РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ

А В Т О М Е Т Р И Я

2005, том 41, № 5

УДК 547 : 541. 6 : 51-7

Л. И. Макаров
(Новосибирск)

ОСОБЫЕ ВЕРШИНЫ ВЗВЕШЕННОГО ГРАФА ВЫБОРКИ*

Рассмотрены особые объекты из множества (выборки), представленного в виде полного взвешенного графа с заданной матрицей расстояний между его вершинами. Особые объекты (вершины) определяются по экстремальным значениям критериев, зависящих от длин ребер графа.

Введение. При исследовании некоторого множества (выборки) объектов часто используется его представление в виде взвешенного графа. При этом вершины графа соответствуют объектам, а для каждого ребра указана его длина (вес) – расстояние между соответствующими объектами.

В качестве расстояния обычно используют количественную оценку степени различия объектов. Например, расстояние между объектами, имеющими векторное описание в системе количественных признаков, задают как расстояние между их векторами. Для объектов, имеющих внутреннюю структуру, используют графовые модели описания. Так, моделью структуры химического соединения служит молекулярный граф, вершины которого соответствуют атомам или группам атомов соединения, а ребра – химическим связям между его атомами. В этом случае расстояние между объектами зависит от величины их наибольшего общего структурного фрагмента. Чем больше такой фрагмент, тем больше структурное подобие объектов и меньше расстояние между ними. Графовые модели химических структур применяются при исследовании взаимосвязи структур и свойств химических соединений, изучении строения соединений по их инфракрасным (ИК) спектрам [1–6] и т. д.

Способ выбора подмножества (выборки) из некоторого множества объектов зависит от специфики задачи. Для исследования взаимосвязи структур и свойств соединений в выборку включаются соединения, обладающие заданным свойством. При изучении строения соединений выборку формируют из соединений, имеющих ИК-спектры, близкие к ИК-спектру исследуемого соединения.

При описании выборки объектов в виде взвешенного графа естественно возникает задача нахождения его особых вершин, которые обладают близки-

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 05-01-00816).

ми к экстремальным значениями некоторых критериев – функций, зависящих от длин ребер графа.

Для каждой вершины графа используются три критерия, основанные на:

1) минимальном или максимальном значении длин инцидентных ей ребер;

2) сумме длин всех этих ребер;

3) степени вершины после удаления из графа ребер с длиной, большей заданного порога.

Эти критерии позволяют формализовать некоторые интуитивные понятия, связанные с расстояниями между объектами выборки, например «центр и периферия выборки», «компактные группы объектов», «граничные объекты» и т. д. Если выборка разбита на таксоны, то особые объекты можно определять в каждом таксоне и, кроме того, определять взаимное расположение таксонов по расстояниям между их объектами.

Особые вершины взвешенного графа. Пусть $G(V, X)$ – полный неориентированный граф без петель и кратных ребер, имеющий множество вершин $V \quad \{v_i\}$, $|V| \quad p$, и множество ребер $X \quad \{x_{ij}\}$, $x_{ij} \quad (v_i, v_j)$, $v_i, v_j \quad V$, $i, j \quad \{1,2,\ldots, p\}$. Граф $G(V, X)$ называют взвешенным, если каждому его ребру x_{ij} приписано положительное действительное число $c_{ij} \quad c(v_i, v_j) \quad 0$ – длина ребра, и задают матрицей весов $\|c_{ij}\|$ порядка p, где $c_{ii} \quad 0$.

Для оценки структурного подобия химических соединений используют разные функции расстояний между их молекулярными графами, например функцию $f_{kl} \quad p_k \quad p_l \quad 2p_{kl}$, где p_k и p_l – количества вершин k- и l-го молекулярных графов, а p_{kl} – количество вершин их наибольшего общего подграфа [6]: $0 \quad p_{kl} \quad \min(p_k, p_l)$.

Особые вершины графа по критерию 1.

Обозначения:

$c_i^x \quad \max\limits_j c_{ij}$ – длина наибольшего ребра, инцидентного i-й вершине;

$c_i^n \quad \min\limits_j c_{ij}$ – длина наименьшего ребра, инцидентного i-й вершине;

$c^x \quad \sum\limits_i c_i^x \Big/ p$ – средняя длина наибольших ребер вершин;

$c^n \quad \sum\limits_i c_i^n \Big/ p$ – средняя длина наименьших ребер вершин;

$c \quad \sum\limits_i \sum\limits_j c_{ij} \Big/ (p(p \quad 1))$ – средняя длина ребер графа;

$c \quad \min\limits_i c_i^x$, $c \quad \max\limits_i c_i^x$.

Определения особых вершин.

Центр графа – множество вершин $V \quad \{v_i\}$, имеющих $c_i^x \quad c$ [7].

Ядро графа – множество V^C вершин v_i, имеющих $c_i^x \quad l \quad (c^x \quad c)/2$.

Периферия графа – множество V^P вершин v_i, для которых $c_i^x \quad l \quad (c^x \quad c)/2$.

Слой графа – множество V^L вершин v_i, для которых выполняется условие $l \quad c_i^x \quad l$.

Край графа – множество V^E вершин v_i, для которых выполняются условия $c_i^x \quad l$ и $\min_j c(v_i, v_j) \quad c$. Вершинами Края являются вершины Периферии, удаленные от каждого из центров v_i на расстояние, превышающее c.

Опора графа – множество V^S вершин v_i, длины ребер которых удовлетворяют условию $c^n \quad c_{ij} \quad c^x$ для всех j.

Очевидно, что $V \quad V^C, V^E \quad V^P, V^S \quad V^P \quad 0$.

В частных случаях некоторые множества особых вершин могут совпадать. Например, если взвешенный граф $G(V, X)$ можно представить на евклидовой плоскости в виде правильного многоугольника, то для всех вершин выполняются равенства $c_i^x \quad c_j^x \quad c^x \quad c \quad c$, $c_i^n \quad c_j^n \quad c^n$; $i, j \quad \{1,2,\ldots, p\}$, и поэтому $V \quad V^C \quad V^P \quad V, V^L \quad V^E \quad V^S \quad 0$.

Эти же соотношения выполняются и в случае полного взвешенного графа с ребрами одинаковой длины. Его можно считать представлением обыкновенного полного графа.

Интерпретация множеств особых вершин может быть следующей.

Центры графа выборки соответствуют объектам, наибольшая непохожесть которых на другие объекты минимальна. Традиционно такие объекты называют типичными или эталонами выборки.

Диаметральные вершины v_i, имеющие $c_i^x \quad c$, соответствуют объектам, входящим в пары максимально непохожих друг на друга объектов выборки, $v_i \quad V \quad V^P$.

Вершины Периферии соответствуют таким объектам выборки, для которых в ней имеются объекты, мало похожие на них.

Вершины Опоры соответствуют объектам, которые в достаточной мере похожи на все объекты выборки.

Особые вершины графа по критерию 2.

Обозначения:

$D_i \quad \sum_j c_{ij}$ – дистанция i-й вершины;

$D \quad \sum_i D_i \Big/ p$ – средняя дистанция вершин графа;

$D^n \quad \min_i D_i$ – минимальная дистанция вершин в графе;

$D^x \quad \max_i D_i$ – максимальная дистанция вершин в графе.

Определения особых вершин.

Медиана графа – множество V^M вершин v_i, имеющих дистанцию $D_i \quad D^n$ [7].

Д-Ядро графа – множество V^K вершин v_i, имеющих дистанцию $D_i \quad (D \quad D^n)/2$.

Д-Периферия графа – множество V^O вершин v_i, имеющих диапазон $D_i \quad (D \quad D^x)/2$.

Д-Слой графа – множество V^D вершин v_i, для которых выполняется условие $(D \quad D^n)/2 \quad D_i \quad (D \quad D^x)/2$.

Каждая вершина Медианы графа выборки соответствует объекту, в среднем наиболее похожему на все остальные объекты, так как средняя длина ее ребра минимальна.

Вершины Д-Периферии в среднем более удалены от всех остальных вершин, чем вершины Д-Ядра и Д-Слоя.

Вершины Д-Ядра в среднем менее удалены от всех остальных вершин, чем вершины Д-Слоя и Д-Периферии.

Очевидно, что $V^M \quad V^K$, $V^D \quad V^O \quad 0$, $V^K \quad V^O \quad 0$, $V^K \quad V^D \quad 0$.

Особые вершины графа по критерию 3.

Обозначения:

d_i – число вершин v_j, соединенных с вершиной v_i ребрами длиной c_{ij} $(c^n \quad \min_i c_i^n)/2$;

$d \quad \sum_i d_i \Big/ p$ – среднее значение d_i.

Определение особых вершин.

Компакты вершин графа – множество V^T вершин v_i, имеющих $d_i \quad d$.

Вершины Компактов имеют достаточно много близких к ним вершин. Это соответствует наличию в выборке групп очень похожих друг на друга объектов, что может характеризовать неоднородность распределения объектов в выборке.

Очевидно, что $V^T \quad V^S \quad 0$, т. е. объекты Опоры достаточно похожи на все объекты выборки, но в выборке нет объектов очень похожих на них.

Представители множеств особых вершин. Во множествах особых вершин графа можно выделить представителей этих множеств, т. е. вершин, расстояние между которыми превышает значение некоторого заданного порога, например с – средней длины ребер графа.

Поиск представителей множеств особых вершин графа можно осуществить некоторым простым приближенным методом, например «жадным» алгоритмом [8]. Рассмотрим схему такого алгоритма для нахождения представителей любого из множеств вершин Ядра, Периферии или Слоя взвешенного графа.

Пусть символ $Z \quad \{C, P, L\}$. Тогда V^Z при выбранном Z обозначает соответствующее множество особых вершин. Пусть h^Z – величина, которая принимает значения: $h^C \quad c$, $h^P \quad c$, $h^L \quad (l \quad l)/2$. Тогда схема алгоритма при заданном значении Z и $V^Z \quad 0$ имеет следующий вид.

Шаг n $(n \quad 1,2,\ldots)$.

Во множестве вершин $V(n)$, $V(1) \quad V^Z$, находится очередной представитель V^Z – вершина $v_i^Z(n) \quad V(n)$ – одна из тех вершин, для которых достигается $\min_i \left| c_i^x \quad h^Z \right|$.

Очевидно, что $v_i^C(1) \quad V$, $v_i^P(1) \quad V$.

Далее определяется множество $V(n \quad 1) \quad V(n) \setminus V_n$, где V_n – множество вершин $v_j \quad V(n)$, для которых выполняется неравенство $c(v_j, v_i^Z(n)) \quad c$.

Если $V(n \quad 1) \quad 0$, то переход к $(n \quad 1)$-му шагу, иначе конец.

Представителей других множеств особых вершин можно установить с помощью аналогичных алгоритмов.

Особые вершины таксона графа относительно других таксонов. Если для взвешенного полного графа $G(V, X)$ проведена таксономия (разбиение) [4, 5] его вершин на подмножества (таксоны) V_k, $k = 1, 2, \ldots, K$, то в каждом таксоне можно выделить особые вершины относительно вершин других таксонов.

Граница (вершины, близкие к другим таксонам) V_k^N k-го таксона – множество вершин v_i, имеющих $\min_i \min_j c_{ij}$, $v_i \in V_k$, $v_j \in (V \setminus V_k)$.

Провинция (вершины, удаленные от других таксонов) V_k^R k-го таксона – множество вершин v_i, имеющих $\max_i \min_j c_{ij}$, $v_i \in V_k$, $v_j \in (V \setminus V_k)$.

Вершины Границы таксона V_k соответствуют его объектам, которые наиболее похожи на некоторые объекты других таксонов. Вершины Провинции таксона V_k соответствуют его объектам, наименее похожим на объекты других таксонов.

Расстояния между таксонами. Пусть для полного взвешенного графа $G(V, X)$ произведено разбиение на таксоны V_k, $k = 1, 2, \ldots, K$. Тогда можно построить взвешенный граф таксонов $G^*(V^*, X^*)$, вершины которого соответствуют таксонам, а каждому ребру приписана величина расстояния между соответствующей парой таксонов. Расстояние $r(kt)$ между парой таксонов V_k, $|V_k| = p_k$, и V_t, $|V_t| = p_t$, взвешенного графа $G^*(V^*, X^*)$ может быть определено разными способами (везде $c_{ij} = c(v_i, v_j)$, $v_i \in V_k$, $v_j \in V_t$).

1. Расстояние $r_1(kt) = \min_i \min_j c_{ij}$ – наименьшее расстояние между вершинами таксонов.

2. Расстояние $r_2(kt) = \sum_i \sum_j c_{ij} \Big/ (p_k p_t)$ – средняя длина всех ребер, соединяющих все вершины таксонов.

3. Расстояние $r_3(kt) = \sum_i \min_j c_{ij} + \sum_j \min_i c_{ji} \Big/ (p_k + p_t)$ – средняя длина кратчайших ребер, соединяющих все вершины каждого таксона с вершинами другого.

4. Расстояние $r_4(kt) = c(v_i^k, v_j^t)$ – расстояние между фиксированными вершинами (например, центрами или медианами) таксонов $v_i^k \in V_k$ и $v_j^t \in V_t$.

Легко видеть, что если для всех трех циклов в графе G выполняются все условия метрики, то в графе G^* для расстояний $r_1(kt)$, $r_2(kt)$, $r_3(kt)$ может не выполняться неравенство треугольника. Таким образом, эти расстояния не являются метриками.

Взвешенный граф таксонов может быть использован с целью укрупнения таксонов в итеративном процессе таксономии и поиска особых объектов выборки. При этом таксоны, полученные на очередном шаге процесса, рассматриваются как объекты выборки для следующего шага.

Заключение. В исследуемых выборках возможно присутствие «сомнительных» объектов: объектов с недостоверными значениями характеристик, ошибочно включенных в выборку, объектов с уникальными свойствами и т. д. Можно предположить, что вершины, соответствующие таким объектам во взвешенном графе выборки, связаны с другими вершинами (объектами) ребрами относительно большей длины, чем средняя длина ребер графа.

Поэтому можно считать, что сомнительные объекты соответствуют Периферии, Краю и Д-Периферии графа выборки.

Установление сомнительных вершин взвешенного графа выборки позволяет провести коррекцию состава выборки с целью повышения ее репрезентативности. Анализируя состав выборки, можно выделить типичные объекты, которые достаточно похожи на все остальные или на некоторые большие подмножества объектов. К таким объектам выборки можно отнести представителей ее Ядра, Д-Ядра, Компактов и для каждого таксона объекты его Провинции.

Информация об особых вершинах графа выборки может быть использована в прикладных исследованиях, например, для изучения неизвестной структуры химического соединения по его инфракрасному спектру.

Проблема состоит в нахождении фрагментного состава молекулы исследуемого соединения по его спектральным характеристикам, в поиске существенных фрагментов, определяющих особенности спектра, и установлении структуры его молекулы исходя из полученных фрагментов [2–4].

Такие исследования проводятся в Новосибирском институте органической химии СО РАН, имеющем базу данных о спектрах и структурах более чем 30000 химических соединений [2]. В базе данных Института применяется бинарное векторное описание молекулярных графов. Каждая компонента вектора графа соответствует определенному фрагменту, а ее значение равно единице, если этот фрагмент входит в граф, и равно нулю, если фрагмент в молекулярном графе не содержится. В качестве используемого набора фрагментов выбирается множество всех ограниченных заданным размером фрагментов, входящих в молекулярные графы соединений базы данных.

Схему методики исследования, использующую данные об особых вершинах графа выборки, можно представить в следующем виде.

1. По полученному исходному ИК-спектру исследуемого соединения, имеющего неизвестную структуру, в базе данных находится выборка порядка 30–50 соединений, имеющих спектр, близкий к исходному. Для структур соединений этой выборки находятся их попарные расстояния и строится ее взвешенный граф. Если необходимо, производится экспертная коррекция выборки с помощью удаления из нее сомнительных соединений, соответствующих вершинам Периферии, Краю и Д-Периферии взвешенного графа выборки.

2. В модифицированном графе выборки находятся представители множеств особых вершин (соединений). По известному фрагментному составу каждого соединения-представителя определяется их суммарный фрагментный состав. Эти фрагменты используются при компьютерной генерации соединений, структурно-близких к исследуемому соединению.

3. Производится компьютерная генерация множества структур соединений, которые потенциально могут быть созданы с использованием фрагментов выборки. Для этого множества сгенерированных структур строится взвешенный граф, в котором определяются представители множеств особых вершин (структур). Полученный набор структур-представителей упорядочивается по степени возможной близости к исследуемой структуре. Из упорядоченного набора структур-представителей эксперт выбирает наиболее предпочтительные для их химического синтеза.

Использование особых вершин графа выборки при решении этой проблемы позволит снизить количество генерируемых структур соединений,

предлагаемых для синтеза, и тем самым значительно сократить время и трудоемкость установления структуры соединения по его ИК-спектру.

СПИСОК ЛИТЕРАТУРЫ

1. Стьюпер Э., Брюгге У., Джурс П. Машинный анализ связи химической структуры и биологической активности. М.: Мир, 1982.
2. Пиоттух-Пелецкий В. Н., Дерендяев Б. Г., Молодцов С. Г., Богданова Е. Ф. Полные наборы фрагментных составов структур при интерпретации ИК-спектров с помощью поисковой системы. Ч. 4. Формирование наиболее вероятной гипотезы о строении изучаемого соединения // Журнал структурной химии. 1997. 38, № 4. С. 785.
3. Varmuza K., Penchev P. N., Scsibrany H. Maximum common substructures of organic compounds exhibiting similar infrared spectra // Journ. Chem. Inf. Comput. Sci. 1998. 38, N 3. P. 420.
4. Дерендяев Б. Г., Макаров Л. И., Богданова Т. Ф., Пиоттух-Пелецкий В. Н. Таксоно- мия структур соединений, отобранных из базы данных по ИК-спектроскопии // Журнал структурной химии. 2001. 42, № 2. С. 325.
5. Макаров Л. И. Отделимость вершин взвешенного графа и алгоритмы их классификации // Автометрия. 1997. № 5. С 81.
6. Макаров Л. И. Методика и алгоритмы прогноза свойств химических соединений по общим фрагментам молекулярных графов // Журнал структурной химии. 1998. 39, № 1. С. 115.
7. Кристофидес Н. Теория графов. Алгоритмический подход. М.: Мир, 1978.
8. Рейнгольд Э., Нивергельд Ю., Део Н. Комбинаторные алгоритмы. Теория и практика. М.: Мир, 1980.

Институт математики им. С. Л. Соболева СО РАН,
E-mail: makarov@math.nsc.ru

Поступила в редакцию
18 января 2005 г.