяет вершину $v_{i\max}$ в отдельный кластер, т.е. $P = P \cup \{c_{i\max}\}$ и удаляет ее из c . Блок *Split* выполняется пока в компоненте c есть отрицательные ребра. Вычислительная сложность алгоритма *SGClust1* установлена в теореме 2.

Теорема 2. Алгоритм *SGClust1* находит решение задачи k -кластеризации знакового графа при $\alpha = 1$ за время, не превышающее $O(n^2 + nm)$.

Пример. На рис. 2–4 приведены примеры работы алгоритмов *SGClust0* и *SGClust1*. На рис. 2 изображен исходный граф Σ . Результат работы алгоритма *SGClust0* приведен на рис. 3. Получены три кластера между которыми нет положительных ребер, но внутри кластера $\{6,7,8,9\}$ есть отрицательное ребро, что является допустимым. На рис. 4 изображено разбиение на 4 кластера, полученное с помощью алгоритма *SGClust1*. Т.к. наличие отрицательных ребер внутри кластера не является допустимым, вершина 7 выделена в отдельный кластер.

Рис. 2. Граф Σ

Рис. 3. Результат работы алгоритма *SGClust0*

Рис. 4. Результат работы алгоритма *SGClust1*

3. Вычислительные эксперименты

Для оценки эффективности алгоритмов были проведены вычислительные эксперименты, результаты которых приведены в табл. 1 и 2. Для проверки алгоритма *SGClust0* использовались графы, для которых было известно k , т.е. количество кластеров, для которого граф k -сбалансирован. Алгоритм *SGClust1* проверялся на графах, которые не являются k -кластеризуемым. Эксперименты проводились на ПК с процессором Intel® Core™ i7-10510U CPU @ 1.80ГГц и оперативной памятью объемом 8ГБ.

Таблица 1

Алгоритм *SGClust0*

Число вершин	Число ребер	Сбалансированность	Количество кластеров	Время работы
9	9	1- и 3-сбалансирован	1, 2 или 3	менее 1 мс
25	50	5-сбалансирован	5	менее 1 мс
36	90	несбалансирован	2	1 мс
30	30	1- или 10-сбалансирован	1, 2, ..., 10	1 мс
30	435	10- сбалансирован	10	1 мс
30	435	несбалансирован	10	1 мс

Таблица 2

Алгоритм *SGClust1*

Число вершин	Число ребер	Количество кластеров	Время работы
4	4	2	менее 1 мс
5	10	5	менее 1 мс
6	6	3	менее 1 мс
7	8	3	менее 1 мс
10	45	5	1 мс

Результаты вычислительных экспериментов показывают, что алгоритмы *SGClust0* и *SGClust1* работают корректно.

Заключение

В работе исследуются два частных случая задачи кластеризации знакового графа. Разработаны алгоритмы *SGClust0* для кластеризации в случае минимизации положительной ошибки и *SGClust1* для кластеризации в случае минимизации отрицательной ошибки. Приведены теоремы, показывающие, что разработанные алгоритмы находят решение за полиномиальное время.

Проведенные вычислительные эксперименты подтвердили эффективность разработанных алгоритмов. Перспективным направлением исследований является разработка алгоритма для решения задачи при произвольных $0 < \alpha < 1$.

ЛИТЕРАТУРА

1. Kartwright D., Harary F. Structural balance: a generalization of Heider's theory // The Psychological Review. – 1956. – V.63. – No.5. – P. 277–293.
2. Nikhil Bansal, Avrim Blum, Shuchi Chawla Correlation clustering // Machine Learning. – 2004. – V.56. – P. 89–113.
3. Doreian P., Mrvar A. Structural balance and partitioning signed graphs // In A. Ferligoj & A. Kramberger (Eds.), Developments in data analysis (Metodoloskizvezki 12, pp. 195-208). Ljubljana, Slovenia: FDV.
4. Zaslavsky T. Balance and clustering in signed graphs // Unpublished manuscript, Binghamton University, New York, 2010.
5. Ибрагимова Э.И., Семенова Д.В. Распознавание k -кластеризуемости знаковых графов // Информационные технологии и математическое моделирование (ИТММ-2020). Материалы XIX Международной конференции имени А.Ф. Терпугова. – 2020. – С. 468–472.