

ПРИКЛАДНАЯ ТЕОРИЯ ГРАФОВ

DOI 10.17223/20710410/24/6

УДК 519.172.1+517.553

АНАЛИТИЧЕСКАЯ СЛОЖНОСТЬ КЛАСТЕРНЫХ ДЕРЕВЬЕВ¹

А. И. Нормов, Т. М. Садыков

*Российский экономический университет им. Г. В. Плеханова, г. Москва, Россия***E-mail:** SadykovTM@rsute.ru

Вводится понятие аналитической сложности двоичного дерева — его неотрицательной целочисленной характеристики, отражающей комбинаторную структуру дерева и свойства его «наиболее экономичного» аналитического представления. Изучаются свойства аналитической сложности дерева и предлагается алгоритм её вычисления. Развита в работе методика применяется для сравнения кластерных деревьев.

Ключевые слова: кластерный анализ, двоичное дерево, аналитическая сложность.

Введение

Анализ данных большого объёма требует, как правило, их систематизации и классификации. Формируемые классы (кластеры) объектов многочисленной генеральной совокупности данных, подлежащей изучению, могут сравниваться друг с другом путём их отождествления с элементами некоторой комбинаторной структуры (графа, дерева, матрицы и пр.). Эта группа методов анализа данных включает в себя, в числе прочих, дискриминантный анализ, кластеризацию и факторный анализ данных [1]. При этом возникает задача сравнения свойств и характеристик комбинаторных структур, сопоставляемых большим совокупностям исследуемых данных при помощи перечисленных методов. В частности, актуальной является задача сравнения структур кластерных деревьев, которые строятся при помощи кластерного анализа [1, 2].

В настоящей работе вводится понятие аналитической сложности двоичного дерева. Аналитическая сложность функции $f(x, y)$ двух переменных — это её числовая характеристика со значениями в $\mathbb{N} \cup \{\infty\}$, содержащая информацию о возможных представлениях $f(x, y)$ в виде суперпозиций функций одной переменной и фиксированной функции $s(x, y)$ двух переменных, в качестве которой всюду в дальнейшем используется сумма аргументов: $s(x, y) = x + y$ [3]. В работе изучаются свойства аналитической сложности двоичных деревьев и предлагается алгоритм её вычисления. Под аналитической сложностью двоичного дерева понимается аналитическая сложность функции двух переменных, сопоставленной этому дереву некоторым каноническим образом. Полученные результаты применяются для анализа и сравнения структур кластерных деревьев.

¹Второй автор поддержан грантами РФФИ №13-01-12417-офи-м2 и 14-01-00544-а и грантом Правительства РФ для проведения исследований под руководством ведущих ученых в Сибирском федеральном университете (договор №14.Y26.31.0006).

1. Предварительные сведения

Следуя принятой в теории графов терминологии [4], под *степенью* вершины графа будем понимать количество рёбер, инцидентных этой вершине. *Степенью графа* (в частности, дерева) называется наибольшая из степеней его вершин. Дерево называется *двоичным*, если это упорядоченное дерево, каждый узел которого имеет не более двух потомков. Под числом *уровней* в дереве с заданным корнем будем понимать количество рёбер, связывающих корень с максимально удалённой от него вершиной.

Типичной задачей анализа данных (вычисляемых или полученных в результате измерений) является задача построения так называемого кластерного дерева по набору данных об изучаемой совокупности объектов и по заданной метрике (расстоянию) в пространстве их параметров [1]. Алгоритмы кластеризации широко известны и реализованы в многочисленных компьютерных системах статистического анализа данных (например, в системе *Statistica 12*). Если измерения выполнены с достаточно высокой точностью, а используемая метрика позволяет различать все параметры рассматриваемых объектов, то построенное по этим данным кластерное дерево будет двоичным. Действительно, повышение точности измерений соответствует малому возмущению набора данных, содержащего информацию об изучаемой совокупности объектов, а значит, и расстояний между ними в смысле используемой метрики. По теореме Сарда [5, § 2] множество конфигураций, три или более элемента которых лежат на одинаковом расстоянии друг от друга (в смысле произвольной фиксированной достаточно гладкой функции расстояния), имеет нулевую меру. Это и означает, что соответствующее данной конфигурации кластерное дерево будет двоичным. Преобразование кластерного дерева, соответствующее повышению точности измерений и отображающее вершину высокой степени в набор вершин степени 3, иллюстрируется рис. 1.

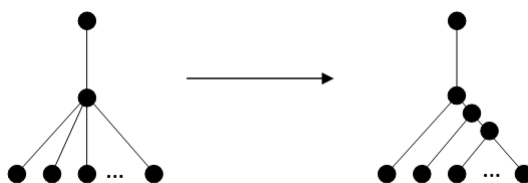


Рис. 1. Преобразование дерева степени $n > 3$ с двумя уровнями в двоичное дерево с $n - 1$ уровнями

Результат кластеризации существенным образом зависит от метрики (т. е. расстояния в пространстве параметров изучаемых объектов), с использованием которой сравниваются объекты кластеризуемого множества. При этом от выбора метрики зависит как геометрическая, так и комбинаторная структура кластерного дерева. Характер этой зависимости иллюстрируется следующим примером.

Пример 1. Простейшим случаем, в котором при использовании разных метрик получаются двоичные деревья с разной комбинаторной структурой, является случай, когда кластеризуемое множество состоит из четырёх элементов, каждый из которых характеризуется двумя числовыми параметрами. Рассмотрим, например, множество точек плоскости $M = \{(0, 0), (4, 0), (6, 3), (8, 3)\}$ (рис. 2).

Кластеризация множества M с использованием стандартного евклидова расстояния и метрики Манхэттена (в которой расстояние между точками плоскости с евклидовыми координатами (x_1, x_2) и (y_1, y_2) равно $|x_1 - y_1| + |x_2 - y_2|$) даёт кластерные деревья, изображённые на рис. 3.

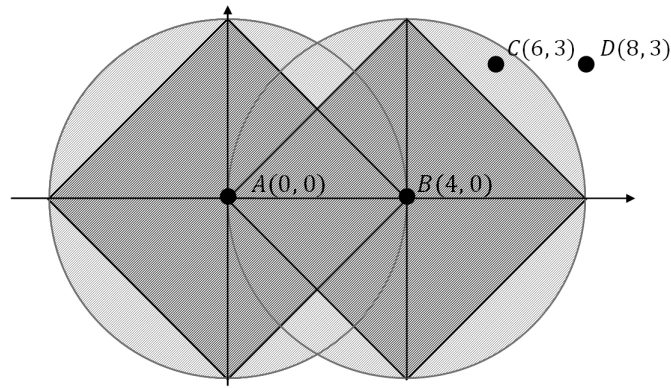


Рис. 2. Конфигурация точек плоскости, дающая кластерные деревья с различной комбинаторной структурой в зависимости от выбора метрики

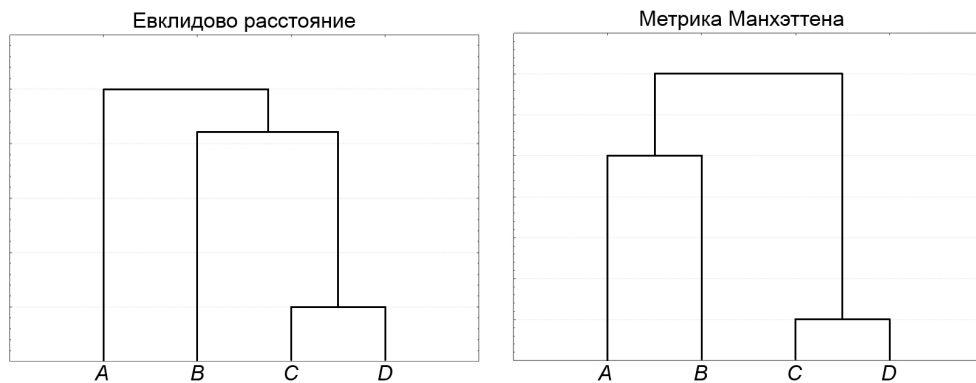


Рис. 3. Кластерные деревья, построенные с использованием разных метрик по одному и тому же набору исходных данных

Причина зависимости комбинаторной структуры кластерного дерева от выбора метрики в пространстве параметров иллюстрируется рис. 2. В то время как в метрике Манхэттена точки множества M группируются в пары похожих элементов, в евклидовой метрике каждая последующая точка лежит на всё большем расстоянии от кластера, образованного всеми предыдущими точками.

Всюду в дальнейшем предполагается, что сравнение объектов кластеризуемого множества производится с использованием произвольной, но фиксированной метрики в пространстве их параметров.

2. Аналитическая сложность функции двух переменных

Аналитическая сложность голоморфной функции $f(x, y)$ двух комплексных переменных — это её числовая характеристика со значениями в $\mathbb{N} \cup \{\infty\}$, содержащая информацию о «наиболее экономичном» представлении $f(x, y)$ в виде суперпозиций функций одной комплексной переменной и фиксированной функции $s(x, y)$ двух переменных. В качестве последней всюду в дальнейшем используется сумма аргументов: $s(x, y) = x + y$. В работе [3] предложена иерархия классов аналитической сложности $Cl_0 \subset Cl_1 \subset \dots \subset Cl_n \subset \dots \subset Cl_\infty$ для функций двух комплексных переменных. Согласно этой иерархии, тривиальный класс функций нулевой аналитической сложности образуют функции, которые зависят лишь от одной из переменных или являются постоянными. В качестве «материала» для построения функций более высокой (т. е. ненулевой) аналитической сложности используются функции нулевого класса и функ-

ция $s(x, y)$, всевозможные конечные композиции которых и образуют множество функций конечной аналитической сложности. Стратификация этого множества осуществляется путём учёта числа применений функции $s(x, y)$, необходимого для образования нужной композиции. Будем исходить из следующего индуктивного определения.

Определение 1 [3]. Нулевой класс аналитической сложности Cl_0 состоит из функций, зависящих не более чем от одной из своих переменных. Функция $f(x, y)$ лежит в множестве Cl_n функций аналитической сложности $n > 0$, если для некоторой точки $(x_0, y_0) \in \mathbb{C}^2$ и некоторого ростка $\mathfrak{f}(x, y) \in \mathcal{O}(U(x_0, y_0))$ этой функции, голоморфного в окрестности этой точки, найдутся ростки голоморфных функций $a, b \in Cl_{n-1}$ и $c \in Cl_0$, такие, что $\mathfrak{f}(x, y) = c(a(x, y) + b(x, y))$. Если такое представление невозможно ни для какого конечного n , то функция f имеет бесконечную аналитическую сложность.

Данное определение является корректным в том смысле, что аналитическая сложность голоморфной функции двух переменных не зависит от ростка функции, для которого ищется наиболее экономичное представление из определения 1. Это следует из того, что для каждого класса аналитической сложности существует дифференциальный критерий принадлежности функции этому классу, отождествляющий множество функций заданной аналитической сложности с множеством всех решений системы дифференциальных многочленов с целыми коэффициентами. В силу принципа консерватизма аналитическое продолжение решения такой системы уравнений вдоль любого пути также должно быть её решением и, следовательно, иметь ту же аналитическую сложность. Аналитическую функцию $f(x, y)$, $x, y \in \mathbb{C}$, назовём функцией аналитической сложности n , если $f(x, y) \in Cl_n$, но $f(x, y) \notin Cl_{n-1}$. Всюду в дальнейшем под сложностью функции двух переменных будем понимать её аналитическую сложность.

Для Cl_1 известен дифференциальный критерий принадлежности функции этому классу [3]: аналитическая функция $f(x, y)$ лежит в Cl_1 в том и только в том случае, когда

$$f'_x(f'_y)^2 f'''_{xxy} - (f'_x)^2 f'_y f'''_{xxy} + f''_{xy}(f'_x)^2 f''_{yy} - f''_{xy}(f'_y)^2 f''_{xx} = 0. \quad (1)$$

Получить это уравнение можно следующим образом: продифференцируем функцию $f(x, y) = c(a(x) + b(y))$ по переменным x и y :

$$f'_x = a'(x) c'(a(x) + b(y)), \quad f'_y = b'(y) c'(a(x) + b(y)).$$

Отсюда видно, что функция $\ln(f'_x/f'_y)$ представима в виде суммы функции переменной x и функции переменной y , а значит, имеет место равенство $(\ln(f'_x/f'_y))''_{xy} = 0$. Вычисляя производные в левой части, заключаем, что числитель полученного выражения равен левой части уравнения (1). Подобное уравнение существует для любого Cl_n , но уже для Cl_2 его вычисление является нетривиальной (и на данный момент нерешённой) задачей [6].

Пример 2. Приведём важные для дальнейшего примеры функций двух переменных конечной аналитической сложности.

а) Функция x^y лежит в первом классе аналитической сложности. Это следует из представления $x^y = \exp(\exp(\ln \ln x + \ln y))$ и того обстоятельства, что данная функция существенно зависит от обеих переменных. Другой способ убедиться в том, что $x^y \in Cl_1$, состоит в применении дифференциального критерия (1): подстановка x^y вместо f даёт тождественный нуль.

б) Согласно «а» и определению 1, для произвольной функции одной переменной $h(t)$ композиция $h(x^y)$ также лежит в Cl_1 .

в) Аналогичные рассуждения показывают, что аналитическая сложность функции $x^{(y^x)}$ равна 2. Действительно, применение дифференциального критерия (1) даёт $x^{(y^x)} \notin Cl_1$. Представление, аналогичное приведённому в примере «а», позволяет убедиться в том, что аналитическая сложность функции $x^{(y^x)}$ не превосходит двух.

3. Аналитическая сложность двоичного дерева и алгоритм её вычисления

Опишем процедуру сопоставления двоичному дереву неотрицательного целого числа — его аналитической сложности. Эта величина является классификационным параметром кластерного дерева (ср. с [7]) и играет в дальнейшем важную роль. С этой целью заданному двоичному дереву T некоторым каноническим способом сопоставляется аналитическая функция двух переменных $E_T(x, y)$. Аналитической сложностью двоичного дерева T по определению называется аналитическая сложность функции $E_T(x, y)$. Как мы увидим, эта величина корректно определена (и при этом конечна), т. е. зависит лишь от самого двоичного дерева T . Тем самым любому двоичному дереву сопоставлена целочисленная неотрицательная характеристика, которая используется, в частности, для классификации кластерных деревьев.

Будем рассматривать аналитические функции с двумя независимыми переменными x, y . Отметим, что если функция E допускает представление в виде $E = \tilde{C}(A_{m-1}(x, y) + B_{n-1}(x, y))$, где A_{m-1} — функция сложности $m - 1$, B_{n-1} — функция сложности $n - 1$, а $\tilde{C}$ — произвольная непостоянная функция одной переменной, то сложность E равна $\max(m, n)$.

Алгоритм 1. Построение аналитической функции конечной сложности по двоичному дереву

Зафиксируем вершину двоичного дерева степени не выше 2, которую будем считать корнем. Для сопоставления дереву аналитической функции двух переменных присвоим каждой вершине значение в зависимости от её степени в соответствии со следующими правилами:

- 1) вершинам степени 1, за исключением корня дерева, сопоставляем последовательно значение x или y , чередуя эти переменные;
- 2) вершинам степени 2, за исключением корня, сопоставляем произвольную функцию h одной переменной;
- 3) вершинам степени 3 сопоставляем функцию возведения в степень « $\wedge$ »;
- 4) корню дерева степени 1 сопоставляем произвольную функцию h , а корню степени 2 — функцию « $\wedge$ ».

После сопоставления значений вершинам необходимо выполнить «сборку» функции. Она начинается с корня дерева и выполняется следующим образом:

- 1) если корню сопоставлено значение « $\wedge$ », то ему соответствует функция $E_1^{E_2}$, где E_1, E_2 — функции левого и правого потомков соответственно;
- 2) если корню присвоено значение h , то ему соответствует функция $h(E)$, где E — функция потомка;
- 3) далее по аналогии применяем последовательно данную операцию ко всем поддеревам двоичного дерева; листьям соответствуют функции одной переменной x и y .

На рис. 4 приведён пример построения функции по дереву.

Отметим, что алгоритм 1 обращает команду `TreeForm[E[T]]` в среде компьютерной алгебры *Mathematica 9.0* [8] и позволяет задать комбинаторную структуру двоичного дерева при помощи формулы от двух переменных.

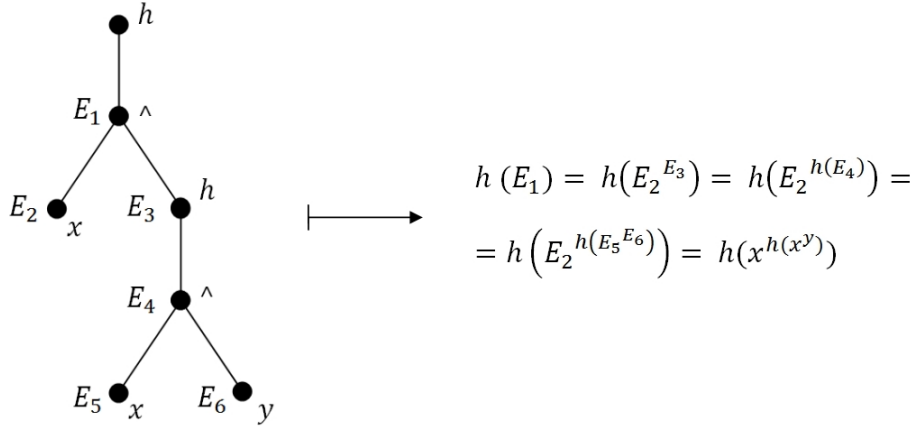


Рис. 4. Построение функции $E(T)$ конечной аналитической сложности по двоичному дереву T

Определение 2. Аналитической сложностью $C(T)$ двоичного дерева T называется аналитическая сложность функции, которая кодирует это дерево в смысле алгоритма 1.

Пример 3. Для изображённых на рис. 5 двоичных деревьев соответствующие им функции двух переменных и их аналитическая сложность имеют следующий вид: $E(T_1) = h(h(x^y))$, $C(T_1) = 1$; $E(T_2) = x^{(y^x)}$, $C(T_2) = 2$; $E(T_3) = h(h(h(x)))$, $C(T_3) = 0$; $E(T_4) = h(x^{h(x^y)})$, $C(T_4) = 2$; $E(T_5) = h((x^y)^{(x^y)^x})$, $C(T_5) = 3$.

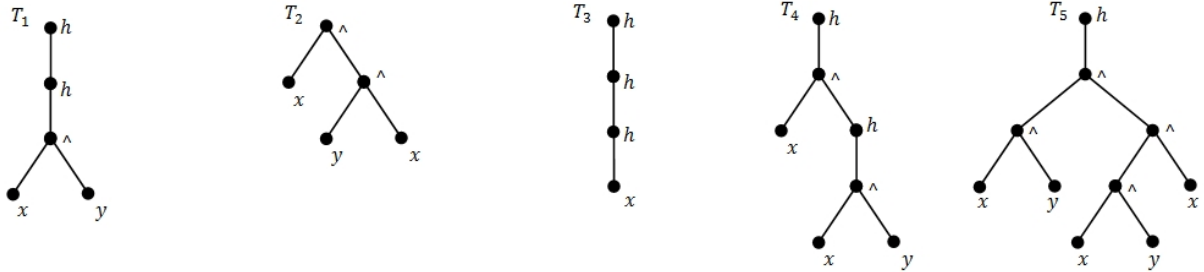


Рис. 5. Примеры двоичных деревьев различной аналитической сложности

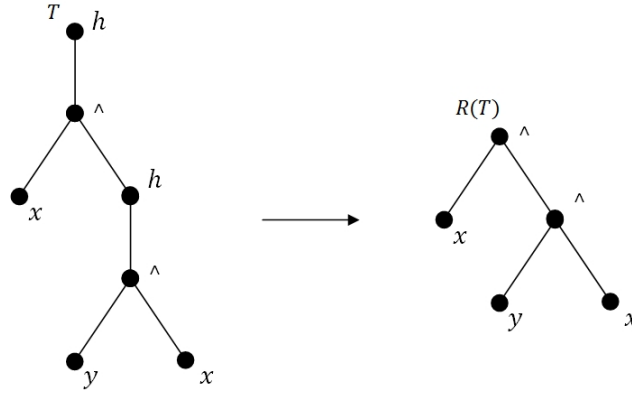
Определение 3. Под *редукцией* $R(T)$ двоичного дерева T будем понимать исключение из дерева отличных от корня вершин степени 2 и корня дерева в случае, если он имеет степень 1.

Например, функции, сопоставляемые деревьям на рис. 6 при помощи алгоритма 1, имеют следующий вид: $E(T) = h(x^{h(y^x)})$, $E(R(T)) = x^{(y^x)}$. Из рассуждений, приведённых в примере 2, следует, что деревья T и $R(T)$ имеют одинаковую аналитическую сложность (равную двум).

Основной результат работы состоит в следующем.

Теорема 1. Аналитическая сложность двоичного дерева T равна числу уровней в дереве $R(T)$.

Доказательство. Дерево $R(T)$ может быть тривиальным (т. е. содержать единственную вершину) лишь в случае, когда исходное дерево T является линейным (т. е. не содержит узлов степени выше двух). Линейному дереву алгоритм 1 сопоставляет

Рис. 6. Редукция $R(T)$ двоичного дерева T

функцию одной переменной $h(h(\dots(x)\dots))$, которая в соответствии с определением 1 имеет нулевую аналитическую сложность. Таким образом, утверждение теоремы имеет место в случае, когда редукция даёт тривиальное дерево.

Пусть редукция $R(T)$ нетривиальна. Если исходное дерево T содержит один уровень, то T есть двоичное дерево с тремя вершинами и двумя рёбрами. Такому дереву алгоритм 1 сопоставляет функцию двух переменных x^y , которая имеет аналитическую сложность равную 1.

Если редукция дерева T нетривиальна, то в соответствии с алгоритмом 1 степень корня редуцированного дерева $R(T)$ равна 2. Линейная же часть дерева T от корня до первого узла степени 2 не влияет ни на результат редукции, ни на аналитическую сложность T . Таким образом, можем без ограничения общности считать, что корень дерева T имеет степень 2.

Предположим, что утверждение теоремы имеет место для всех двоичных деревьев, таких, что соответствующие им редуцированные деревья содержат не более k уровней. Обозначим через T_1 и T_2 поддеревья исходного дерева T , смежные с его корнем. По предположению, аналитическая сложность каждого из деревьев T_i равна числу уровней в $R(T_i)$. Обозначим через $L(T)$ число уровней в дереве T , а через E_i — функцию, сопоставляемую поддереву T_i при помощи алгоритма 1. Из определения редуцированного дерева следует равенство

$$L(R(T)) = \max(L(R(T_1)), L(R(T_2))) + 1. \quad (2)$$

С другой стороны, в соответствии с определением 1 аналитической сложности функции двух переменных

$$C(T) = C(E_1^{E_2}) = (\exp(\exp(\ln \ln E_1 + \ln E_2))) = \max(C(E_1), C(E_2)) + 1. \quad (3)$$

Согласно индуктивной гипотезе, $C(E_i) = L(R(T_i))$, $i = 1, 2$. Это означает, что выражения (2) и (3) совпадают. ■

Отметим некоторые следствия доказанной теоремы.

Следствие 1. Аналитическая сложность дерева с n вершинами не превышает числа $\left\lfloor \frac{n-1}{2} \right\rfloor$, где $\lfloor x \rfloor$ — наибольшее целое число, не превосходящее x .

Доказательство. На рис. 7 изображено одно из двоичных деревьев максимальной аналитической сложности при заданном количестве n вершин.

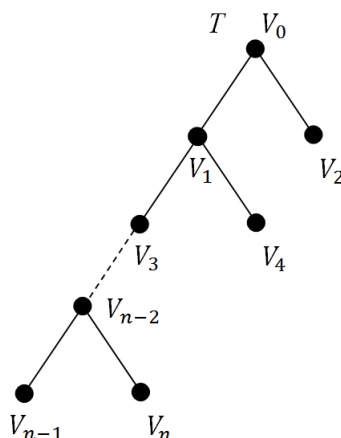


Рис. 7. Двоичное дерево максимальной аналитической сложности при заданном числе вершин

Поскольку редукция дерева на рис. 7 совпадает с ним самим, по теореме 1 его аналитическая сложность равна числу уровней в нём. Следовательно, для дерева на рис. 7 имеет место неравенство $2C(T) + 1 \leq n$. Если исходное дерево допускает нетривиальную редукцию или содержит меньшее количество уровней при фиксированном числе вершин, его аналитическая сложность уменьшается. ■

Следствие 2. Аналитическая сложность дерева T , полученного путём отождествления корня двоичного дерева T_1 с вершиной степени 1 двоичного дерева T_2 , не превосходит суммы их аналитических сложностей $C(T_1) + C(T_2)$.

Доказательство. Действительно, операция редукции может быть применена к каждому из деревьев T_1 и T_2 по отдельности, а значит, согласно теореме 1,

$$C(T) = L(R(T)) \leq L(R(T_1)) + L(R(T_2)) = C(T_1) + C(T_2).$$

Равенство здесь имеет место в том и только в том случае, когда корень дерева T_1 отождествляется с вершиной степени 1 дерева T_2 , находящейся на уровне с максимальным номером. ■

В заключение отметим нижнюю границу скорости роста аналитической сложности двоичного дерева как функции количества его вершин.

Следствие 3. Пусть T — максимально ветвящееся двоичное дерево с k уровнями, т.е. такое, что каждая его вершина, за исключением корня, имеет степень либо 1, либо 3 и все вершины, за исключением уровня $k - 1$, имеют одинаковую степень. Аналитическая сложность дерева T равна $\lfloor \log_2(n + 1) - 1 \rfloor$.

Заключение

Результаты работы дают эффективную методику сравнения кластерных деревьев, полученных на основе реальных данных, и использованных для этого параметров объектов кластеризуемого множества. Малая аналитическая сложность дерева даёт возможность сделать вывод о высокой регулярности его комбинаторной структуры. Близкая аналитическая сложность двух кластерных деревьев, построенных по одному и тому же набору данных для разных метрик, может указывать на сопоставимую эффективность этих метрик для описания исследуемой совокупности объектов (при дополнительном исследовании решающих правил, основанных на этих метриках).

Наконец, близость аналитической сложности кластерного дерева к одной из оценок, данных в следствиях 1 и 3, может указывать либо на недостаточность (для оценки из следствия 1), либо на избыточность (для оценки из следствия 3) множества параметров, учитываемых в метрике, используемой для кластеризации.

Авторы признательны В. К. Белошанке за плодотворные обсуждения вопросов аналитической сложности.

ЛИТЕРАТУРА

1. *Дюран Б., Оделл П.* Кластерный анализ. М.: Статистика, 1977. 128 с.
2. *Миркин Б. Г.* Методы кластерного анализа для поддержки принятия решений. М.: Изд. дом Национального исследовательского университета «Высшая школа экономики», 2011. 84 с.
3. *Белошанка В. К.* Об аналитической сложности функций двух переменных // Российский журнал математической физики. 2007. Т. 14. № 3. С. 243–249.
4. *Зыков А. П.* Основы теории графов. М.: Наука, 1987. 383 с.
5. *Арнольд В. И., Варченко А. Н., Гусейн-Заде С. М.* Особенности дифференцируемых отображений. Т. 1. М.: Наука, 1984. 293 с.
6. *Красиков В. А., Садыков Т. М.* Об аналитической сложности дискриминантов // Труды Математического института им. В. А. Стеклова. 2012. Т. 279. С. 86–101.
7. *Чикалов И. В.* Алгоритм построения деревьев решений с минимальным суммарным весом вершин // Вестник ННГУ. Математическое моделирование и оптимальное управление. 2000. Т. 11. № 2. С. 200–204.
8. *Морозов А. А., Таранчук В. Б.* Программирование задач численного анализа в системе Mathematica. Минск: БГПУ, 2005. 145 с.