© 2024 г. Б.Г. МИРКИН, д-р техн. наук (bmirkin@hse.ru)
(Национальный исследовательский университет
“Высшая школа экономики”, Москва;
университет Лондона, Биркбек),
А.А. ПАРИНОВ, (aparinov@hse.ru)
(Национальный исследовательский университет
“Высшая школа экономики”, Москва)

АГЛОМЕРАТИВНЫЙ КОНСЕНСУСНЫЙ КЛАСТЕР-АНАЛИЗ С АВТОМАТИЧЕСКИМ ВЫБОРОМ ЧИСЛА КЛАСТЕРОВ[1]

Представлены теоретические и вычислительные результаты, связанные с оригинальной моделью консенсусного кластерного анализа, основанной на так называемом проективном расстоянии между разбиениями. Это расстояние определяется как сумма квадратов элементов разности бинарной матрицы инциденций одного разбиения и ее ортогональной проекции на подпространство, порождаемое столбцами матрицы инциденций другого разбиения. Оказывается, при достаточном количестве разбиений предлагаемый метод агломеративного кластеринга правильно вычисляет не только консенсусное разбиение, но число кластеров в нем.

Ключевые слова: консенсусный кластер-анализ, проективное расстояние, консенсусная матрица, агломеративный кластер-анализ, средне-взвешенный критерий.

DOI: 10.31857/S0005231024030014, **EDN:** UCGYKT

1. Введение

Проблема консенсусного кластерного анализа состоит в следующем. Задана некоторая совокупность разбиений данного множества объектов, иногда называемая кластерным ансамблем. Требуется сформировать некое “усредненное” разбиение, наиболее согласованное с имеющейся совокупностью. Впервые эта проблема была сформулирована как математическая задача аппроксимации в работе [1] с использованием введенного им расстояния между разбиениями в связи с предложенным им общим подходом к анализу данных неколичественного вида. Аксиоматическая характеризация расстояния Миркина была опубликована в данном журнале в статье [2]. Оказалось, что использование этого расстояния в задаче аппроксимации не совсем адекватно, так как не удовлетворяет так называемому тесту Мучника (см. раздел 7.6.4 в [3]). Поэтому в [3, 4] предложена более адекватная мера близости между

[1] Статья выполнена при поддержке Российского научного фонда (проект № 22-11-00323) в НИУ ВШЭ, Москва, Россия. Исследование выполнено с использованием суперкомпьютерного комплекса НИУ ВШЭ.

разбиениями, называемая в данной работе проективным расстоянием, которая до сего времени не подвергалась эмпирической верификации. Интерес к данной проблематике со стороны международного сообщества пробудился уже в новом веке с публикацией статей [5, 6]. В этих и последующих статьях по консенсусному кластер-анализу (см., например, обзоры в [7, 8]) мотивацией послужила вполне практическая и насущная проблематика. Дело в том, что кластер-анализ широко используется во многих практических приложениях — маркетинге, банковском деле, биоинформатике, искусственном интеллекте и пр. Между тем, результаты применения алгоритмов зависят от параметров, задаваемых пользователем "на глазок", таких как количество кластеров или порог значимости расстояния. Поэтому возникает совокупность более или менее равноважных кластерных разбиений, полученных при различных параметризациях, и, следовательно, задача консенсусного кластерного анализа.

Цель данной статьи — представить и обосновать метод формирования консенсусного разбиения, основанный на использовании проективного расстояния между разбиениями. Метод использует иерархическую схему агломеративного кластер-анализа на основе так называемой консенсусной матрицы связей между объектами со следующими особенносями: (1) сдвиг величин связи так, чтобы сумма всех связей стала нулевой; (2) использование средневзвешенной внутренней связи в качестве максимизируемого критерия; (3) обнуление диагональных элементов матрицы связи после каждого объединения. Критерий средневзвешенной внутренней связи реализует цель минимизации суммарного проективного расстояния между консенсусным разбиением и заданным ансамблем. Проективное расстояние между двумя разбиениями определяется как сумма квадратов элементов разности двух матриц: матрицы инциденций разбиения из данного ансамбля и его ортогональной проекции на линейное подпространство, порожденное матрицей инциденций искомого консенсусного разбиения [3, 9]. Вычислительный эксперимент основан на новом генераторе ансамбля "синтетических" разбиений. Генератор включает вероятность "мутации", которая определяет и разнообразие генерируемых разбиений и их близость к исходному "истинному" разбиению. В качестве конкурентных алгоритмов используются наиболее популярные схемы максимизации суммарных внутренних связей, включая так называемый критерий модулярности [10] и алгоритм Лувен [4].

2. Консенсусная матрица и методы ее сдвига

При заданном ансамбле разбиений $R_1, \ldots, R_M$ на множестве N объектов I консенсусное разбиение обычно формируется с использованием так называемой консенсусной матрицы размерности $N \times N$, $A = (a_{ij})$, (i, j)-й элемент которой, a_{ij}, определяется как количество таких разбиений ансамбля, R_m $(m = 1, \ldots, M)$, в которых и i, и j принадлежат одному и тому же классу $(i, j \in I)$. Любое разбиение R множества I можно взаимнооднозначно пред-

ставить через ее бинарную $N \times N$ матрицу смежности $r = (r_{ij})$, в которой $r_{ij} = 1$, если i и j находятся в одном и том же классе R, и $r_{ij} = 0$ в противном случае. Как известно, $A = \sum_m r_m$, где r_m – матрица смежности R_m $(m = 1, \ldots, M)$.

Элементы консенсусной матрицы выражают степень сходства между объектами согласно заданным разбиениям R_m $(m = 1, \ldots, M)$. Наиболее сходными являются те объекты, которые входят в один и тот же класс во всех разбиениях ансамбля. Для таких объектов консенсусная связь равна $a_{ij} = M$. Напротив, самые отличающиеся объекты — те, которые входят в разные классы во всех разбиениях ансамбля без исключения. Для них $a_{ij} = 0$. Элементы матрицы A неотрицательны. Для дальнейшего анализа полезно преобразовать эту матрицу так, чтобы часть элементов стала отрицательной, а среднее значение связи стало равным нулю.

В литературе предложено два способа такого преобразования: сдвиг модулярности [10] и сдвиг шкалы [3], определяемые следующим образом.

- **Сдвиг модулярности.** Это преобразование использует понятие случайного взаимодействия. Суммарные связи $a_{i.} = \sum_j a_{ij}$ и $a_{.j} = \sum_i a_{ij}$ рассматриваются как "энергетические заряды" строки i и столбца j соответственно. Случайное взаимодействие зарядов выражается их произведением, точнее величиной $a_{i.}a_{.j}/a_{..}$, где $a_{..} = \sum_i a_{i.} = sum_j a_{.j}$ — суммарный заряд, так чтобы взаимодействие также выражалось в единицах заряда. Это приводит к следующему преобразованию модулярности:

  $$a_{ij} \leftarrow a_{ij} - a_{i.}a_{.j}/a_{..}, \tag{1}$$

  поэтому нетрудно доказать, что после применения сдвига модулярности сумма всех связей, а значит, и их средняя величина, становится равной нулю.

- **Сдвиг шкалы**

  Это преобразование состоит в вычитании из всех элемнтов матрицы одной и той же пороговой величины, равной величине средней связи $\overline{a} = \frac{\sum\limits_{i,j} a_{ij}}{N^2} = \frac{a_{..}}{N^2}$:

  $$a_{ij} \leftarrow a_{ij} - \overline{a}. \tag{2}$$

  Конечно, после этого сдвига нуля шкалы в точку среднего и само среднее, и суммарная связь становятся тоже нулевыми.

Пример. Рассмотрим множество, состоящее из шести объектов $I = \{1, 2, 3, 4, 5, 6\}$ и пять разбиений на этом множестве, представленных в табл. 1 цифровыми метками классов.

Консенсусная матрица для этих данных представлена в табл. 2.

Таблица 1. Пять разбиений на множестве шести объектов, представленные цифровыми метками классов в соответствующих столбцах

№	R1	R2	R3	R4	R5
1	1	1	2	3	3
2	1	1	1	3	2
3	1	2	1	2	2
4	2	2	2	2	3
5	2	2	2	1	1
6	2	3	2	1	1

Таблица 2. Консенсусная матрица пяти разбений, представленных в табл. 1

	1	2	3	4	5	6
1	5	3	1	2	1	1
2	3	5	3	0	0	0
3	1	3	5	2	1	0
4	2	0	2	5	3	2
5	1	0	1	3	5	4
6	1	0	0	2	4	5

Таблица 3. Консенсусная матрица (слева), матрица случайных взаимодействий (в середине) и результат ее модулярного сдвига (справа)

	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
1	5	3	1	2	1	1	2,22	1,88	2,05	2,39	2,39	2,05	2,78	1,12	–1,05	–0,39	–1,39	–1,05
2	3	5	3	0	0	0	1,88	1,59	1,74	2,03	2,03	1,74	1,12	3,41	1,26	–2,03	–2,03	–1,74
3	1	3	5	2	1	0	2,05	1,74	1,89	2,21	2,21	1,89	–1,05	1,26	3,11	–0,21	–1,21	–1,89
4	2	0	2	5	3	2	2,39	2,03	2,21	2,58	2,58	2,20	–0,39	–2,03	–0,21	2,42	0,42	–0,21
5	1	0	1	3	5	4	2,39	2,03	2,21	2,58	2,58	2,21	–1,39	–2,03	–1,21	0,42	2,42	1,79
6	1	0	0	2	4	5	2,05	1,74	1,89	2,21	2,21	1,89	–1,05	–1,74	–1,89	–0,21	1,79	3,11

Таблица 4. Консенсусная матрица после сдвига шкалы

	1	2	3	4	5	6
1	1,96	–0,04	–2,04	–1,04	–2,04	–2,04
2	–0,04	1,96	–0,04	–3,04	–3,04	–3,04
3	–2,04	–0,04	1,96	–1,04	–2,04	–3,04
4	–1,04	–3,04	–1,04	1,96	–0,04	–1,04
5	–2,04	–3,04	–2,04	–0,04	1,96	0,96
6	–2,04	–3,04	–3,04	–1,04	0,96	1,96

В следующей табл. 3 представлена и сама эта матрица, и рассчитанная на ее основе матрица случайных взаимодействий, и результат ее модулярного преобразования.

Для получения матрицы связей со сдвигом шкалы, нужно подсчитать среднюю связь, 3.04, между объектами и вычесть ее из всех элементов консенсусной матрицы (см. табл. 4).

Сравнивая табл. 3 и 4, нельзя не заметить, что модулярный сдвиг оставляет положительными значительно больше связей, чем сдвиг шкалы. В правой части табл. 3 имеются две строки с тремя положительными элементами каждая — это строки 2 и 5. Напротив, в матрице после сдвига шкалы остается только один внедиагональный положительный элемент (см. табл. 4). Это важно потому, что только положительные связи могут “заставить” объекты объединяться в одном кластере.

3. Критерии разбиения

Из различных критериев кластер-анализа, опубликованных в литературе, рассмотрим два критерия, использующих внутрикластерные связи. В качестве связей рассматриваются элементы консенсусной матрицы, полученной сдвигом шкалы или модулярности. Один из этих критериев — это сумма связей внутри классов разбиения. Иными словами, для всякого разбиения $R = \{R_1, \dots, R_K\}$ объектов множества I на K классов/кластеров внутренняя связь в кластере R_k $(k = 1, \dots, K)$ рассчитывается как сумма всех связей a_{ij} для таких пар (i, j), в которых оба объекта, и i, и j, принадлежат R_k. Таким образом, суммарная внутренняя связь в R выражается формулой

$$f(R) = \sum_{k=1}^{K} \sum_{i,j \in R_k} a_{ij}. \tag{3}$$

Задача состоит в том, чтобы найти разбиение R, максимизирующее эту величину. Очевидно, что при неотрицательных связях a_{ij} данный критерий ведет к тривиальному решению: максимум суммы внутренних связей достигается на разбиении, состоящем из единственного универсального кластера, включающего все объекты. Именно поэтому предварительное преобразование связей путем сдвига шкалы или модулярности оказывается существенным. Следует отметить, что популярный критерий модулярности разбиения [10, 13] есть не что иное, как суммарная внутренняя связь (3) после сдвига модулярности [3].

Другой рассматриваемый здесь критерий — это средневзвешенная внутренняя связь [3]:

$$g(R) = \sum_{k=1}^{K} \frac{\sum\limits_{i,j \in R_k} a_{ij}}{N_k}, \tag{4}$$

где N_k – это количество объектов в кластере R_k.

Критерий (4) похож на критерий (3). Разница в том, что суммы связей внутри кластеров делятся здесь на их численности. По-другому говоря, в качестве оценки внутрикластерных связей здесь выступает произведение средней внутренней связи и его численности, что и объясняет название критерия. Критерий средневзвешенной внутренней связи возникает в контексте аппроксимирующего кластер-анализа [12]. Хотя он и не обязательно ведет к универсальному кластеру как оптимальному решению при неотрицательных связях, все равно полезно его применять после сдвига связей к нулевому среднему, как видно далее. Именно этот критерий является наиболее адекватным в задаче консенсусного кластер-анализа, как будет объяснено в разделе 5.

4. Агломеративные алгоритмы

Имеется несколько подходов, обычно применяемых для получения локально-оптимальных решений: объединение (агломерация) малых кластеров в большие, разделение больших кластеров на меньшие части, последовательное формирование кластеров по одному, обмен объектов из разных кластеров и т.п. Здесь рассматриваются только агломеративные алгоритмы, поскольку предполагается показать, что этот подход, примененный к средневзвешенному критерию (4), вполне эффективен для достижения цели консенсусного кластер-анализа.

Ниже показаны три агломеративных алгоритма для кластер-анализа.

Первый из них — обычный агломеративный алгоритм применительно к критерию суммарной связи (3).

Алгоритм AgSu:

1. Инициализация. В качестве начального принимается разбиение на синглтоны – кластеры, состоящие из одного объекта каждый, $S = \{S_1, \ldots, S_N\}$, $S_k = \{k\}$, $k = 1, \ldots, N$. Определяем матрицу связей между ними, $B = (b_{ij})$, как равную $A = (a_{ij})$, так что связь b_{ij} между синглтонами $\{i\}$ и $\{j\}$ равна a_{ij}.

2. Общий шаг. При заданном разбиении $S = \{S_1, \ldots, S_m\}$ множества объектов I и $m \times m$ матрице $B = (b_{st})$ суммарных связей между кластерами $(s, t = 1, \ldots, m)$ найти максимальную величину $b_{s*t*} = \max_{s,t} b_{st}$. Если $b_{s*t*} > 0$, следует соединить кластеры S_{s*} и S_{t*} в объединенный кластер $S_{s*} \leftarrow S_{s*} \cup S_{t*}$ и пересчитать величины связей путем арифметического прибавления строки $t*$ к строке $s*$: $b_{s*t} \leftarrow b_{s*t} + b_{t*t}$ для всех $t = 1, \ldots, m$, после чего подобным же образом сложить столбцы: $b_{ss*} \leftarrow b_{ss*} + b_{st*}$ для всех $s = 1, \ldots, m$. После этого удалить строку и столбец $t*$ из матрицы B и уменьшить m на 1. Если $b_{s*t*} < 0$, вычисления прекращаются. Если $m < 3$, тоже стоп.

Этот алгоритм действительно локально-оптимальный: на каждом шаге объединение производится оптимальным образом, поскольку увеличивает значение критерия максимально возможным образом. Действительно, разность между значениями критерия (3) после объединения кластеров s и t и до этого равна $2b_{st}$.

Сформулируем агломеративный алгоритм, предназначенный для максимизации средневзвешенного критерия (4). Единственное отличие от алгоритма AgSu состоит в том, что теперь объединение кластеров должно максимизировать критерий (4), а не (3).

Рассмотрим какое-нибудь разбиение $S = \{S_1, \ldots, S_m\}$ и разбиение $S(s, t) = \{S_1, \ldots, S_{s-1}, S_s \cup S_t, \ldots, S_m\}$, полученное объединением классов S_s и S_t в этом разбиении. Разность между $g(S(s, t))$ и $g(S)$ может быть выражена следующим образом:

$$\Delta(s, t) = g(S(s, t)) - g(S) = \frac{2b_{st} - N_t b_s s/N_s - N_s b_t t/N_t}{N_s + N_t}, \tag{5}$$

где b_{st}, b_{ss}, b_{tt} – элементы матрицы B суммарных связей между кластерами или внутри кластеров. Эта формула доказывается элементарными преобразованиями формулы для разности $g(S(s,t)) - g(S)$.

Алгоритм AgSa:

1. Инициализация. В качестве начального принимается разбиение на синглтоны – кластеры, состоящие из одного объекта каждый, $S = \{S_1, \ldots, S_N\}$, $S_k = \{k\}$, $k = 1, 2, \ldots, N$. Определяем матрицу связей между ними, $B = (b_{ij})$, как равную $A = (a_{ij})$, так что связь b_{ij} между синглтонами $\{i\}$ и $\{j\}$ равна a_{ij}.

2. Общий шаг. При заданном разбиении $S = \{S_1, \ldots, S_m\}$ множества объектов I и $m \times m$ матрице $B = (b_{st})$ суммарных связей между кластерами $(s, t = 1, \ldots, m)$, найти максимальную величину $\Delta(s*, t*) = \max_{s,t} \Delta(s,t)$ согласно (5). Если $\Delta(s*, t*) > 0$, следует соединить кластеры S_{s*} и S_{t*} в объединенный кластер $S_{s*} \leftarrow S_{s*} \cup S_{t*}$ и пересчитать величины связей путем арифметического прибавления строки $t*$ к строке $s*$: $b_{s*t} \leftarrow b_{s*t} + b_{t*t}$ для всех $t = 1, \ldots, m$, после чего подобным же образом сложить столбцы: $b_{ss*} \leftarrow b_{ss*} + b_{st*}$ для всех $s = 1, \ldots, m$. После этого удалить строку и столбец $t*$ из матрицы B и уменьшить m на 1. Если $\Delta(s*, t*) < 0$, вычисления прекращаются. Если $m < 3$, тоже стоп.

Алгоритмы AgSu и AgSa имеют то преимущество, что после сдвига число кластеров определяется автоматически — объединение кластеров прекращается, как только все внедиагональные величины b_{st} для алгоритма AgSu, или $\Delta(s,t)$ для алгоритма AgSa, становятся отрицательными. В этом случае никакое дальнейшее объединение кластеров не может увеличить значения критерия.

Как хорошо известно, агломеративные алгоритмы имеют тот недостаток, что используют сравнительно медленные вычисления — число шагов при поиске максимального элемента матрицы имеет порядок N^2, особенно на первых шагах. Некоторые усилия были предприняты по сокращению вычислений за счет использования таких свойств критериев кластеризации, которые позволяют применить результаты предыдущих вычислений [18]. Позже появилась работа [5] с революционной идеей, что вовсе нет никакой нужды в изнуряющем поиске максимума в матрице. Авторы использовали суммарный критерий (3) после сдвига модулярности, чтобы сформулировать и обосновать свою идею, названную ими Лувенский алгоритм: давайте возьмем один элемент s пары (s,t) случайным образом, так что максимум b_{st} определяется только перебором t порядка N, не N^2. Конечно, эту идею можно использовать с любым критерием, не только (3). Далее сформулируем Лувенский алгоритм применительно к произвольному критерию $c(R)$, который надо максимизировать на множестве всех разбиений множества I. Будем считать, что вычисление разности $\Delta c(s,t) = c(S(s,t)) - c(S)$ — несложная операция.

Лувенский алгоритм GAL.

1. Инициализация. Рассматривай тривиальное разбиение на N одиночных кластеров $R = \{\{1\}, \ldots, \{N\}\}$ в качестве начального разбиения.

2. Общий шаг. При заданном разбиении $R = \{R_1, \ldots, R_K\}$ организуй последовательный просмотр кластеров в произвольном порядке $1, \ldots, K$.

2.1. Для каждого конкретного s найти $t*$, максимизирующий разность $\Delta c(s,t)$ по всем $t = 1, \ldots, K$.

2.2. Объедини кластеры R_s и R_{t*}, в качестве R рассматривай разбиение $R(s,t*)$ и уменьши K на 1.

2.3. Проверка: Если $\Delta c(s,t*) < 0$ или $K < 3$, то останов. В противном случае переходи к следующему кластеру.

2.4. Если все кластеры пройдены, начинай шаг 2 с текущим R.

Определим одну дополнительную операцию, которую можно выполнять перед началом работы агломеративного алгоритма и/или на любом его шаге:

ZD: Обнуление диагональных элементов матрицы связи.

Согласно этой операции каждый диагональный элемент текущей матрицы связей B заменяется нулем.

Было проверено, стоит ли выполнять ZD перед каждым агломеративным шагом. Оказалось, что это помогает в алгоритме AgSa и не помогает — точнее, ухудшает результат — при других рассматривавшихся алгоритмах. Поэтому далее применяем ZD перед каждым шагом объединения в алгоритме AgSa, а в других алгоритмах — только вначале.

5. Вычислительный эксперимент

В настоящее время эксперименты с методами консенсусного кластер-анализа носят, если так можно выразиться, опосредованный характер. Берется таблица данных, либо реальная, скажем из хранилища данных в Ирвайнском кампусе Университета Калифорнии, либо "синтетическая", содержащая кластерную структуру, порожденную тем или иным генератором кластерных структур. К этой таблице повторно применяется один кластерный алгоритм или несколько алгоритмов при различных, как правило случайных, значениях параметров алгоритма. Например, это может быть метод к-средних при случайных инициализациях и постоянном числе кластеров, равно одному в сгенерированной таблице данных. Результаты этих применений и образуют исходный ансамбль разбиений. Таким образом, проблематика консенсусного кластер-анализа здесь комбинируется со спецификой взятой таблицы данных и алгоритма или алгоритмов получения разбиений. Это порождает вопросы, связанные с качеством формируемых кластерных ансамблей, их репрезентативностью, их разнообразием, полнотой и прочее [11, 14, 19]. Тематика консенсусного кластер-анализа не имеет никакого отношения ни к качеству данных, ни к качеству алгоритмов кластеризации, используемых для получения ансамблей разбиений. Эксперименты должны быть организованы таким образом, чтобы генератор данных непосредственно генерировал ансамбль разбиений множества I так, чтобы и разноообразие ансамбля и его репрезентативность было легко контролировать.

Здесь предлагаем именно такую организацию вычислительного эксперимента для консенсусного кластер-анализа. Генератор "синтетических" данных начинает с генерации "истинного" разбиения. Этим процессом управляют три параметра: численность множества объектов N, количество классов в разбиении K, минимальный размер класса m. Этот последний параметр особенно полезен, когда в применяемых алгоритмах кластер-анализа используются вероятностные соображения. Для оценки параметров такого алгоритма может понадобиться не менее m элементов. Для выполнения условия распределяем mK объектов по классам, помещая в каждый ровно m объектов. Остающиеся $N - mK$ объектов случайно распределяются по K кластерам. Это можно сделать в Матлаб с помощью команды $randi(K, T)$, которая приписывает каждый из $T = N - Km$ объектов какому-то из K разных классов.

После того, как получено истинное разбиение S, генерируется ансамбль разбиений $R_1, \ldots, R_M$ для его представления. Для этого задается вероятность "мутации", p, $0 < p < 1$. Для генерации R_1 переназначаем $100p\%$ сущностей в любой случайно выбранный кластер. Другие разделы генерируются аналогично. Увеличивая p, увеличиваем разнообразие ансамбля. То, что такой ансамбль является репрезентативным для истинного раздела, следует из **генерации**.

Кому-то такой механизм мутации может показаться слишком упрощенным. Например, все разделы, созданные с его помощью, имеют то же количество кластеров, что и истинные. Действительно, можно предложить более сложные схемы мутации, например со случайными слияниями и разделениями кластеров исходных данных. Отметим, однако, хорошие свойства предложенного генератора данных. Во-первых, увеличивая вероятность мутации p, действительно можно создавать достаточно разнообразные разделы. Во-вторых, уменьшая число разделов в ансамблях, можно создавать действительно сложные ситуации для консенсусных алгоритмов кластеризации, например делая их число меньше, чем число частей в истинном разделе, $M < K$.

В последующих расчетах были использованы два значения количества объектов $N = 1000$ и $N = 3000$, три значения количества кластеров $K = 4$, $K = 9$ и $K = 15$, и два значения размера ансамбля разбиений $M = 40$ и $M = 10$. Они сведены в табл. 5, в которой также перечислены рассматриваемые алгоритмы.

Качество результатов оценивается по двум характеристикам: количеству полученных кластеров и ARI (Adjusted Rand index), индексу сходства между

Таблица 5. Параметры экспериментов

N	K	M	m	p	Типа сдвига данных	Алгоритм
1000	4	10	2	0,8	Модульный сдвиг	Агломеративный
3000	9	40			Сдвиг масштаба	Лувена
	15					

истинным и полученным алгоритмом разбиением [6]. ARI основан на количестве пар объектов, которые совпадают в сравниваемых разделах, т.е. либо принадлежат к одному кластеру, либо к разным кластерам в обоих разделах:

$$(6) \quad ARI\,(A,B) = \frac{\binom{N}{2} * \sum_{s=1}^{K_A}\sum_{t=1}^{K_B}\binom{n_{st}}{2} - \sum_{s=1}^{K_A}\binom{a_s}{2}\sum_{t=1}^{K_B}\binom{b_t}{2}}{\frac{1}{2}\binom{N}{2}\left[\sum_{s=1}^{K_A}\binom{a_s}{2} + \sum_{t=1}^{K_B}\binom{b_t}{2}\right] - \sum_{s=1}^{K_A}\binom{a_s}{2}\sum_{t=1}^{K_B}\binom{b_t}{2}}.$$

В (6) A и B – два разбиения множества сущностей с частями K_A и K_B соответственно; a_s и b_t – кардинальности частей в A и B соответственно; n_{st} – частоты в совместном распределении AB; $\binom{n}{2}$ – биномиальный член, равный $n(n-1)/2$.

Чем ближе значение ARI к единице, тем более похожи разделы; $ARI = 1,0$ показывает, что $A = B$. Если один из разделов состоит только из одной части, самого множества I, то $ARI = 0$. ARI может быть и отрицательным, что случается довольно редко, как, скажем, при специально определенных "дуальных" парах разделов [7].

После создания ансамбля разбиений и вычисления соответствующей консенсусной матрицы происходят вычисления одним из восьми вариантов обработки в зависимости от варианта преобразования матрицы (модульный сдвиг или сдвиг масштаба), используемого критерия (суммарный или средневзвешенный) и применяемого алгоритма (агломерация или Лувен). Результаты представлены в табл. 6–9 в зависимости от размера данных N, и ансамблей разбиений M.

Эти таблицы наглядно демонстрируют следующее.

1. Результаты при 1000 и 3000 объектах практически совпадают, это означает, что количество рассматриваемых объектов мало влияет на консенсусные решения.

Таблица 6. Результаты применения алгоритма консенсус кластеризации при $N = 1000$, $M = 40$

		Суммарный критерий				Средневзвешенный критерий			
		Лувен		Агломерация		Лувен		Агломерация	
		Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль
4	ARI	0,84/0,09	0,88/0,03	0,89/0,03	0,90/0,01	0,98/0,01	0,98/0,01	0,99/0,0	0,99/0,0
	#	3,8/0,44	4/0	4/0	4/0	8,6/1,1	9/1,4	4/0	4/0
9	ARI	0,44/0,03	0,43/0,04	0,45/0,01	0,50/0,03	0,99/0,0	0,99/0,00	1,0/0,0	1,0/0,0
	#	4,2/0,45	4,2/0,45	4/0	4,8/0,45	11,6/1,5	11,8/1,3	9/0	9/0
15	ARI	0,29/0,01	0,28/0,01	0,33/0,02	0,34/0,01	0,99/0,0	0,99/0,0	1/0	1/0
	#	4/0	4/0	4,8/0,45	5/0	17,4/0,5	17,6/0,89	15/0	15/0

Таблица 7. Результаты применения алгоритма консенсус кластеризации при $N = 3000$, $M = 40$

		Суммарный критерий				Средневзвешенный критерий			
		Лувен		Агломерация		Лувен		Агломерация	
		Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль
4	ARI	0,88/0,01	0,87/0,01	0,88/0,01	0,88/0,01	0,98/0,00	0,98/0,00	0,99/0,0	0,99/0,0
	#	4/0	4/0	4/0	4/0	12,2/1,1	13/1	4/0	4/0
9	ARI	0,40/0,02	0,40/0,03	0,43/0,05	0,41/0,02	0,99/0,0	0,99/0,00	1,0/0,0	1,0/0,0
	#	4,2/0,45	3,8/0,45	4,2/0,45	4/0	14,6/0,55	13,8/0,45	9/0	9/0
15	ARI	0,26/0,01	0,27/0,01	0,29/0,02	0,28/0,02	1,0/0,0	0,99/0,0	1/0	1/0
	#	4/0	4/0	4,4/0,55	4,4/0,55	19,4/1,1	19,8/1,3	15/0	15/0

Таблица 8. Результаты применения алгоритма консенсус кластеризации при $N = 1000$, $M = 10$

		Суммарный критерий				Средневзвешенный критерий			
		Лувен		Агломерация		Лувен		Агломерация	
		Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль
4	ARI	0,44/0,02	0,43/0,02	0,41/0,03	0,40/0,02	0,70/0,02	0,70/0,02	0,67/0,03	0,66/0,03
	#	3/0	4/0	4/0	3,2/0,45	13,8/0,84	14/1	4/0	4/0
9	ARI	0,18/0,02	0,16/0,01	0,21/0,05	0,21/0,02	0,73/0,01	0,79/0,01	0,73/0,01	0,74/0,01
	#	3/0	4/1	3/0	3,8/0,84	20,2/1,1	20,0/1,2	9/0	9/0
15	ARI	0,12/0,01	0,11/0,01	0,14/0,01	0,13/0,01	0,81/0,01	0,81/0,01	0,76/0,03	0,76/0,03
	#	3/0	4/0,71	3,4/0,55	4,2/0,45	26,2/1,9	26,2/1,3	14,2/0,84	14,4/0,55

Таблица 9. Результаты применения алгоритма консенсус кластеризации при $N = 3000$, $M = 10$

		Суммарный критерий				Средневзвешенный критерий			
		Лувен		Агломерация		Лувен		Агломерация	
		Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль	Масштаб	Модуль
4	ARI	0,39/0,03	0,40/0,01	0,41/0,01	0,40/0,02	0,71/0,01	0,71/0,01	0,69/0,01	0,68/0,01
	#	3/0	3/0	3/0	3,2/0,45	12,2/3,63	13,4/2,88	4/0	4/0
9	ARI	0,16/0,01	0,16/0,02	0,17/0,01	0,17/0,02	0,79/0,01	0,79/0,01	0,72/0,01	0,72/0,01
	#	3/0	3,6/0,55	3/0	3,6/0,55	23,4/0,89	23,6/0,89	9/0	9/0
15	ARI	0,10/0,01	0,10/0,01	0,10/0,01	0,11/0,01	0,83/0,01	0,83/0,01	0,77/0,01	0,77/0,01
	#	3/0	4,4/0,89	3,4/0,55	4,6/0,55	30,0/1,9	31,6/2,7	15/0	15/0

2. Вопреки ожиданиям, результаты при двух разных нормализациях — модульной и со сдвигом масштаба — во многом схожи, так что вопрос о том, какую из них использовать, становится нерелевантным при консенсусной кластеризации.

3. Восстановление кластеров всегда лучше при использовании полусреднего критерия, а не суммарного. Чем больше число кластеров, тем больше разница.

4. Чем больше размер ансамбля, тем лучше: результаты восстановления данных при $M = 10$ значительно хуже, чем при $M = 40$. Особенно разрушительным является эффект при суммарном критерии, при котором уровень восстановления кластеров сохраняется в среднем на уровне ARI, равном 0,4 при $K = 4$, 0,2 при $K = 9$ и 0,1 при $K = 15$.

5. При $M = 40$ агломерация по полусреднему критерию приводит к идеальным результатам при $K = 9$, 15 и почти идеальным при $K = 4$. Лувенский алгоритм при полусреднем критерии достигает почти таких же хороших результатов в восстановлении кластеров. Однако он терпит неудачу в отношении количества кластеров. Напротив, при $M = 10$ алгоритм Лувена всегда выигрывает в восстановлении кластеров, хотя по-прежнему переоценивает их количество.

6. Среди рассматриваемых методов есть один, который всегда правильно восстанавливает количество кластеров: агломерация по критерию полусреднего. Он работает даже при уменьшении ARI до величины порядка 0,7. Единственный случай, в котором он может потерпеть неудачу, пусть и незначительную, — это случай $K = 15$, $M = 10$, т.е. $M < K$, при меньшем количестве объектов ($N = 1000$, но не при $N = 3000$).

6. Обсуждение

Некоторые из приведенных выше эмпирических результатов могут быть объяснены теоретическими соображениями, связанными с общей концепцией консенсусной кластеризации, основанной на индексах расстояния между разделами. Учитывая индекс $d(R, S)$, оценивающий несходство между любыми разделами R и S из I, можно определить понятие консенсусного раздела следующим образом. Если задан ансамбль разделов $R_1, R_2, \ldots, R_M$ из I, то консенсусным разделом является любой раздел R из I, который минимизирует суммарное расстояние $D(R) = \sum_{m=1}^{M} d(R_m, R)$. Обычно расстояние $d(R_m, R)$ определяется как расстояние несовпадения, или расстояние Миркина, между соответствующими $N \times N$ двоичными матрицами r_m и r, элементы которых $r_m(i, j)$ или $r(i, j)$ равны 1, если i и j находятся в одной части Rm или R соответственно; в противном случае они равны 0. Расстояние Миркина — это количество несовместимых пар (i, j), таких, что i и j находятся в одной части одного из разделов, а в другом разделе i и j принадлежат разным частям [17]. Очевидно, что это половина расстояния $L1$ между бинарными матрицами разделов. Нетрудно доказать, что консенсусным разделом с расстоянием Миркина является тот, который максимизирует суммарный критерий

$$F(R) = \sum_{k=1}^{K} \sum_{i,j \in R_k} \left(a_{ij} - \frac{M}{2} \right), \tag{7}$$

где a_{ij} элементы консенсусной матрицы для ансамбля $R_1, \ldots, R_M$. Действительно,

$$D(R) = \sum_{m=1}^{M} d(R_m,\ R) = \sum_{m=1}^{M} \sum_{i,j=1}^{N} |r_m(i,j) - \ r(i,j)|/2 =$$

$$= \sum_{i,j=1}^{N} \sum_{m=1}^{M} |r_m(i,j) - \ r(i,j)|/2. \tag{8}$$

очевидно, что внутренняя сумма равна

$$\sum_{m=1}^{M} |r_m(i,j) - r(i,j)| = \sum_{m=1}^{M} |r_m(i,j) - r(i,j)|^2 =$$

$$= \sum_{m=1}^{M} (r_m(i,j) + r(i,j) - 2r(i,j)\, r_m(i,j)) = a_{ij} + Mr(i,j) - 2a_{ij} r(i,j), \tag{9}$$

тогда как $\sum_{m=1}^{M} r_m(i,j) = \ a_{ij}$.

Эти манипуляции корректны, поскольку значения элементов $r(i,j)$ и $r_m(i,j)$ здесь равны нулю или единице, так что квадратичная операция оставляет их инвариантными. Отбросив первый элемент, a_{ij}, который здесь постоянен, и умножив остаток на –1/2, можно увидеть, что задача минимизации $D(R)$ действительно эквивалентна задаче максимизации $F(R)$ в (7).

Видно, что критерий (7) действительно является внутрикластерным суммарным критерием (3) с предварительным сдвигом сходства на $M/2$. К сожалению, у определенного таким образом консенсусного разбиения есть недостаток: оно не проходит так называемый тест Мучника [9]. Этот тест требует проверить для любого разбиения $T = T_1, \ldots, T_K$ на I является ли T консенсусным разбиением для ансамбля его K дихотомических представлений $T_k = T_k, I - T_k$ $(k = 1, 2, \ldots, K)$. Если да, то расстояние проходит тест; если нет, то расстояние не проходит тест.

Посмотрим, удовлетворяет ли консенсус расстояний Миркина этому тесту. Возьмем разбиение $T = T_1, \ldots, T_7$ с $K = 7$ частями, так что $K/2 = 3,5$. Рассмотрим записи матрицы консенсуса a_{ij} в этом случае. Предположим сначала, что i и j принадлежат одной и той же части T. Тогда они должны принадлежать одной и той же части в каждом T^m, поскольку каждая часть T содержится в любой части T^m. Это означает, что $a_{ij} = 7$ для таких i и j. Рассмотрим теперь, что i принадлежит, скажем, T_1, а j — другой части, скажем T_2. Тогда i и j принадлежат разным частям как в T_1, так и в T_2. Однако они принадлежат одной и той же части $I - T_3$ в T_3, потому что $I - T_3$ содержит и T_1, и T_2. Аналогично, эти i и j принадлежат одной и той же части в T^m для всех остальных $m = 4, 5, 6, 7$. Таким образом, $a_{ij} = 5$ для этих i

и j. Следовательно, все записи в матрице консенсуса здесь равны либо 5, либо 7, причем обе больше, чем $K/2 = 3,5$. Таким образом, для максимизации критерия (7) выгодно собрать все сущности в универсальном разбиении I, состоящем из единственной части самого I, но не из разбиения T. Следовательно, тест Мучника действительно провален.

Существует и другая мера расстояния, которая называется проективным расстоянием [11, 12]. Рассмотрим номинальный признак над множеством сущностей I, представленный разбиением $S = S_l$, и другой номинальный признак, представленный разбиением $R = R_k$. Определим $N \times L$ фиктивную матрицу Y, соответствующую разделу S, матрицу инцидентности раздела, приписав каждой категории S_l в S бинарную переменную y_l, фиктивную, которая является просто $1/0$ N-мерным вектором, элементы которого $y_{il} = 1$, если $i \in Sl$ и $y_{il} = 0$, в противном случае $(l = 1, \ldots, L)$. Аналогично определим $N \times K$ матрицу инцидентности X, столбцы которой x_k — $0/1$-векторы, соответствующие категориям S_k из S. Проективное расстояние определяется как суммарная квадратичная разность между Y и его ортогональной проекцией на линейное пространство, охватывающее столбцы X [12, 11]. Используя символ $\| \|^2$ для обозначения суммы квадратов (квадратичной нормы), проективное расстояние между R и S определяется по формуле $\delta(X, Y) = \|Y - P_X Y\|^2$, где P_X – ортогональный проектор $P_X = X(X^T X)^{-1} X_T$ на линейное пространство, охватывающее столбцы X. Заметим, что эта мера расстояния несимметрична. Точный смысл расстояния $\delta(X, Y)$ разобран в [11, с. 319]. Здесь сосредоточимся на суммарном расстоянии $\triangle(R) = \sum_{m=1}^{M} \delta(X, Y_m)$, которое должно быть минимизировано относительно неизвестного разбиения R, представленного матрицей X, для получения проективной дистанционной консенсусной кластеризации. Матрицы Y_m представляют здесь разделы R_m заданного ансамбля разделов.

Оказывается, эта задача эквивалентна задаче максимизации полусреднего критерия $g(R)$ в (4). Чтобы доказать это, рассмотрим матрицы инцидентности X и Y_m разделов R и R_m соответственно. Эти бинарные матрицы обозначают через 1 принадлежность объектов (строк) к кластерам. Обозначим общее число кластеров во всех разбиениях ансамбля $(m = 1, 2, \ldots, M)$ через L и сформируем $N \times L$ матрицу $Y = (Y_1, Y_2, \ldots, Y_M)$, состоящую из всех L столбцов этих матриц. Столбцы Y соответствуют всем кластерам в разбиениях $R_1, R_2, \ldots, R_M$. Тогда критерий $\triangle(R) = \sum_{m=1}^{M} \delta(X, Y_m)$ можно переформулировать как $\triangle(X) = \|Y - P_X Y\|^2$, или, что эквивалентно, как $\triangle(X) = Tr((Y - P_X Y)(Y - P_X Y)^T)$, где Tr обозначает след квадратной матрицы, что есть сумма ее диагональных элементов. Раскрывая скобки в последнем выражении, получается $\triangle(Y) = = Tr(YY^T - P_X YY^T - YY^T P_X + P_X YY^T P_X) = Tr(YY^T - P_X YY^T)$. Действительно, операция Tr коммутативна, так что $Tr(P_X YY^T) = Tr(YY^T P_X)$ и $Tr(P_X YY^T P_X) = Tr(P_X P_X YY^T) = Tr(P_X YY^T)$. Последнее уравнение следует из того, что $P_X P_X = P_X$, что легко доказать непосредственно. Заметим теперь, что матрица YY^T равна консенсусной матрице A. Очевидно,

что $a_{ii} = L$ для всех $i \in I$, так что $Tr(YY^T) = NL$. С другой стороны, (i,i)-й диагональный элемент матрицы $P_X A$ равен сумме произведений $p_{ij}a_{ij}$, где p_{ij} — либо 0, если i и j находятся в разных кластерах, либо $1/N_k$, если i и j принадлежат одному кластеру S_k. На этом доказательство завершено.

Теперь можно доказать, что консенсусное разбиение, определенное с помощью проективного расстояния, действительно проходит тест Мучника. Рассмотрим снова разбиение $T = T_1, \ldots, T_K$ на I и ансамбль его K дихотомических представлений $T^k = T_k, I - T_k$ $(k = 1, 2, \ldots, K)$. Матрица консенсуса A здесь состоит из $a_{ij} = K$, если i и j принадлежат некоторому T_k $(k = 1, 2, \ldots, K)$, или $a_{ij} = K - 2$, если i и j принадлежат разным частям T. Рассмотрим средневзвешенный критерий (4) для разбиения $R = R_1, R_2, \ldots, R_m$. Обозначим среднее сходство внутри R_k через a_k. Тогда значение (4), очевидно, равно сумме $N_k a_k$, где N_k — количество объектов в R_k. Максимальное значение a_k в этом случае равно K, и оно достигается, когда R_k входит в часть T, поскольку в этом случае все внутрикластерные значения $a_{ij} = K$. Если же, напротив, R_k пересекает несколько частей T, то некоторые внутрикластерные значения a_{ij} будут равны $K - 2$, так что $a_k < K$. Это доказывает, что максимальное значение критерия (4) в рассматриваемом случае равно NK (как сумма всех значений $N_k K$), и оно достигается при любом R, либо совпадающем с T, либо являющимся более гранулированной версией T, полученной путем деления некоторых его частей. Подтверждением полученных результатов можно считать доказанные факты: средневзвешенный критерий (4) воплощает хорошую концепцию консенсусной кластеризации с использованием проективного расстояния между разделами, тогда как суммарный критерий (3) относится к плохой концепции консенсуса кластеризации с использованием расстояния Миркина. Именно поэтому критерий (4) в экспериментах в подавляющем большинстве случаев превосходит критерий (3).

Также предстоит объяснить два других эмпирически наблюдаемых факта:

1. Почему столь разные преобразования данных, как сдвиг модульности и сдвиг масштаба, приводят к очень похожим результатам консенсусной кластеризации?

2. Почему эвристика постоянного обнуления диагональных записей настолько эффективна при определении нужного числа кластеров с помощью критерия полусреднего?

Следует отметить, что видимое "противоречие" между высокими значениями ARI и неправильным количеством кластеров (см. результаты Лувена для полусреднего критерия в табл. 6 и 7 выше) легко объясняется нечувствительностью индекса ARI к лишним мелким кластерам. Возьмем, например, разбиение R множества из 1000 человек на две равные по размеру части. Сделаем из одной из частей 20 одиночных кластеров и обозначим полученное таким образом разбиение I на 22 кластера через S. Индекс ARI между R и S равен 0,96, что не так уж далеко от единицы.

7. Заключение

Основной целью данной работы является выдвижение полусреднего критерия консенсусной кластеризации (4), модифицированного постоянным обнулением главной диагональной эвристики, в качестве критерия, который должен использоваться при консенсусной кластеризации для восстановления как скрытого разбиения, так и количества кластеров в нем. Отметим, что этот критерий возникает при консенсусной кластеризации со специально разработанной системой оценки несходства между разделами — проективным расстоянием. В отличие от традиционно используемых расстояния несовпадения или расстояния Миркина между разделами (см., например, в [5]), проективное расстояние, как показано, проходит естественный тест на валидность (тест Мучника). В представленных экспериментах агломеративная кластеризация с критерием (4) демонстрирует очень сильную тенденцию к восстановлению как скрытого разбиения, так и количества кластеров в нем. Сравнивается производительность этого метода с наиболее популярным методом кластеризации — кластеризацией по модулю. К сожалению, кластеризация по модульному принципу оказывается менее чем удовлетворительной и не должна применяться в качестве инструмента кластеризации по консенсусу. Другим вкладом данной работы является новый дизайн вычислительных экспериментов с консенсусными методами кластеризации. Вместо традиционных подходов к созданию ансамблей разделов, опосредованных наборами данных и применяемыми методами кластеризации, предлагается простой вероятностный механизм мутации для создания репрезентативного ансамбля разделов, разнообразие которого контролируется значением вероятности мутации. В эксперименты не включаются реальные наборы данных, такие как те, что находятся в знаменитом репозитории UC Irvine Machine Learning, поскольку нет прямых доказательств того, что признаки в этих наборах действительно связаны с истинными разделами. Будущая работа должна включать объяснение наблюдаемых странностей, разработку более реалистичных механизмов мутации и адаптацию подхода к большим наборам данных. Интересным направлением могут стать подходы, связанные с методами анализа формальных понятий (FCA) [16].

СПИСОК ЛИТЕРАТУРЫ

1. *Миркин Б.Г.* Об одном подходе к обработке нечисловых данных / Математические методы моделирования и решения экономических задач (Ред. К.А. Багриновский). Новосибирск, ИЭиОПП СО АН СССР, 1969. С. 141–150.
2. *Миркин Б.Г., Черный Л.Б.* Об измерении близости между различными разбиениями конечного множества объектов // АиТ. 1970. № 5. С. 120–127.
3. *Mirkin B.* Clustering: A Data Recovery Approach // Chapman and Hall, 2012. V. 19. https://doi.org/10.1201/9781420034912
4. *Миркин Б.Г., Мучник И.Б.* Геометрическая интерпретация показателей качества классификации / Методы анализа многомерной экономической информации (Ред. Б.Г. Миркин). Новосибирск. Наука, Сибирское отделение. 1981. С. 3–11.

5. *Strehl A., Ghosh J.* Cluster Ensembles — A Knowledge Reuse Framework for Combining Multiple Partitions // J. Machin. Learning Res. 2002. P. 583–617. https://doi.org/10.1162/153244303321897735
6. *Monti S., Tamayo P., Mesirov J., et al.* Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data // Machine Learning. 2003. P. 91–118. https://doi.org/10.1023/A:1023949509487
7. *Ünlü R., Xanthopoulos P.* Estimating the number of clusters in a dataset via consensus clustering // Expert Syst. Appl. 2019. https://doi.org/10.1016/j.eswa.2019.01.074
8. *Alguliyev R., Aliguliyev R., Sukhostat L.* An efficient algorithm for big data clustering on a single machine // CAAI Transactions on Intelligence Technology. 2020. https://doi.org/10.1049/trit.2019.0048
9. *Liu P., Zhang K., Wang P., et al.* A clustering-and maximum consensus-based model for social network large-scale group decision making with linguistic distribution // Inform. Sci. 2022. P. 269–297.
10. *Newman M.E.* Modularity and community structure in networks // Proc. Nation. Acad. Sci. 2006. P. 8577–8582.
11. *de Amorim R.C., Shestakov A., Mirkin B., et al.* The Minkowski central partition as a pointer to a suitable distance exponent and consensus partitioning // Patt. Recognit. 2017. P. 62–72.
12. *Blondel V.D., Guillaume J.L., Lambiotte R., et al.* Fast unfolding of communities in large networks // J. Statist. Mechan.:Theory Experiment. 2008. No. 10. P. 10008–10016.
13. *Brandes U., Delling D., Gaertler M., et al.* On modularity clustering // IEEE Transaction. Knowledge. 2007. P. 172–188.
14. *Fern X., Lin W.* Cluster ensemble selection // Statist. Anal. Data Mining: The ASA Data Sci. J. 2008. No. 1. P. 128–141. https://doi.org/10.1002/sam.10008
15. *Guénoche A.* Consensus of partitions: a constructive approach // Advances in Data Analysis and Classification. 2011. No. 5(3). P. 215–229.
16. *Hubert L.J., Arabie P.* Comparing partitions // J. Classifikat. 1985. No. 2. P. 193–218.
17. *Kovaleva E.V., Mirkin B.G.* Bisecting K-means and 1D projection divisive clustering: A unified framework and experimental comparison // J. Classifikat. 2015. P. 414–442.
18. *Murtagh F., Contreras P.* Algorithms for hierarchical clustering: an overview // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012. No. 32. P. 86–97.
19. *Pividori M., Stegmayer G., Milone D.H.* Diversity control for improving the analysis of consensus clustering // Inform. Sci. 2016. No. 361. P. 120–134.
20. *Gnatyshak D., Ignatov D.I., Mirkin B.G., et al.* A Lattice-based Consensus Clustering Algorithm // CLA. CEUR Workshop Proceedings. 2016. V. 1624. P. 45–56.

Статья представлена к публикации членом редколлегии А.А. Галяевым.

Поступила в редакцию 08.07.2023

После доработки 21.10.2023

Принята к публикации 20.01.2024