КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2024 Т. 16 № 3 С. 673–693
DOI: 10.20537/2076-7633-2024-16-3-673-693

Ки&М

АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ

УДК: 519.854.3, 616.8-07

Задачи и алгоритмы оптимальной кластеризации многомерных объектов по множеству разнородных показателей и их приложения в медицине

Ю. А. Мезенцев[1,a], О. М. Разумникова[1], И. В. Эстрайх[1], И. В. Тарасова[2], О. А. Трубникова[2]

[1]ФГБОУ ВО Новосибирский государственный технический университет,
Россия, 630073, г. Новосибирск, пр. К. Маркса, д. 20
[2]ФГБНУ НИИ комплексных проблем сердечно-сосудистых заболеваний,
Россия,650002, г. Кемерово, Сосновый бул., д. 6, корп. 2

E-mail: [a] mesyan@yandex.ru

Получено 15.10.2023, после доработки — 22.03.2024.
Принято к публикации 03.04.2024.

Работа посвящена описанию авторских формальных постановок задачи кластеризации при заданном числе кластеров, алгоритмам их решения, а также результатам применения этого инструментария в медицине.

Решение сформулированных задач точными алгоритмами реализаций даже относительно невысоких размерностей до выполнения условий оптимальности невозможно за сколько-нибудь рациональное время по причине их принадлежности к классу NP.

В связи с этим нами предложен гибридный алгоритм, сочетающий преимущества точных методов на базе кластеризации в парных расстояниях на начальном этапе с быстродействием методов решения упрощенных задач разбиения по центрам кластеров на завершающем этапе. Для развития данного направления разработан последовательный гибридный алгоритм кластеризации с использованием случайного поиска в парадигме роевого интеллекта. В статье приведено его описание и представлены результаты расчетов прикладных задач кластеризации.

Для выяснения эффективности разработанного инструментария оптимальной кластеризации многомерных объектов по множеству разнородных показателей был выполнен ряд вычислительных экспериментов с использованием массивов данных, включающих социально-демографические, клинико-анамнестические, электроэнцефалографические и психометрические данные когнитивного статуса пациентов кардиологической клиники. Получено экспериментальное доказательство эффективности применения алгоритмов локального поиска в парадигме роевого интеллекта в рамках гибридного алгоритма при решении задач оптимальной кластеризации. Результаты вычислений свидетельствуют о фактическом разрешении основной проблемы применения аппарата дискретной оптимизации — ограничения доступных размерностей реализаций задач. Нами показано, что эта проблема снимается при сохранении приемлемой близости результатов кластеризации к оптимальным.

Прикладное значение полученных результатов кластеризации обусловлено также тем, что разработанный инструментарий оптимальной кластеризации дополнен оценкой стабильности сформированных кластеров, что позволяет к известным факторам (наличие стеноза или старший возраст) дополнительно выделить тех пациентов, когнитивные ресурсы которых оказываются недостаточны, чтобы преодолеть влияние операционной анестезии, вследствие чего отмечается однонаправленный эффект послеоперационного ухудшения показателей сложной зрительно-моторной реакции, внимания и памяти. Этот эффект свидетельствует о возможности дифференцированно классифицировать пациентов с использованием предлагаемого инструментария.

Ключевые слова: оптимальная кластеризация, парные расстояния, центры кластеров, гибридный алгоритм, локальный поиск, роевой интеллект

COMPUTER RESEARCH AND MODELING
2024 VOL. 16 NO. 3 P. 673–693
DOI: 10.20537/2076-7633-2024-16-3-673-693

Ки&М

ANALYSIS AND MODELING OF COMPLEX LIVING SYSTEMS

UDC: 519.854.3, 616.8-07

Tasks and algorithms for optimal clustering of multidimensional objects by a variety of heterogeneous indicators and their applications in medicine

Yu. A. Mezentsev[1,a], O. M. Razumnikova[1], I. V. Estraykh[1], I. V. Tarasova[2], O. A. Trubnikova[2]

[1]Novosibirsk State Technical University,
20 K. Marx ave., Novosibirsk, 630073, Russia
[2]Research Institute of Complex Problems of Cardiovascular Diseases,
6/2 Sosnovy blvd., Kemerovo, 650002, Russia

E-mail: [a] mesyan@yandex.ru

Received 15.10.2023, after completion — 22.03.2024.
Accepted for publication 03.04.2024.

The work is devoted to the description of the author's formal statements of the clustering problem for a given number of clusters, algorithms for their solution, as well as the results of using this toolkit in medicine.

The solution of the formulated problems by exact algorithms of implementations of even relatively low dimensions before proving optimality is impossible in a finite time due to their belonging to the NP class.

In this regard, we have proposed a hybrid algorithm that combines the advantages of precise methods based on clustering in paired distances at the initial stage with the speed of methods for solving simplified problems of splitting by cluster centers at the final stage. In the development of this direction, a sequential hybrid clustering algorithm using random search in the paradigm of swarm intelligence has been developed. The article describes it and presents the results of calculations of applied clustering problems.

To determine the effectiveness of the developed tools for optimal clustering of multidimensional objects according to a variety of heterogeneous indicators, a number of computational experiments were performed using data sets including socio-demographic, clinical anamnestic, electroencephalographic and psychometric data on the cognitive status of patients of the cardiology clinic. An experimental proof of the effectiveness of using local search algorithms in the paradigm of swarm intelligence within the framework of a hybrid algorithm for solving optimal clustering problems has been obtained. The results of the calculations indicate the actual resolution of the main problem of using the discrete optimization apparatus — limiting the available dimensions of task implementations. We have shown that this problem is eliminated while maintaining an acceptable proximity of the clustering results to the optimal ones.

The applied significance of the obtained clustering results is also due to the fact that the developed optimal clustering toolkit is supplemented by an assessment of the stability of the formed clusters, which allows for known factors (the presence of stenosis or older age) to additionally identify those patients whose cognitive resources are insufficient to overcome the influence of surgical anesthesia, as a result of which there is a unidirectional effect of postoperative deterioration of complex visual-motor reaction, attention and memory. This effect indicates the possibility of differentiating the classification of patients using the proposed tools.

Keywords: optimal clustering, paired distances, cluster centers, hybrid algorithm, local search, swarm intelligence

Citation: *Computer Research and Modeling*, 2024, vol. 16, no. 3, pp. 673–693 (Russian).

Введение

Задачи кластеризации входят в число наиболее востребованных практикой задач оптимизации и имеют как самостоятельное значение для классификаций различного рода объектов, так и сопряженное при распознавании образов, прогнозировании и решении других задач, относимых в настоящее время к проблематике искусственного интеллекта. Прикладной аспект кластеризации превалирует и в данной статье. Формально кластеризация представляет собой автоматическое разбиение множеств объектов любой природы на подмножества (пересекающиеся или непересекающиеся) по множествам характеризующих их признаков. Обычно из нормализованных значений этих признаков формируются метрики, являющиеся мерами близости между классифицируемыми объектами. Актуальность темы обусловлена принадлежностью большинства задач кластеризации к разряду труднорешаемых, в теории сложности называемых также NP-трудными. Для практики это означает, что оптимальные решения подобного рода задач, реализаций актуальных размерностей, требуют фактически бесконечного времени вычислений компьютеров любой перспективной производительности.

Основным недостатком распространенных постановок и алгоритмов решения задач оптимальной кластеризации является непосредственное применение эвристических алгоритмов, которые не только не гарантируют оптимальности разбиений исходных множеств на кластеры, но и могут приводить к сколь угодно плохим результатам. Решение же точными алгоритмами реализаций задачи даже относительно невысокой размерности до доказательства оптимальности невозможно по причине принадлежности к классу NP.

Модели и алгоритмы кластеризации, представленные в статье, относятся к одному из разделов дискретной оптимизации и могут быть описаны средствами смешанного целочисленного программирования (далее — milp, от mixed integer linear programming). Показана принадлежность сформулированных milp к классу NP, приведены эффективные алгоритмы их приближенного решения. Под эффективностью алгоритма понимается полиномиальная зависимость трудоемкости вычислений от размерности задачи. Описаны подробности формальных постановок и алгоритмов решения milp-кластеризации по парным расстояниям между объектами (кластеризация в парных расстояниях, далее — КПР) и по центрам формируемых кластеров (кластеризация по центрам кластеров, далее — КЦК), а также детализирован прикладной аспект применения разработанного аппарата в медицине. Основной мотивацией авторов явились актуальность и неразрешенность проблематики кластеризации как в целом, так и в приложениях к медицине, несмотря на наличие множества публикаций по этой теме, включая работы недавнего и текущего времени, например [Кельманов, Пяткин, 2013; Eremeev et al., 2019; Pyatkin, 2023].

Рассмотренные в статье модели и алгоритмы оптимальной кластеризации использованы для определения подгрупп пациентов по множеству демографических, клинико-анамнестических и психометрических показателей скорости реакции, внимания и памяти для прогноза послеоперационной когнитивной дисфункции (ПОКД) у пациентов с сердечно-сосудистыми заболеваниями (ССЗ) и предложения соответствующих индивидуальным особенностям методов активации когнитивных ресурсов.

Разобраны особенности организации психометрических и клинико-демографических показателей в двух выделенных с применением разработанного аппарата кластерах, сформированных на основе до- и послеоперационных (операции коронарного шунтирования, далее — КШ) показателей когнитивных функций. Для этого весь массив данных также проанализирован с использованием иерархического кластерного анализа методом Уорда, определившего лучшую группировку параметров.

Постановка задачи оптимальной кластеризации многомерных объектов по множеству показателей

Данный раздел посвящен описанию авторской постановки задачи кластеризации при заданном числе кластеров, алгоритмам и программным средствам ее решения. Основной особенностью постановки является ориентация на минимизацию суммарных расстояний между всеми объектами, включаемыми в кластер. В отличие от стандартного подхода, используемого в методе k-центров (например, [Айвазян и др., 1989]), результаты применения описываемого инструментария не зависят от выбора начальных приближений к решению. С одной стороны, его применение приводит к усложнению постановки с образованием NP-трудной задачи и соответствующих алгоритмов, но, с другой стороны, позволяет получать наиболее качественные разбиения на кластеры. При этом, как показывают вычислительные эксперименты, подход совместим с методом k-центров для поиска начального приближения. А последовательное применение обоих подходов с обучающей выборкой для первого снимает проблему размерности при сохранении качества разбиений. Подробности подхода описаны ниже в следующем разделе.

Приведем формальную постановку задачи оптимальной кластеризации.

Будем именовать ее оптимальной **m**-кластеризацией (разбиением множества объектов в метрическом пространстве на **m** кластеров) для общего случая. Введем обозначения: i, $j = \overline{1, n}$ — номера объектов, l, $k = \overline{1, m}$ — номера кластеров. Через $p_{i,t}$ обозначим t-й нормализованный показатель ($t = \overline{1, T}$), характеризующий объект i, $i = \overline{1, n}$; через $p^k_{i,t}$ — тот же объект в кластере k. Нормализация осуществляется в соответствии с выражениями

$$p_{i,t} = \alpha_t \frac{\widetilde{p}_{i,t} - \widetilde{p}_{\min,t}}{\widetilde{p}_{\max,t} - \widetilde{p}_{\min,t}}$$

либо

$$p_{i,t} = \alpha_t \frac{\widetilde{p}_{\max,t} - \widetilde{p}_{i,t}}{\widetilde{p}_{\max,t} - \widetilde{p}_{\min,t}}$$

(в зависимости от того, что по содержательному смыслу лучше — меньшие значения показателя или большие). Здесь $\widetilde{p}_{i,t}$ — исходное значение t-го показателя, $\widetilde{p}_{\min,t}$ — минимальное значение t-го показателя, $\widetilde{p}_{\max,t}$ — максимальное значение t-го показателя, $\alpha_t \geqslant 0$ — весовой коэффициент для t-го показателя. Заметим, что тривиальный случай равенства $\widetilde{p}_{\max,t} = \widetilde{p}_{\min,t}$ означает равенство всех значений $\widetilde{p}_{i,t}$ для объекта i, $t = \overline{1, T}$. При таких обстоятельствах все значения $p_{i,t}$ также приравниваются константе из диапазона (0, 1], например весовому коэффициенту α_t.

Через $c_{i,j}$ обозначим расстояния между объектами i и j, $i, j = \overline{1, n}$; $c^k_{i,j}$ — расстояния между объектами i и j в кластере k, например евклидово:

$$c^k_{i,j} = \left(\sum_{t=1}^{T} \left(p^k_{i,t} - p^k_{j,t} \right)^2 \right)^{1/2}.$$

Определим переменные y^k_i (идентифицирующие принадлежность объектов i, $j = \overline{1, n}$ кластеру k, $k = \overline{1, m}$) и зависимые переменные $x^k_{i,j} = y^k_i \cdot y^k_j$, $i = \overline{1, n}$.

Тогда задача кластеризации состоит в определении булевых переменных $y^k = \left\| y^k_i \right\|$ и $x^k = \left\| x^k_{i,j} \right\|$ при выполнении ряда условий.

Условия выбора:

$$y_i^k = \begin{cases} 1, \text{ если объект } i \text{ принадлежит кластеру } k, \\ 0 \text{ — в противном случае,} \end{cases} \quad i = \overline{1,\, n}, \tag{1}$$

$$\sum_{k=1}^{m} y_i^k = 1, \quad i = \overline{1,\, n}, \tag{2}$$

линеаризующие посредством замены $x_{i,j}^k = y_i^k \cdot y_j^k$, $i = \overline{1,\, n}$, неравенства

$$0 \leqslant y_i^k + y_j^k - 2x_{i,j}^k \leqslant 1, \quad k = \overline{1,\, m}, \quad i,\, j = \overline{1,\, n}, \quad i \neq j,$$

при несимметричной матрице расстояний, преобразуемые в

$$0 \leqslant y_i^k + y_j^k - x_{i,j}^k - x_{j,i}^k \leqslant 1, \quad k = \overline{1,\, m}, \quad i,\, j = \overline{1,\, n}, \quad i \neq j, \tag{3}$$

$$x_{i,j}^k = \begin{cases} 1, \text{ если кластеру } k \text{ принадлежат объекты } i,\, j\colon y_i^k = 1,\ y_j^k = 1, \\ 0 \text{ — в противном случае,} \end{cases} \quad i,\, j = \overline{1,\, n}, \quad i \neq j. \tag{4}$$

Добавим в задачу условия, реализующие минимаксный критерий:

$$\sum_{j=1}^{n} \sum_{i=1}^{n} c_{i,j} x_{i,j}^k \leqslant \lambda, \quad k = \overline{1,\, m}, \quad i \neq j, \quad \lambda \to \min, \tag{5}$$

имеющий смысл минимизации максимальной по всем кластерам суммы расстояний между всеми объектами каждого кластера.

В совокупности условия (1)–(5) являются вариантом формализации задачи оптимальной кластеризации. Минимаксный критерий (5) эффективен при разбиении исходного множества на максимально однородные подмножества с одновременной минимизацией сумм расстояний между объектами.

Кроме критерия (5), в зависимости от содержательного смысла задачи кластеризации, в ряде случаев более приемлемым является аддитивный критерий, который удобно представлять в виде

$$\sum_{j=1}^{n} \sum_{i=1}^{n} c_{i,j} x_{i,j}^k = \lambda^k, \quad k = \overline{1,\, m}, \quad i \neq j, \tag{6}$$

$$\sum_{k=1}^{m} \lambda^k \to \min. \tag{7}$$

Здесь λ^k — сумма расстояний между всеми парами объектов в кластере k, $k = \overline{1,\, m}$.

Решение варианта задачи **m**-кластеризации (1)–(4), (6), (7) позволяет находить разбиения множества объектов с заданными расстояниями между всеми парами объектов на заданное число (m) подмножеств (кластеров), которое гарантирует минимизацию суммы минимальных суммарных расстояний между всеми парами объектов по всем кластерам.

В совокупности условия (6) и (7) реализуют аддитивный критерий кластеризации, наиболее часто используемый для разбиения множеств объектов на максимально непохожие (отдаленные друг от друга по суммам расстояний между объектами) подмножества.

Формального доказательства NP-трудности представленных постановок задач нами не приводится, поскольку это опосредует сведение представленных постановок к любой известной задаче с NP-полнотой, что само по себе может оказаться труднорешаемой задачей. Заметим

только, что при значительном упрощении ограничивающих условий любой из приведенных выше задач получается NP-трудная задача смешанного целочисленного программирования. В частности подзадача (1)–(2), (5) интерпретируется как NP-трудная задача оптимизации расписаний несвязанных параллельных машин по критерию $C_{\max}$. Доказательство ее NP-трудности можно найти, например, в работах [Pinedo, 2008; Lenstra, Shmoys, Tardos, 1987]. С учетом же остальных условий сложность представленных задач увеличивается на много порядков, что показывают вычислительные эксперименты на реальных данных.

Следует отметить, что для сокращения трудоемкости обеих постановок milp можно заменить зависимые булевы переменные $x_{i,j}^k$ (условие (4)) непрерывными:

$$0 \leqslant x_{i,j}^k \leqslant 1, \quad k = \overline{1, m}, \quad i, j = \overline{1, n}, \quad i \neq j. \tag{8}$$

Гипотеза об эквивалентности задач оптимальной **m**-кластеризации (1)–(5); (1)–(3), (5), (8); (1)–(4), (6), (7) и (1)–(3), (6)–(8) соответственно отражена в работе авторов [Мезенцев и др., 2019].

Данное утверждение получило практическое подтверждение путем многократного применения в расчетах при полном отсутствии опровергающих результатов. Ее теоретическая база основана на следующих соображениях. Рассмотрим соотношения $x_{i,j}^k = y_i^k \cdot y_j^k$ и эквивалентные условия (3)

$$0 \leqslant y_i^k + y_j^k - x_{i,j}^k - x_{j,i}^k \leqslant 1, \quad k = \overline{1, m}, \quad i, j = \overline{1, n}, \quad i \neq j,$$

означающие, что $x_{i,j}^k$ истинны только тогда, когда истинны y_i^k и y_j^k. Соответственно, y_i^k и y_j^k истинны одновременно только тогда, когда истинны $x_{i,j}^k$. Заменим условие целочисленности $x_{i,j}^k$ (4) на условие $0 \leqslant x_{i,j}^k \leqslant 1$ (8) и рассмотрим возможные варианты соотношений (3): если в оптимальном решении $y_i^k = 1$ и $y_j^k = 1$, то соотношение $x_{i,j}^k = y_i^k \cdot y_j^k$ выполнится только в случае $x_{i,j}^k = x_{j,i}^k = 1$; если в оптимальном решении $y_i^k = 1$ и $y_j^k = 0$, то соотношение $x_{i,j}^k = y_i^k \cdot y_j^k$ выполнится только в случае $x_{i,j}^k = x_{j,i}^k = 0$. Совершенно аналогично $x_{i,j}^k = x_{j,i}^k = 0$ для случая $y_i^k = 0$ и $y_j^k = 0$. Поэтому если условия $x_{i,j}^k = y_i^k \cdot y_j^k$ и $0 \leqslant y_i^k + y_j^k - x_{i,j}^k - x_{j,i}^k \leqslant 1$, $k = \overline{1, m}$, $i, j = \overline{1, n}$, $i \neq j$, действительно эквивалентны при любых y_i^k для бинарных $x_{i,j}^k$, то они эквивалентны для любых $0 \leqslant x_{i,j}^k \leqslant 1$.

Таким образом, решения релаксаций (1)–(3), (5), (8) и (1)–(3), (6)–(8) совпадают с решениями задач в исходных постановках (1)–(5) и (1)–(4), (6), (7) соответственно.

Такая замена условий приводит к снижению числа булевых переменных в релаксированных задачах на величину $m \cdot n^2$. Общее число булевых переменных (1) в задачах (1)–(3), (5), (8) и (1)–(3), (6)–(8) составляет величину $m \cdot n$ при наличии $m \cdot n^2 + 1$ непрерывных переменных (7) и (8), против $m \cdot n(1 + n)$ булевых переменных в задачах (1)–(5) и (1)–(4), (6), (7). Разница весьма существенна при применении в практических приложениях представленных формальных задач.

Поскольку перспективы разработки приемлемых по точности аппроксимационных эффективных алгоритмов для сформулированных задач не определены из-за принадлежности к классу NP, применим для их решения условно экспоненциальные алгоритмы, а также алгоритмы локального поиска, успешность практического применения которых сильно зависит от фактического числа целочисленных переменных. В этом смысле релаксации (1)–(3), (5), (8) и (1)–(3), (6)–(8) имеют существенные преимущества перед постановками (1)–(5) и (1)–(4), (6), (7).

Трудоемкость задач оптимальной кластеризации многомерных объектов в КПР-постановке и средства решения

В работе [Мезенцев и др., 2019] показана принадлежность обеих постановок milp (1)–(3), (5), (8) и (1)–(3), (6)–(8) классу NP. Это означает, что для представленных выше вариантов за-

дач кластеризации не существует теоретически эффективных алгоритмов. Однако для практических применений в достаточной степени разработаны алгоритмы, которые можно именовать как условно экспоненциальные. Несмотря на недоказанность их эффективности, данные алгоритмы позволяют за разумное время находить оптимальные (либо приближенные к оптимальным) решения представленных задач дискретной оптимизации. Примером такого алгоритма может служить алгоритм бинарных отсечений и ветвлений [Mezentsev, 2017]. В качестве стандартных средств решения могут применяться также модули Gurobi optimization либо IBM CPLEX optimization studio. Средства последнего применены для поиска решения сформулированных выше задач кластеризации (1)–(3), (5), (8) и (1)–(3), (6)–(8). В частности, для решения обеих задач использованы язык OPL (optimization programming language) и вычислительные модули CPLEX, реализующие барьерный метод в качестве оценочного и алгоритмы смешанного программирования (ветвей и отсечений, ветвей и границ). Ниже приводится содержательная интерпретация результатов обработки данных медико-психологических исследований на основе формальных постановок (1)–(3), (5), (8) и (1)–(3), (6)–(8) средствами IBM CPLEX optimization studio.

Применение инструментария КПР привело к обнадеживающим практическим результатам. Для ряда частных случаев [Разумникова и др., 2021] удалось достичь приемлемого быстродействия с оценкой отклонений от оптимумов, не превышающей 3 %. Однако число булевых переменных (классифицируемых объектов с учетом числа классов) в работе [Разумникова и др., 2021] не превышало 400. Эта же оценка получена посредством множества других практических расчетов с использованием персональных компьютеров средней производительности [Авдеенко, Мезенцев, 2020]. В общем же случае проблема размерности для множества прикладных задач кластеризации остается актуальной. Поэтому важнейшей задачей при применении постановок КПР является снижение размерностей решаемых задач, а также, как показано ниже, широкое применение алгоритмов локального поиска, роевого интеллекта в частности. Возможные пути — применение методов декомпозиции, релаксации, применение гибридных моделей. Например, в работе [Авдеенко, Мезенцев, 2020] декомпозиционный подход применялся для 10-кластеризации 120 объектов. Это породило NP-трудную задачу КПР, содержащую 1200 булевых, 144 000 непрерывных переменных, при соответствующем выражениям (1)–(3), (6), (8) числе ограничений. В результате прямое применение методов ветвей и отсечений к порожденной задаче КПР оказалось возможным только с ограничениями на время счета, что привело к некоему результату, близость которого к оптимальному оказалось невозможно оценить. В противном случае получение точного решения требует фактически бесконечного времени. В этой же работе предложена иерархическая декомпозиция задачи, приводящая к существенному снижению размерности, что позволило улучшить исходное решение примерно на 20 % в терминах критерия качества (6)–(7) при приемлемом времени счета.

Связь КПР с другими постановками и возможности расширения применений

Координаты оптимального решения задачи КПР (1)–(3), (6)–(8) обозначим через $\widetilde{y}^k = \left\|\widetilde{y}_i^k\right\|$, $\widetilde{x}^k = \left\|\widetilde{x}_{i,j}^k\right\|$, $\widetilde{\lambda}^k$, через $\widetilde{n}^k$ — число элементов кластера k в оптимальном решении, вычислим координаты центров кластеров $\widetilde{z}^k$ для этих решений:

$$\widetilde{z}^k = \frac{1}{\widetilde{n}^k} \sum_{i=1}^{n} \sum_{t=1}^{T} p_{i,t}^k \widetilde{y}_i^k.$$

В постановках КПР (1)–(3), (5), (8) и (1)–(3), (6)–(8) применена метрика λ^k — сумма расстояний между всеми парами объектов во всех кластерах k, $k = \overline{1, m}$. Довольно часто для фор-

мирования критерия качества кластеризации используются меры расстояний объектов классифицируемого множества до центров кластеров. Примером могут служить суммы квадратов евклидовых расстояний до центров кластеров с максимизацией расстояний между центрами, которые реализованы в методе k-центров и в кластеризации на основе нейронной сети Кохонена, реализованных в Statistica 12 Advanced.

Для корректного сравнения результатов кластеризации необходимо сопоставить оценки качества разбиений по обоим критериям и с применением обоих подходов.

Обозначим:

z^k — координата (неизвестная) центра кластера k, $k = \overline{1, m}$;

$z^k = \left\|z_t^k\right\|$, $t = \overline{1, T}$ (T — размерность вектора описания объекта);

$p_i = \left\|p_{i,t}\right\|$ — координата t ($t = \overline{1, T}$) T-мерного объекта i, $i = \overline{1, n}$;

n^k — число элементов в кластере k, $k = \overline{1, m}$.

При этом, очевидно, для любого кластера $p_{i,t}^k = p_{i,t}$, $k = \overline{1, m}$,

$r_{i,t}^k = p_{i,t} - z_t^k$ — координата t вектора r_i^k с началом в центре кластера k;

$r_i^k = \left\|r_{i,t}^k\right\|$ — вектор i с началом в центре кластера k, $k = \overline{1, m}$, $i = \overline{1, n}$.

$$r_i^k = \left(\sum_{t=1}^{T}\left(p_{i,t} - z_t^k\right)^2\right)^{1/2} = \left(\sum_{t=1}^{T} {r_{i,t}^k}^2\right)^{1/2}, \tag{9}$$

r_i^k интерпретируется как евклидово расстояние объекта i до центра кластера k.

Опираясь на введенные обозначения, сформулируем следующую задачу кластеризации.

Найти y_i^k, $i = \overline{1, n}$, $k = \overline{1, m}$, и z_t^k, $k = \overline{1, m}$, $t = \overline{1, T}$:

$$\sum_{i=1}^{n}\sum_{t=1}^{T}\left(p_{i,t} - z_t^k\right)^2 y_i^k \leqslant \beta^k, \quad \sum_{k=1}^{m}\beta^k \to \min \tag{10}$$

при условиях

$$y_i^k = \begin{cases} 1, \text{ если объект } i \text{ принадлежит кластеру } k, \\ 0 \text{ — в противном случае,} \end{cases} \quad i = \overline{1, n}, \tag{11}$$

$$\sum_{k=1}^{m} y_i^k = 1, \quad i = \overline{1, n}, \tag{12}$$

$$z_t^k \geqslant 0, \quad k = \overline{1, m}, \quad t = \overline{1, T}. \tag{13}$$

В ряде случаев возможно дополнение задачи ограничением на числа элементов в кластерах:

$$\sum_{i=1}^{n} y_i^k \geqslant n^k, \quad k = \overline{1, m}. \tag{14}$$

Представленная формальная постановка также является задачей оптимальной кластеризации. При этом, в отличие от КПР, используется иной принцип разбиений, ориентированный на центры кластеров. Он реализован в методе k-центров и описан во множестве работ, посвященных проблемам классификации и кластеризации. Как отмечалось выше, сформулированная задача (9)–(14) и основанные на ней модификации именуются кластеризацией по центрам кластеров (КЦК). Известно, что задача (10)–(13) является NP-полной. Для нее не существует аппроксимационных алгоритмов с априорными оценками близости решений к оптимальным. Существует

множество эвристических алгоритмов, к каковым относятся и упомянутый метод k-центров, ряд разновидностей локального поиска и кластеризация на основе нейронных сетей.

Вместе с тем нам удалось выявить тесную связь задач КПР и КЦК, близость оптимумов КЦК к оценкам центров кластеров при оптимальных решениях КПР, благодаря чему естественным развитием темы оптимальной кластеризации явилось совместное применение обоих подходов.

Рассмотрим связь задачи (10)–(13) и задачи КПР (1)–(3), (6)–(8), а также соотношение координат оптимальных решений любых реализаций задач кластеризации в обеих постановках.

Пусть известно оптимальное решение задачи (10)–(13) $\widehat{z}^k$, $\widehat{y}^k$. Покажем связь $\widehat{y}^k$ с оптимальным решением задачи КПР (1)–(3), (6)–(8) $\overline{y}^k = \left\|\overline{y}_i^k\right\| \left(\overline{x}^k = \left\|\overline{x}_{i,j}^k\right\|\right)$.

Расширенная задача КПР и связь с задачей КЦК

Если вычисленные постфактум центры кластеров $z^k \left(\overline{z}^k = \frac{1}{\overline{n}^k} \sum\limits_{i=1}^{n} \sum\limits_{t=1}^{T} p_{i,t}^k \overline{y}_i^k\right)$ ввести в состав объектов кластеризации, то вполне очевидно, что при повторном КПР-разбиении они войдут в состав элементов соответствующих кластеров. А центры кластеров $\overline{z}^k$ от этого расширения не изменятся. Такую задачу КПР будем именовать расширенной.

В этом контексте координаты дополняющих объектов (для которых положим $i = 0$) и парные расстояния между исходными и дополняющими объектами обозначим как $p_0^k = \overline{z}^k$, $c_{i,0}^k = r_i^k$, $k = \overline{1, m}$.

Отметим свойства оптимальных решений расширенных КПР, которые по аналогии с исходной КПР (1)–(3), (6)–(8) обозначим как $\widetilde{\overline{y}}^k = \left\|\widetilde{\overline{y}}_i^k\right\| \left(\widetilde{\overline{x}}^k = \left\|\widetilde{\overline{x}}_{i,j}^k\right\|\right)$, $i,\ j = \overline{0, \overline{n}^k}$, $k = \overline{1, m}$. Вполне очевидно, что $\widetilde{\overline{y}}_i^k = \overline{y}_i^k$, $\widetilde{\overline{x}}_{i,j}^k = \overline{x}_{i,j}^k$, $i,\ j = \overline{1, \overline{n}^k}$, $\widetilde{\overline{z}}^k = \overline{z}^k = \frac{1}{\overline{n}^k} \sum\limits_{i=0}^{n} \sum\limits_{t=1}^{T} p_{i,t}^k \widetilde{\overline{y}}_i^k$, $k = \overline{1, m}$, то есть решения обеих КПР совпадают для всех $i,\ j = \overline{1, \overline{n}^k}$, $k = \overline{1, m}$.

Задача о минимальном реберном покрытии графа расширенной задачи КПР

Все расстояния между вычисленными центрами кластеров $\overline{z}^k \left(\widehat{z}^k\right)$ можно положить равными бесконечности $\left(c_{0,0}^k = \infty\right)$ и в дальнейшем из рассмотрения исключить.

Сформулируем следующую задачу минимального реберного покрытия полного графа исходной задачи КПР:

$$\Lambda_0 = \sum_{k=1}^{m} \sum_{i=1}^{n} c_{i,0}^k x_{i,0}^k \to \min, \tag{15}$$

$$\sum_{k=1}^{m} x_{i,0}^k = 1, \quad i = \overline{1, n}, \tag{16}$$

$$\sum_{i=1}^{n} x_{i,0}^k \geqslant n^k, \quad k = \overline{1, m}, \tag{17}$$

$$x_{i,0}^k = \begin{cases} 1, \text{ если объект } i \text{ принадлежит кластеру } k, \\ 0 \text{ — в противном случае,} \end{cases} \quad i = \overline{1, n}. \tag{18}$$

Задача полиномиально разрешима, поскольку является частным случаем задачи о назначениях, в соответствии с чем огораничения на переменные можно заменить на

$$0 \leqslant x_{i,0}^k \leqslant 1, \quad k = \overline{1,\, m}, \quad i = \overline{1,\, n}. \tag{19}$$

Таким образом, с одной стороны, фактически получена задача (15)–(17), (19), эквивалентная приведенной выше (10)–(14) при известных центрах кластеров $\widehat{z}^k$, где $y_i^k = x_{i,0}^k$, $k = \overline{1,\, m}$, $i = \overline{1,\, n}$.

С другой стороны, задачу (15)–(17), (19) можно рассматривать как частный случай задачи КПР (1)–(3), (6)–(8) при замене определения m клик с минимальным суммарным расстоянием определением m минимальных реберных покрытий этих клик.

Обозначим также через $\widehat{x}_{i,0}^k$, $k = \overline{1,\, m}$, $i = \overline{1,\, n}$, $\widehat{\Lambda}_0 = \sum\limits_{k=1}^{m} \sum\limits_{i=1}^{n} c_{i,0}^k \widehat{x}_{i,0}^k$ оптимальное решение и его оценку для задачи (15)–(17), (19). Учитывая приведенные выше обозначения (10), можно также записать: $\widehat{\Lambda}_0 = \sum\limits_{k=1}^{m} \widehat{\beta}^k$, где $\widehat{\beta}^k$ — вычисляемая оценка кластера k.

Отметим соотношения решений задач (1)–(3), (6)–(8) и (15)–(19) по критериям (6)–(7) и (15).

1. Допустимость. $\widehat{y}_i^k = \widehat{x}_{i,0}^k$, $k = \overline{1,\, m}$, $i = \overline{1,\, n}$, являются координатами допустимого решения задачи (1)–(3), (6)–(8) в силу выполнения ограничений (1), (2). Тогда условия (3) и (8) обеспечивают выполнение равенств $\widehat{x}_{i,j}^k = \widehat{y}_i^k \cdot \widehat{y}_j^k$, $k = \overline{1,\, m}$, $i,\, j = \overline{1,\, n}$, $i \neq j$, что означает выполнение всех ограничений (1)–(3), (8) для значений $\widehat{y}_i^k$, $\widehat{x}_{i,j}^k$, $k = \overline{1,\, m}$, $i,\, j = \overline{1,\, n}$, $i \neq j$.

2. Оптимальность. Покажем, что на этих допустимых решениях при выполнении ряда условий достигается минимум целевой функции (6), (7). Для этого рассмотрим модифицированную целевую функцию (6)–(7) Λ в сравнении с Λ_0:

$$\widetilde{\Lambda} = \sum_{k=1}^{m} \sum_{j=1}^{n} \sum_{i=1}^{n} c_{i,j}^k \widetilde{x}_{i,j}^k, \quad \widehat{\Lambda} = \sum_{k=1}^{m} \sum_{j=1}^{n} \sum_{i=1}^{n} c_{i,j}^k \widehat{x}_{i,j}^k, \quad \widehat{\Lambda}_0 = \sum_{k=1}^{m} \sum_{i=1}^{n} c_{i,0}^k \widehat{x}_{i,0}^k, \quad \widetilde{\Lambda}_0 = \sum_{k=1}^{m} \sum_{i=1}^{n} c_{i,0}^k \widetilde{x}_{i,0}^k.$$

Отмечаем, что $\widetilde{\Lambda} = \widehat{\Lambda}$ и $\widetilde{\Lambda}_0 = \widehat{\Lambda}_0$ при $\widetilde{x}_{i,j}^k = \widehat{x}_{i,j}^k$ и $\widetilde{x}_{i,0}^k = \widehat{x}_{i,0}^k$. Обратное тоже верно с допущением неединственности оптимальных решений.

В то же время, учитывая линейную связь значений $c_{i,0}^k$ с $\widetilde{x}_{i,j}^k$, $k = \overline{1,\, m}$, $i,\, j = \overline{1,\, n}$, $i \neq j$, можно предположить наличие тесной связи между оптимальными решениями обеих задач оптимальной кластеризации (КПР и КЦК) $\widetilde{y}_i^k$ и $\widehat{x}_{i,0}^k$, $k = \overline{1,\, m}$, $i = \overline{1,\, n}$, что подтверждается экспериментально. Это не означает равенства оптимальных решений в общем случае, это означает близость оптимальных решений, измеряемую в терминах критериев качества $\widetilde{\Lambda}$ и $\widehat{\Lambda}$, $\widetilde{\Lambda}_0$ и $\widehat{\Lambda}_0$. Из данного постулата непосредственно вытекает:

1) корректность сравнений результатов решений обеих задач КПР и КЦК,

2) возможность применения линеаризованной постановки КПР (1)–(3), (6)–(8) с получением исходных значений координат центров кластеров по обучающей выборке для эффективного приближенного решения задач кластеризации больших данных.

Последовательный гибридный алгоритм кластеризации

Основным недостатком используемых постановок КЦК вида (10)–(14) является непосредственное применение известных эвристических алгоритмов (включая алгоритм k-центров), которые не только не гарантируют оптимальности разбиений исходных множеств на кластеры, но и могут приводить к сколь угодно плохим результатам.

В связи с этим нами предложен гибридный алгоритм, сочетающий преимущества точных методов на базе КПР на начальном этапе с быстродействием методов решения упрощенных задач КЦК вида, представленного выше, на завершающем этапе. Пользуясь отмеченным выше свойством задачи КПР (1)–(3), (6)–(8) находить близкие к оптимальным решения реализаций этих задач, содержащих до 400 переменных y_i^k, а также выявленной связью задач КПР и КЦК [Мезенцев и др., 2019], можно предложить алгоритм с двухэтапной схемой вычислений. На первом этапе по любой обучающей выборке исходных данных решается подзадача КПР и по полученным решениям вычисляются центры кластеров. Далее при известных оценках центров решается подзадача КЦК (10)–(14) (класса ЛП) для всей выборки. Соответствующая подзадача становится задачей ЛП, которая полиномиально разрешима, что также показано ранее в работе [Разумникова и др., 2021].

Таким образом, с одной стороны, снимается проблема размерности КПР, которая NP-трудна; с другой стороны, разрешается проблема точности КЦК. Поэтому соответствующий последовательный гибридный алгоритм (ПГА) эффективен. Его практические применения и апостериорный анализ решений доказывают вычислительную эффективность и достаточную для практических нужд близость к оптимумам результатов кластеризации как по критерию КПР (6), (7), так и КЦК (10). Приводим запись ПГА.

Алгоритм A_e (разбиения множества n объектов на m подмножеств)

Ввод исходных данных. Определение параметров алгоритма: размера обучающей выборки n_v, числа кластеров m; обнуление начальных значений центров кластеров реализации исходной задачи КЦК (10)–(14). Задание начальных значений критериев качества кластеризации:

$$\widetilde{\Lambda} = 0, \quad \widehat{\Lambda} = 0, \quad \widetilde{\Lambda}_0 = 0, \quad \widehat{\Lambda}_0 = 0.$$

1. Выборка обучающего подмножества n_v объектов из исходного множества n, которые пронумеруем от 1 до n_v ($i = \overline{1, n_v}$). При этом число связанных булевых переменных y_i^k, равное $m \cdot n_v$, должно удовлетворять определенным выше ограничениям прямых алгоритмов решения реализаций задач КПР (1)–(3), (6)–(8). Репрезентативность обучающей выборки можно обеспечить формированием подмножества с использованием некоторой сетки, накладываемой на исходное множество, равномерной например.

2. Формирование реализации подзадачи КПР (1)–(3), (6)–(8) с характеристиками $y_i^k \in \{0, 1\}$, $0 \leqslant x_{i,j}^k \leqslant 1$, $k = \overline{1, m}$, $i, j = \overline{1, n_v}$, $i \neq j$, $\Lambda_v = \sum\limits_{k=1}^{m} \sum\limits_{j=1}^{n_v} \sum\limits_{i=1}^{n_v} c_{i,j}^k x_{i,j}^k \to \min$; определение оптимальных решений либо приближений к оптимальным решениям любым известным алгоритмом [Mezentsev, 2017], например, которые обозначим как $\widetilde{y}_i^k$, $\widetilde{x}_{i,j}^k$, $k = \overline{1, m}$, $i, j = \overline{1, n_v}$, $i \neq j$, $\widetilde{\Lambda}_v = \sum\limits_{k=1}^{m} \sum\limits_{j=1}^{n_v} \sum\limits_{i=1}^{n_v} c_{i,j}^k \widetilde{x}_{i,j}^k$.

3. Вычисление центров кластеров в обучающей выборке в соответствии с $\widetilde{z}_{vt}^{k} = \frac{1}{\widetilde{n}_v^k} \sum\limits_{i=1}^{n_v} p_{i,t}^{k} \widetilde{y}_i^k$, $t = \overline{1, T}$, $k = \overline{1, m}$, $\widetilde{z}_v^k = \frac{1}{\widetilde{n}_v^k} \sum\limits_{i=1}^{n_v} \sum\limits_{t=1}^{T} p_{i,t}^{k} \widetilde{y}_i^k$, где $\widetilde{n}_v^k$ — число элементов в кластере k, $k = \overline{1, m}$, обучающей выборки.

4. Задание начальных значений координат центров кластеров реализации задачи КЦК (10)–(14): $\widehat{z}^k = \widetilde{z}_v^k$, $k = \overline{1, m}$, номера итерации алгоритма $l = 0$, присваивание $\widehat{\Lambda}_0^l := \widehat{\Lambda}_0$.

5. Увеличение номера шага $l := l + 1$.

6. Вычисление новых значений $c_{i,0}^k$, $k = \overline{1, m}$, $i = \overline{1, n}$. Решение полиномиально разрешимой задачи (15)–(17), (19). Фиксация результатов $\widehat{x}_{i,0}^k$, $k = \overline{1, m}$, $i = \overline{1, n}$, $\widehat{\Lambda}_0^l = \sum\limits_{k=1}^{m} \sum\limits_{i=1}^{n} c_{i,0}^k \widehat{x}_{i,0}^k$. Вычисление новых значений центров кластеров $\widehat{z}_t^k = \frac{1}{\widehat{n}^k} \sum\limits_{i=1}^{n_v} p_{i,t}^k \widehat{x}_{i,0}^k$, $t = \overline{1, T}$, $k = \overline{1, m}$.

7. Сравнение значений $\widehat{\Lambda}_0$ на текущем и предшествующем шаге. Если $\widehat{\Lambda}_0^l - \widehat{\Lambda}_0^{l-1} = 0$, переход к п. 8. Иначе — возврат к п. 6.

8. Найдено решение исходной задачи $\widehat{y}_i^k \left(\widehat{x}_{i,0}^k\right)$, $k = \overline{1, m}$, $i = \overline{1, n}$, $\widehat{\Lambda}_0$, $\widehat{\Lambda}$.

Заметим, что алгоритм в общем случае не гарантирует оптимальности решений задач (10)–(14) и (1)–(3), (6)–(8). Однако на реальных данных показывает хорошие результаты в сравнении с другими средствами оптимальной кластеризации по точности и быстродействию, что позволяет надеяться на его широкое применение в практических целях.

Описанные далее результаты оптимальной кластеризации получены при помощи ПГА A_e.

Реализация последовательного гибридного алгоритма кластеризации с использованием локального поиска в парадигме роевого интеллекта

Общая схема реализованного алгоритма аналогична A_e ПГА, рассматривается как одна из его реализаций и реализуется по двухэтапной схеме следующим образом.

1. Генерация начального решения КПР (1)–(3), (6)–(8) и поиск локально наилучшего алгоритмом муравьиной колонии (АМК). По завершении каждой итерации t алгоритма муравьиной колонии производится попытка улучшить некоторый процент текущих лучших решений (параметр алгоритма) с помощью направленного локального поиска в парадигме роевого интеллекта (см. ниже АМК A_m).

2. Лучшее из полученных на первом этапе решений выбирается в качестве начального для решения задачи КЦК (15)–(17), (19) любым алгоритмом ЛП, включая алгоритм k-центров.

Алгоритм муравьиной колонии A_m

Используемые обозначения:
t — номер итерации алгоритма,
p_0 — пороговая вероятность,
N — количество объектов,
K — количество кластеров,
$(1 - \rho)$ — скорость испарения феромона,
L — количество лучших решений.

0. Ввод исходных данных и параметров алгоритма: T_0 — количество итераций, $(1-\rho)$ — скорость испарения феромона, p_0 — пороговая вероятность, L — количество лучших решений.

1. Формируется матрица феромонов **T** размером $N \times K$. Сначала все элементы инициализируются малым числом. На каждой итерации матрица обновляется.

2. Инициализация муравьев. На этом шаге генерируются случайные решения (муравьи) по следующей схеме. Генерируется случайное число в диапазоне от 0 до 1 из равномерного распределения. Если данное число меньше пороговой вероятности p_0 (параметр алгоритма), то объект распределяется в кластер с максимальным количеством феромонов для данного объекта по матрице феромонов. В противном случае принадлежность объекта кластеру определяется в соответствии с попаданием вновь сгенерированного случайного числа в интервал вероятностей, сформированных как нормализованная матрица феромонов **P** с элементами
$$p_{i,j} = \frac{\tau_{i,j}}{\sum\limits_{k=1}^{K} \tau_{i,j}}, \quad j = \overline{1,\ K}.$$

3. С целью улучшения полученных решений к определенному проценту муравьев ant_l применяется процедура локального поиска. Производится сортировка решений в порядке возрастания значений целевой функции КПР (6)–(7) и к некоторому количеству L (параметр алгоритма) текущих лучших решений применяется следующая процедура. Для каждого решения генерируется N случайных чисел. Если число этой последовательности с номером i меньше пороговой вероятности локального поиска p_0 (параметр алгоритма), то объекту с номером i в данном решении назначается кластер отличный от того, которому он был назначен в исходном решении. Выбор нового кластера производится случайным образом, при этом все кластеры, кроме исходного, имеют одинаковую вероятность назначения. Изменения сохраняются только в том случае, если целевая функция (6)–(7) на данном решении уменьшается.

4. Обновляется матрица феромонов. При этом учитываются как ранее найденные, так и новые решения, полученные на текущей итерации t. В обновлении матрицы участвуют только L лучших решений. Данная модификация муравьиного алгоритма носит название «метод элитных муравьев». Обновление производится по следующей формуле:
$$\tau_{i,j}(t+1) = (1-\rho)\tau_{i,j}(t) + \sum_{k=1}^{K} \Delta\tau_{i,j}, \quad i = \overline{1,\ N}, \quad j = \overline{1,\ K},$$
где $(1-\rho)$ — скорость испарения феромона (параметр алгоритма). Чем выше скорость испарения, тем быстрее забывается информация о решениях, полученных на предыдущих итерациях. Вторая часть суммы эквивалентна $\frac{1}{F_l}$, где F_l — значение целевой функции (6)–(7) на решении с номером l.

Шаги 2–4 циклически повторяются в соответствии с заданным количеством итераций T_0. Количество итераций T_0 — параметр алгоритма, задаваемый при вводе исходных данных.

Результаты вычислительных экспериментов с ПГА

Параметры АМК A_m, которые использовались при всех запусках программы, реализующей ПГА A_e:

$p_0 = 0{,}98$ — пороговая вероятность,

$(1 - \rho) = 0{,}01$ — скорость испарения феромона,

$p_l = 0{,}01$ — вероятность для локального поиска,

$ant_l = 1{,}0$ — процент муравьев для локального поиска.

Для всех описываемых ниже вычислительных экспериментов данные нормализованы в соответствии с выражениями

$$p_{i,t} = \alpha_t \frac{\widetilde{p}_{i,t} - \widetilde{p}_{\min,t}}{\widetilde{p}_{\max,t} - \widetilde{p}_{\min,t}}$$

либо

$$p_{i,t} = \alpha_t \frac{\widetilde{p}_{\max,t} - \widetilde{p}_{i,t}}{\widetilde{p}_{\max,t} - \widetilde{p}_{\min,t}}$$

(в зависимости от того, что по содержательному смыслу лучше — меньшие значения показателя или большие (подробности см. выше в п. 1)). Результаты сведены в табл. 1–8. Для IBM Cplex optimization studio во всех экспериментах время счета ограничено 2,5 часами (9000 с).

Для выяснения эффективности разработанного метода оптимальной кластеризации многомерных объектов по множеству разнородных показателей был выполнен ряд вычислительных экспериментов с использованием массивов данных, включающих социально-демографические, клинико-анамнестические, электроэнцефалографические и психометрические данные, отражающие состояние когнитивных функций, пациентов кардиологической клиники. Подробнее описание выборки и методик тестирования скорости сложной сенсомоторной реакции, показателей систем внимания и памяти было представлено ранее (например, в [Разумникова и др., 2022a; Разумникова и др., 2022b; Тарасова и др., 2017]). Сбор данных выполнялся в соответствии со стандартами Хельсинкской декларации, исследование когнитивного статуса пациентов получило одобрение Этического комитета ФГБНУ «Научно-исследовательский институт комплексных проблем сердечно-сосудистых заболеваний» (Кемерово, Россия).

В первом эксперименте были использованы данные, отражающие интегральные показатели когнитивного статуса (MMSE норм, CCI_1 норм) и активности коры головного мозга (мощности тета1-ритма, рассчитанного для фронтальной области правого полушария (FR_R_Z1_T1)). Эти показатели были зарегистрированы у 122 пациентов, для вычислений заданы три кластера (результаты вычислений приведены в табл. 1). В качестве целевой функции (ЦФ; критерия качества кластеризации) во всех экспериментах применен КПР-критерий (6)–(7).

Таблица 1. Результаты 3-кластеризации 122 объектов по двум показателям когнитивного статуса и мощности тета1-ритма

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК
Значение ЦФ	867,5	**862,43**	865,64
Точность	0,47	0,0	0,37
Время счета (с)	44	2150	9000

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 862,43.

Отличием второго эксперимента была замена параметра ЭЭГ на мощность альфа1-ритма фронтальной области левого полушария (FR_L_Z1_A1), количество кластеров также 3 (результаты вычислений приведены в табл. 2).

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 1128,59.

Таблица 2. Результаты 3-кластеризации 122 объектов по двум показателям когнитивного статуса и левополушарной мощности альфа1-ритма

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК
Значение ЦФ	1135,94	1128,63	**1128,59**
Точность	0,65	0,004	0,0
Время счета (с)	44	2150	9000

В третьем эксперименте наряду с MMSE норм и CCI_1 норм для выделения трех кластеров применяли показатели мощности альфа1-ритма для фронтальной области правого полушария (FR_R_Z1_A1) (результаты вычислений приведены в табл. 3).

Таблица 3. Результаты 3-кластеризации 122 объектов по двум показателям когнитивного статуса и правополушарной мощности альфа1-ритма

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК
Значение ЦФ	1144,92	**1115,65**	1120,14
Точность	2,62	0,0	0,4
Время счета (с)	44	2150	9000

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 1115,65.

Для четвертого эксперимента использовали показатели MMSE норм, CCI_1 норм и показатели мощности бета2-ритма для фронтальной области левого полушария (FR_L_Z1_B2) при количестве кластеров 3 (результаты вычислений приведены в табл. 4).

Таблица 4. Результаты 3-кластеризации 122 объектов по трем показателям с FR_L_Z1_B2

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК
Значение ЦФ	972,44	972,41	**968,26**
Точность	0,43	0,43	0,0
Время счета (с)	44	2150	9000

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 968,26.

Пятый эксперимент (табл. 5): количество пациентов — 122, количество показателей — 3 (в составе MMSE норм, CCI_1 норм, FR_R_Z1_B2), количество кластеров — 3.

Таблица 5. Результаты 3-кластеризации 122 объектов по двум показателям когнитивного статуса и правополушарной мощности бета2-ритма

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК
Значение ЦФ	985,03	**972,25**	977,24
Точность	1,31	0,0	0,51
Время счета (с)	44	2150	9000

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 972,25.

Общий вывод по пяти экспериментам: в среднем наилучшие решения удается получить, применяя предложенные алгоритмы ПГА A_e с надстройкой в виде A_m. Трудоемкость КПР не позволяет за сколько-нибудь приемлемое время получать оптимальные по аддитивному критерию разбиения точными методами (реализованными в IBM Cplex optimization studio) даже при кластеризации данных умеренной размерности. Вычислительные эксперименты доказывают эффективность локального поиска в парадигме роевого интеллекта, в частности алгоритма муравьиной колонии.

Данный вывод подтверждается и нижеследующими результатами (табл. 7, 8) двух вычислительных экспериментов с программными реализациями ПГА A_e задачи 3-разбиения с использованием клинических и психометрических параметров, приведенных в табл. 6.

Таблица 6. Показатели, участвующие в кластеризации

	Все индексы 1 — до операции, а 10 — на десятый день после операции	
1	стенозы	0 — нет, 1 — до 50 %, 2 — более 50 %
2	ЛТ	Личностная тревожность до КШ
3	СТ	Ситуативная тревожность до КШ
4	ВОЗРАСТ	
5	СЗМР_СЭ	Зрительно-моторная реакция выбора, скорость реакции
6	СЗМР_КО	Зрительно-моторная реакция выбора, количество ошибок
7	ПНП_СЭ	Уровень функциональной подвижности нервных процессов, скорость реакции
8	ПНП_КО	Уровень функциональной подвижности нервных процессов, количество ошибок
9	ПНП_ППС	Уровень функциональной подвижности нервных процессов, количество пропущенных положительных сигналов
10	РГМ_СЭ	Работоспособность головного мозга, скорость реакции
11	РГМ _КО	Работоспособность головного мозга, количество ошибок
12	РГМ_ППС	Работоспособность головного мозга, количество пропущенных положительных сигналов
13	КП_1мин	Корректурная проба Бурдона, количество обработанных символов за 1 мин теста (врабатываемость)
14	КП_4мин	Корректурная проба Бурдона, количество обработанных символов за 4 мин теста (истощаемость)
15	ЗП_Ч	Количество запомненных чисел
16	ЗП_Слоги	Количество запомненных слогов
17	ЗП_Слова	Количество запомненных слов

Содержательно задача состоит в наилучшем разбиении группы пациентов со стенозами на подгруппы по множеству показателей до операции КШ с целью прогноза тех когнитивных показателей, которые имеют максимальную вероятность снижения после оперативного вмешательства.

Результаты 3-кластеризации в эксперименте 6 с использованием 17 показателей когнитивного и личностного статуса пациентов (табл. 6), зарегистрированных до операции коронарного шунтирования, приведены в табл. 7.

Таблица 7. Результаты 3-кластеризации 199 объектов по 17 показателям когнитивного и личностного статуса пациентов, зарегистрированных до операции КШ

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК	КЦК (1 000 000 запусков)
Значение ЦФ	5814,55	5990,86	6154,8	**5694,91**
Точность	2,1	5,2	8,08	0,0
Время счета (с)	102	4994	9000	1220

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 5694,91.

3-кластеризация в эксперименте 7 (табл. 8) также проводилась по 17 показателям (стенозы, ЛТ_10, СТ_10, ВОЗРАСТ, СЗМР-СЭ_10, СЗМР-КО_10, ПНП-СЭ_10, ПНП-КО_10, ПНП-ППС_10, РГМ-СЭ_10, РГМ-КО_10, РГМ-ППС_10, КП-1м_10, КП-4м_10, ЗП ч_10, ЗП слоги_10, ЗП слова_10), приведенным в табл. 6, но после операции коронарного шунтирования.

Таблица 8. Результаты 3-кластеризации 199 объектов по 17 показателям когнитивного и личностного статуса пациентов, зарегистрированных после операции КШ

	КПР 1000 муравьев, 1000 итераций + КЦК	КПР 100 муравьев, 500 000 итераций + КЦК	Cplex КПР + КЦК	КЦК, случ. начальное решение (1 000 000 запусков)
Значение ЦФ	**5296,28**	5568,66	6081,42	5354,03
Точность	0,0	5,14	14,82	1,09
Время счета (с)	102	4994	9000	1220

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 5296,28.

Отмечаем значительное преимущество ПГА с применением АМК и даже равномерного случайного поиска перед применением Cplex для этих же целей.

Эксперимент 6 дает 8 % улучшения аддитивного КПР-критерия, эксперимент 7 — около 15 % при сокращении времени счета почти на два порядка.

Промежуточные результаты экспериментов позволяют сделать вывод о целесообразности применения представленной схемы ПГА A_e с АМК A_m. В подавляющем большинстве случаев такая комбинация лучше применения в ПГА IBM Cplex optimization studio и по точности, и особенно по времени решения. Поэтому завершающие эксперименты по кластеризации группы пациентов со стенозами на подгруппы по множеству уточненных показателей до операции КШ и после нее проводились только по схеме ПГА A_e с АМК A_m.

В табл. 9–12 представлены результаты 2- и 3-кластеризации пациентов до и после операции шунтирования. Количество пациентов — 163. Количество показателей — 17 (см. табл. 6).

Таблица 9. Результаты 2-кластеризации 163 объектов по 17 показателям до операции КШ

Количество муравьев	Количество итераций	Этап 1, КПР (АМК)	Этап 2, КЦК	Время счета
1000	1000	7114,49	**6614,48**	75
100	500 000	7135	6822,07	3603
0	1 000 000	*	**6614,48**	820

* КЦК (алгоритм *k*-средних) при случайном начальном решении (без КПР и использования муравьиного алгоритма) — 1 000 000 запусков.

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 6614,48.

Таблица 10. Результаты 3-кластеризации 163 объектов по 17 показателям до операции КШ

Количество муравьев	Количество итераций	Этап 1, КПР (АМК)	Этап 2, КЦК	Время счета
1000	1000	4744	4311,59	67
100	500 000	4747,89	4311,59	3734
0	1 000 000		**4286,78**	1084

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 4286,78.

Таблица 11. Результаты 2-кластеризации 163 объектов по 17 показателям после операции КШ

Количество муравьев	Количество итераций	Этап 1, КПР (АМК)	Этап 2, КЦК	Время счета
1000	1000	7014,28	**6493,56**	72
100	500 000	7068,42	6493,56	3667
0	1 000 000		6493,56	807

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 6493,56.

Таблица 12. Результаты 3-кластеризации 163 объектов по 17 показателям после операции КШ

Количество муравьев	Количество итераций	Этап 1, КПР (АМК)	Этап 2, КЦК	Время счета
1000	1000	4624,33	4624,33	71
100	500 000	4539,71	4539,71	3401
0	1 000 000		**4454,76**	1085

Найденное лучшее приближение к оптимальному решению имеет оценку по аддитивному КПР-критерию (6)–(7): 4454,76.

Приведем сравнение первых столбцов матриц координат $y^k = \left\|y_i^k\right\|$ 2-кластеризации до и после операции КШ (с оценками из табл. 9 и 11).

В этом случае 1 означает принадлежность первому кластеру, 0 — второму. Разбиению из табл. 9 соответствует вектор

$$[1001000001110100001100100101101110110101000010010010010111011100100010101000010111$$
$$0010001110110001000100011000000111010000011010010010101101100111101010000110010 1]^T.$$

Оценка разбиения по аддитивному КПР-критерию (6)–(7): 6614,48.

Разбиению из табл. 11 соответствует вектор:

$$[10\mathbf{1}1000001110100\mathbf{1}1100100101101110110101000010010010010111011110\mathbf{1}10001010101$$
$$0010111001000111011\mathbf{1}0010001000110000001110100\mathbf{1}0011010010010101101101\mathbf{1}1111\mathbf{1}110$$
$$100001100101]^T.$$

Оценка разбиения по аддитивному КПР-критерию (6)–(7): 6493,56.

Жирным шрифтом выделены позиции, перешедшие из второго кластера в первый (ГрНс).

Таким образом, состав кластеров, выделенных до и после операции, отличается только восемью пациентами, перешедшими из одного кластера в другой.

Анализ параметров, характеризующих эту малочисленную группу (ГрНС), показал наличие в ней постоперационного дефицита внимания и памяти (снижение показателей КП-4мин и ЗП_Слоги, $0{,}02 < p < 0{,}03$ согласно критерию Вилкоксона; табл. 13). Кроме того, ГрНС отличается от тех пациентов, что представляют «стабильные» кластеры старшим возрастом (61,9 лет по сравнению с 57,7 лет) и повышенной личностной тревожностью (43,4 и 39,4, соответственно), а также большим количеством ошибок при селекции зрительных сигналов и меньшей врабатываемостью при выполнении корректурной пробы ($p < 0{,}05$ согласно критерию Манна – Уитни при

Таблица 13. Психометрические характеристики пациентов, принадлежащих в «нестабильной» группе (ГрНС) по сравнению с пациентами, составившими «стабильные» кластеры (ГрС)

Показатель	До операции		После операции	
	ГрНс	ГрС	ГрНс	ГрС
ПНП_КО	29,3	24,2**		
СЗМР_КО			2,9	2,0*
КП_1мин	80,0	96,7*	51,9	78,4**
КП_4мин			61,3	99,9**
ЗП_Слоги			1,8	3,0**
ЗП_Слова			3,4	4,5*

Приведены медианы показателей: * — $p < 0{,}05$; ** — $p < 0{,}01$. Обозначение показателей — как в табл. 6.

межгрупповом сравнении дооперационных параметров; см. табл. 13). Результаты анализа постоперационных показателей когнитивных функций указывают на усиление когнитивного дефицита за счет ослабления вербальной памяти и истощаемости внимания (табл. 13).

Таким образом, ГрНС можно рассматривать как пациентов без выраженного стеноза сонной артерии, но с прогнозируемым ПОКД. Следовательно, разработанный метод оптимальной кластеризации с оценкой стабильности сформированных кластеров позволяет к известным факторам (наличие стеноза или старший возраст [Разумникова и др., 2022a; Разумникова и др., 2022b]) дополнительно выделить тех пациентов, когнитивные ресурсы которых оказываются недостаточны, чтобы преодолеть влияние операционной анестезии, вследствие чего отмечается однонаправленный эффект послеоперационного ухудшения разных показателей когнитивных функций: сложной зрительно-моторной реакции, внимания и памяти. Этот эффект свидетельствует о возможности более дифференцированно классифицировать пациентов с использованием описанного выше алгоритма. Согласно ранее выполненному нами исследованию закономерностей развития ПОКД с применением стандартного метода k-средних также были выделены «стабильные» и «нестабильные» кластеры, которые различались, однако, более сложными паттернами послеоперационных изменений когнитивных функций [Разумникова и др., 2022a].

Заключение

Таким образом, разработанный гибридный алгоритм оптимальной кластеризации на множестве демографических, клинико-анамнестических и психометрических показателей когнитивного статуса оказывается полезным для прогноза ПОКД у пациентов с ССЗ и возможности выбора на этой основе соответствующих индивидуальным особенностям методов активации когнитивных ресурсов.

Относительно разработанного аппарата дискретной оптимизации применительно к решению задач кластеризации больших данных на примерах вышеописанного анализа разнородных параметров (социально-демографических, клинико-анамнестических, электроэнцефалографических данных) и психометрических показателей когнитивных функций пациентов кардиологической клиники можно сделать следующие выводы.

1. Экспериментально доказана эффективность предложенных формальных постановок КПР, КЦК и гибридного алгоритма оптимальной кластеризации многомерных данных.

2. Экспериментально доказана эффективность применения алгоритмов локального поиска в парадигме роевого интеллекта в рамках гибридного алгоритма при решении задач оптимальной кластеризации. Из полученных результатов следует решение основной проблемы применения аппарата дискретной оптимизации — ограничения доступных размерностей

реализаций задач рассматриваемого класса. Фактически эта проблема снимается при сохранении приемлемой близости результатов к оптимальным.

3. Полученные результаты свидетельствуют о возможности расширения постановок и потенциала прикладного кластерного анализа многомерных данных. В частности, о возможности оптимальной кластеризации с неизвестным либо интервальным числом кластеров, либо наличием объектов, принадлежащих нескольким кластерам одновременно.

Список литературы (References)

Авдеенко Т. В., Мезенцев Ю. А. Кластеризация документов на основе семантической матрицы связей для концептуального индексирования // Вычислительные технологии. — 2020. — Т. 25, № 3. — С. 99–110. — DOI: 10.25743/ICT.2020.25.3.011

Avdeenko T. V., Mezencev Yu. A. Klasterizaciya dokumentov na osnove semanticheskoj matricy svyazej dlya konceptual'nogo indeksirovaniya [Clustering of documents based on the semantic matrix of links for conceptual indexing] // Computational technologies. — 2020. — Vol. 25, No. 3. — P. 99–110. — DOI: 10.25743/ICT.2020.25.3.011 (in Russian).

Айвазян С. А., Бухштабер В. М., Еников И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 608 с.

Ajvazyan S. A., Buhshtaber V. M., Enikov I. S., Meshalkin L. D. Prikladnaya statistika: klassifikaciya i snizhenie razmernosti [Applied statistics: classification and dimension reduction]. — Moscow: Finansy i statistika, 1989. — 608 p. (in Russian).

Кельманов А. В., Пяткин А. В. О сложности некоторых задач кластерного анализа векторных последовательностей // Дискретный анализ и исследование операций. — 2013. — Т. 20, № 2. — С. 47–57.

Kel'manov A. V., Pyatkin A. V. On complexity of some problems of cluster analysis of vector sequences // J. Appl. Ind. Math. — 2013. —Vol. 7, No. 3. — P. 363–369. (Original Russian paper: *Kel'manov A. V., Pyatkin A. V.* O slozhnosti nekotoryh zadach klasternogo analiza vektornyh posledovatel'nostej // Diskretnyj analiz i issledovanie operacij. — 2013. — Vol. 20, No. 2. — P. 47–57.)

Мезенцев Ю. А., Разумникова О. М., Тарасова И. В., Трубникова О. А. О некоторых задачах кластеризации больших данных по минимаксным и аддитивным критериям, применение в медицине и нейрофизиологии // Информационные технологии. — Изд-во «Новые технологии». — 2019. — Т. 25, № 10. — С. 602–608. — https://doi.org/10.17587/it.25.602-608

Mezencev Yu. A., Razumnikova O. M., Tarasova I. V., Trubnikova O. A. O nekotoryh zadachah klasterizacii bol'shih dannyh po minimaksnym i additivnym kriteriyam, primenenie v medicine i nejrofiziologii [On some tasks of clustering big data by minimax and additive criteria, application in medicine and neurophysiology] // Information technology. — Izd-vo "Novye texnologii". — 2019. — Vol. 25, No. 10. — P. 602–608 (in Russian).

Разумникова О. М., Мезенцев Ю. А., Павлов П. С., Тарасова И. В., Трубникова О. А. Применение инструментов дискретной оптимизации для классификации когнитивного дефицита: особенности использования минимаксного и аддитивного критериев // Программные продукты и системы. Тверь: Изд-во МНИИПУ и НИИ «Центрпрограммсистем», 2021. — № 4. — С. 579–588. — DOI:10.15827/0236-235X.136.579-588

Razumnikova O. M., Mezencev Yu. A., Pavlov P. S., Tarasova I. V., Trubnikova O. A. Primenenie instrumentov diskretnoj optimizacii dlya klassifikacii kognitivnogo deficita: osobennosti ispol'zovaniya minimaksnogo i additivnogo kriteriev [Application of discrete optimization tools for classification of cognitive deficits: features of using minimax and additive criteria] // Software products and systems. Tver': Izd-vo MNIIPU i NII "Centrprogrammsistem", 2021. — No. 4. — P. 579–588. — DOI: 10.15827/0236-235X.136.579-588 (in Russian).

Разумникова О. М., Тарасова И. В., Трубникова О. А., Барбараш О. Л. Изменения в структуре когнитивных функций и тревожности у кардиохирургических пациентов в зависимости от выраженности стенозов сонных артерий // Комплексные проблемы сердечно-сосудистых заболеваний. — 2022а. — Т. 11, № 1. — С. 36–48.

Razumnikova O. M., Tarasova I. V., Trubnikova O. A., Barbarash O. L. Izmeneniya v strukture kognitivnyh funkcij i trevozhnosti u kardiohirurgicheskih pacientov v zavisimosti ot vyrazhennosti stenozov sonnyh arterij [Changes in the structure of cognitive functions and anxiety in cardiac surgery patients depending on the severity of carotid artery stenosis] // Kompleksnye problemy serdechno-sosudistyh zabolevanij (Complex problems of cardiovascular diseases). — 2022a. — Vol. 11, No. 1. — P. 36–48 (in Russian).

Разумникова О. М., Тарасова И. В., Трубникова О. А., Барбараш О. Л. Кластеризация показателей когнитивного статуса кардиохирургических пациентов для оценки риска его послеоперационных изменений // Acta Biomedica Scientifica. — 2022b. — Т. 7, № 1. — С. 129–138. — doi:10.29413/ABS.2022-7.1.15

Razumnikova O. M., Tarasova I. V., Trubnikova O. A., Barbarash O. L. Klasterizaciya pokazatelej kognitivnogo statusa kardiohirurgicheskih pacientov dlya ocenki riska ego posleoperacionnyh izmenenij [Clustering of indicators of the cognitive status of cardiac surgery patients to assess the risk of its postoperative changes] // Acta Biomedica Scientifica. — 2022b. — Vol. 7, No. 1. — P. 129–138. — DOI: 10.29413/ABS.2022-7.1.15 (in Russian).

Тарасова И. В., Трубникова О. А., Барбараш О. Л., Барбараш Л. С. Изменения электроэнцефалограммы у пациентов с ранней и стойкой послеоперационной когнитивной дисфункцией при коронарном шунтировании с искусственным кровообращением // Неврологический журнал. — 2017. — № 3. — С. 136–141.

Tarasova I. V., Trubnikova O. A., Barbarash O. L., Barbarash L. S. Izmeneniya elektroencefalogrammy u pacientov s rannej i stojkoj posleoperacionnoj kognitivnoj disfunkciej prikoronarnom shuntirovanii s iskusstvennym krovoobrashheniem [Changes in the electroencephalogram in patients with early and persistent postoperative cognitive dysfunction during coronary bypass surgery with artificial blood circulation] // Nevrologicheskij zhurnal (Neurological Journal). — 2017. — No. 3. — P. 136–141 (in Russian).

Eremeev A. V., Kel'manov A. V., Kovalyov M. Y., Pyatkin A. V. Maximum diversity problem with squared Euclidean distance // Lect. Notes Comput. Sci. — 2019. — Vol. 11548. — P. 541–551.

Lenstra J. K., Shmoys D. B., Tardos E. Approximation algorithms for scheduling unrelated parallel machines. — Centre for Mathematics and Computer Science, Amsterdam, 1987. — Report OS-R8714.

Mezentsev Yu. Binary cut-and-branch method for solving mixed integer programming problems // Constructive Nonsmooth Analysis and Related Topics (Dedicated to the memory of V. F. Demyanov), CNSA, 2017. — https://doi.org/10.1109/cnsa.2017.7973989

Pinedo M. Scheduling theory, algorithms, and systems. — 3rd. ed.. — Springer, 2008. — 672 p.

Pyatkin A. V. PTAS for p-means q-medoids r-given clustering problem // Lect. Notes Comput. Sci. — 2023. — Vol. 13930. — P. 133–141.