УДК 519.237.8+61:007+61:311+004.8

А.С. Акопов[1], А.А. Московцев[1,3], С.А. Доленко[2], Г.Д. Савина[3]

Кластерный анализ в медико-биологических исследованиях

[1] Федеральное государственное бюджетное учреждение «Научно-исследовательский институт общей патологии и патофизиологии» Российской академии медицинских наук, 125315, Москва, ул. Балтийская, 8
[2] Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына Московского государственного университета им. М.В. Ломоносова, 119991, ГСП-1, Москва, Ленинские горы, 1, стр. 2
[3] Государственное бюджетное образовательное учреждение дополнительного профессионального образования «Российская медицинская академия последипломного образования» Министерства здравоохранения Российской Федерации, 123995, Москва, ул. Баррикадная, 2/1

Кластерный анализ является одним из наиболее популярных методов анализа многопараметрических данных. Применение кластерного анализа позволяет выявить внутреннюю структуру данных, сгруппировать отдельные наблюдения по степени их схожести. В обзоре дается определение основных понятий кластерного анализа, а также рассматриваются наиболее популярные алгоритмы кластеризации: k-средних, иерархические алгоритмы, алгоритмы на базе сетей Кохонена. Также приводятся примеры использования данных алгоритмов в медико-биологических исследованиях.

Ключевые слова: *кластерный анализ, многопараметрическая статистика, кластер, k-средних, иерархические алгоритмы, искусственные нейронные сети, сеть Кохонена*

A.S. Akopov[1], A.A. Moskovtsev[1,3], S.A. Dolenko[2], G.D. Savina[3]

Cluster analysis in biomedical researches

[1] Institute of General Pathology and Pathophysiology, RAMS, 8, Baltiyskaya str., 125315, Moscow, Russia
[2] Skobeltsyn Institute of Nuclear Physics, Moscow State University, 1, bld.2, Leninskye Gory, Moscow, GSP-1, 119991, Russia
[3] Russian Medical Academy of Postdegree Education, 2/1, Barrikadnaya str., 123836, Moscow, Russia

Cluster analysis is one of the most popular methods for the analysis of multi-parameter data. The cluster analysis reveals the internal structure of the data, group the separate observations on the degree of their similarity. The review provides a definition of the basic concepts of cluster analysis, and discusses the most popular clustering algorithms: k-means, hierarchical algorithms, Kohonen networks algorithms. Examples are the use of these algorithms in biomedical research.

Key words: *cluster analysis, multi-parameter statistics, cluster, k-means, hierarchical algorithms, artificial neural networks, Kohonen network*

Задача группировки объектов, однородных по тем или иным параметрам, является актуальной для многих областей науки и техники и имеет длительную историю развития инструментов для ее решения. Попытки введения систематики живых организмов относятся к античности [14]. Примерами решения подобных задач в настоящее время в медико-биологических исследованиях служат: выделение групп пациентов на основе набора клинических показателей [8, 57]; разделение экспериментальной выборки на группы единиц со схожими профилями регистрируемых в опыте показателей [11]; построение таксономии микроорганизмов на основе диагностических и биохимических тестов [39, 50]; классификация образцов опухолевых тканей по профилям экспрессии генов [17, 33]; сегментация (т.е. выделение однородных участков изображения) в нейровизуализационных технологиях [21].

Несмотря на разнообразие группируемых объектов (ткани, организмы, биологические виды, пиксели изображения), в этих исследованиях используются сходные методики, объединенные под названием «*кластерный анализ*». Термин был введен в 1936 г. [76], однако прообразы кластерного анализа можно встретить в более ранних работах (метод корреляционных плеяд Чекановского [22]). Бурное развитие методик и алгоритмов кластерного анализа началось со 2-й половины XX века, в связи с развитием вычислительной техники, и формированием направления *datamining* в информационных технологиях. В медико-биологических исследованиях кластерный анализ наиболее широко используется в связи с высокопроизводительными методами исследований, такими, как биологические микрочипы (microarrays) и секвенирование следующего поколения (next generation sequencing).

Для корреспонденции: *Акопов Александр Сергеевич*, инженер лаб. тромболизиса и тромбогенеза ФГБУ «НИИОПП» РАМН. E-mail: asakopov@gmail.com

Datamining (рус. — добыча данных, интеллектуальный или глубинный анализ данных) — совокупность методик, направленных на обнаружение скрытой информации, знаний и закономерностей в крупных массивах данных, как правило, — сведенных в компьютерные базы данных. Помимо кластерного анализа к datamining относят дискриминантный, корреляционный и регрессионный анализ, факторный анализ, анализ временных рядов, анализ выживаемости, анализ связей и многие другие методики анализа данных.

Основные определения

Кластерный анализ (кластеризация) — способ группировки многомерных объектов, основанный на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп как «сгустков» этих точек («Кластер» — англ. — сгусток, гроздь, скопление) [9]. Существует ряд синонимичных терминов: автоматическая классификация, нумерическая, статистическая или цифровая таксономия. С точки зрения теории машинного обучения (ТМО), кластеризация относится к задачам обучения без учителя, т.е. адаптивный алгоритм при обучении использует только входные данные. Это отличает ее от классификации, при проведении которой алгоритм сначала обучается на выборке данных с заданным разделением на классы, корректируя свою работу для соответствия этому разделению [4].

Определения кластера могут различаться в зависимости от задачи, решаемой при проведении анализа [19]. Одним из наиболее распространенных является следующее: *кластер* — подмножество набора данных, в котором каждый элемент ближе или более схож (в каком-либо смысле) с каждым из элементов, принадлежащих данному подмножеству, чем с любым из элементов, ему не принадлежащих [72].

Также разнообразны, в зависимости от прикладной области, термины, обозначающие единицу данных, подвергаемых кластеризации: событие, единица, объект, наблюдение, пример, паттерн, образ, операционная таксономическая единиц. В общем случае, *единица данных* представляет собой вектор (т.е. упорядоченный набор значений) признаков. *Признаками (атрибутами)* являются характеристики изучаемых объектов, например: уровень экспрессии данного гена в образце, значение какого-либо клинического показателя пациента, интенсивность по одному из цветовых каналов пикселя изображения. Вообще неоднородность терминологии характерна для кластерного анализа, так как метод формировался в рамках различных прикладных дисциплин [15].

Признаки, описывающие объект, могут выражаться в различных шкалах, что важно с точки зрения для применимости того или иного алгоритма кластерного анализа [66]. Наиболее общепринятая классификация [71] выделяет 4 вида шкал, в зависимости от вида представляемых данных, возможных преобразований и вводимых отношений между объектами (табл. 1).

Расстояния, метрики, меры сходства

Схожесть объектов является критерием их помещения в один кластер. Более формальными критериями в алгоритмах кластеризации выступают расстояния, метрики, меры сходства и различия, являющиеся функциями двух аргументов (обозначение — x, y), в качестве которых выступают объекты. Кроме них в тех же целях могут использоваться: коэффициенты корреляции или ассоциации, вероятностные меры сходства [68]. *Функция расстояния* $d(x,y)$ определяется следующими очевидными свойствами [23]:

1. Неотрицательность $d(x,y) \geq 0$ (отрицательное расстояние не имеет смысла);

2. Симметричность $d(x,y) = d(y,x)$ (расстояние от первой точки до второй равно расстоянию, измеренному в обратном направлении);

3. Свойство идентификации $d(x,x) = 0$ (расстояние от точки до нее самой равно нулю).

Таблица 1

Сравнение шкал представления данных

Название	Возможные отношения, устанавливаемые для значений признака	Примеры
Абсолютная	Больше/меньше/равно, на сколько и во сколько раз	Рост, вес индивида; концентрация молекул в растворе
Интервальная (Разностей)	Больше/меньше/равно, на сколько	Температура
Ординальная	Больше/меньше/равно	Классификация эссенциальной гипертензии по ВОЗ (I, II, III стадии)
Номинальная	Равно	Пол, цвет

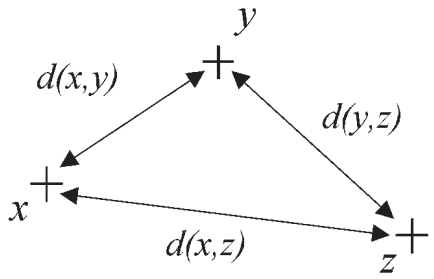

Рис. 1. Иллюстрация аксиомы треугольника: сумма расстояний *d(x,y)* и *d(y,z)* больше *d(x,z)*. Равенство может выполняться при условии расположения точки *y* на прямой, соединяющей *x* и *z*.

Определение *метрики* [6] включает в себя вышеперечисленные свойства (неотрицательность, симметричность, свойство идентификации) и дополняется следующими:

4. Свойство определенности $d(x,y) = 0$, если и только если $x = y$ (расстояние между точками нулевое лишь при совпадении этих точек);

5. Неравенство треугольника:
$d(x,z) \leq d(x,y) + d(y,z)$ (сумма расстояний $d(x,y)$ и $d(y,z)$ больше или равна $d(x,z)$) — см. рис. 1.

Для упрощения дальнейшего изложения опустим различия между понятиями *расстояние* и *метрика*.

Обобщением понятий метрики и расстояния является *мера сходства (различия)* S, которая определяется менее строгим набором свойств [32]:

1. Неотрицательность $S(x,y) \geq 0$;

2. Симметричность $S(x,y) = S(y,x)$;

3. Монотонное возрастание $S(x,y)$ по мере того, как выбираются более схожие (различные), в каком либо смысле, объекты x и y.

Выбор меры сходства между объектами для кластеризации может определяться шкалой, в которой представлены признаки объектов, используемым алгоритмом, характеристиками самих данных (видом распределения признаков). Часто при работе алгоритмов необходимо определять не только расстояния между объектами, но и расстояния между кластерами.

Расстояния между числовыми данными. Одно из наиболее известных расстояний между векторами из n числовых признаков — Евклидово:

$$d^2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

где x_i и y_i — значения i-х признаков объектов x и y.

Данная метрика относится к семейству метрик Минковского, представимых следующей общей формулой:

$$dp(x, y) = \sqrt[p]{\sum_{i=1}^{n} (x_i - y_i)^p},$$

где ρ может принимать значения от 1 до бесконечности (для Евклидовой метрики $\rho = 2$).

К этому семейству также относится Манхэттенская метрика (или «расстояние городских кварталов»), случай для $\rho = 1$:

$$d^1(x, y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|}.$$

Недостатком Евклидовой метрики является тенденция к формированию кластеров одинакового размера и сферических по форме [37], что может приводить к искажению реального распределения данных по кластерам (рис. 2). В то же время, признаки, измеренные по шкале с более широким размахом, будут сильнее влиять на значение расстояния. Манхэттенское расстояние менее чувствительно к таким особенностям данных.

Расстояния между строковыми объектами. В ряде приложений используется расстояние Хэмминга — метрика, используемая для сравнения векторов равной длины, которые образованны признаками с конечным набором значений (например, строки, записанные в определенном алфавите). Например, она может использоваться для сравнения последовательностей нуклеиновых кислот или белков одного размера [51]. Значение метрики Хэмминга определяется как количество несовпадающих значений в соответствующих позициях последовательностей [59]. Например, для двух пентануклеотидных последовательностей: $d_{Ham}(agtcc,agaat) = 3$.

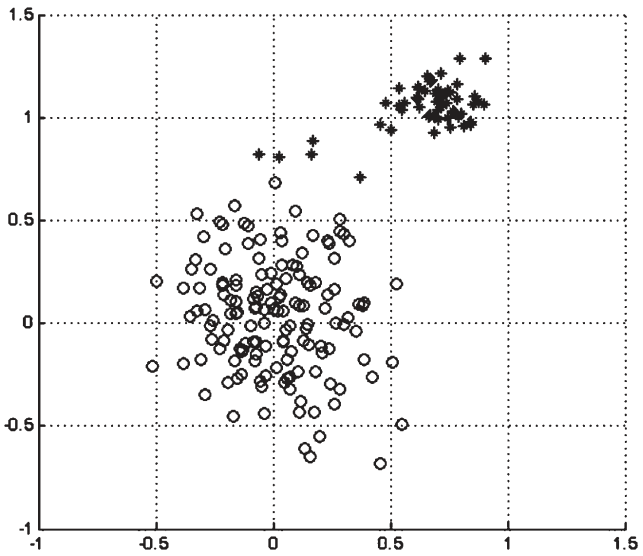

Рис. 2. Разделение данных на кластеры алгоритмом *k*-средних с использованием Евклидовой метрики. Характеристики групп данных: кластер 1 — кружок; кластер 2 — звездочка. Пять точек, принадлежащих кластеру 1 (более разреженному и многочисленному), ошибочно отнесены алгоритмом к кластеру 2.

Таблица 2

Категории алгоритмов кластерного анализа

По структуре получаемых кластеров		
Разделительные (partitional) — объекты распределены между кластерами одного уровня.	Иерархические (hierarchical) — объекты распределены между кластерами, образующими иерархическую структуру. Имеется серия все более мелких разбиений.	
По принадлежности объектов кластеру		
Эксклюзивные (exclusive) или "жесткие" (hard): каждый из объектов относится к одному кластеру.	Перекрывающиеся (overlapping): один объект может быть отнесен сразу к нескольким кластерам.	Нечеткие (fuzzy) и вероятностные (probabilistic): один пример — набор чисел, выражающих степень принадлежности к кластерам. Если сумма чисел равна 1 кластеризация описывает вероятность принадлежности к кластеру.
По способу определения числа кластеров		
Число кластеров определяется пользователем.	Число кластеров определяется в ходе выполнении алгоритма.	
Использование адаптивных алгоритмов		
Не используются.	С использованием искусственных нейронных сетей (ИНС).	С использованием генетических алгоритмов (ГА).

Также использование расстояния Левенштейна, которое, в отличие от предыдущего, не накладывает ограничений на длину сравниваемых последовательностей. Расстояние Левенштейна определяется как минимальное количество удалений и вставок одного символа, а также замены его на другой до получения одной строки из другой [7]. Здесь недостатком является то, что данное расстояние между совершенно различными короткими строками оказывается небольшим, а между достаточно длинными и при том похожими может быть гораздо больше.

Матрицы расстояний (сходства). Естественным представлением кластеризуемых данных является матрица, у которой строки (столбцы) соответствуют кластеризуемым единицам, а столбцы (строки, соответственно) признакам этих единиц. Однако часто при работе алгоритмов кластеризации строится матрица, строки и столбцы которой соответствуют кластеризуемым объектам, а на их пересечении записаны расстояния или значения расстояний (меры сходства) между соответствующими объектами [10].

Категоризация алгоритмов

Существующие классификации кластерного анализа не являются абсолютно исчерпывающими и однозначными, во многом потому, что разработка алгоритмов осуществлялась специалистами разных областей, для специфических задач и типов данных. В основе большинства классификаций лежат разные критерии распределения объектов по кластерам, и даже разное понимание того, что такое кластер [29]. Отечественными авторами предложена достаточно подробная классификация методов кластерного анализа [2].

Многие алгоритмы могут обладать характеристиками, позволяющими отнести их сразу к нескольким классам. Помимо собственно классификации, т.е. разделения всех алгоритмов на однородные группы, возможны и другие способы описания структуры методов кластерного анализа. В табл. 2 приведены характеристики, по которым можно разделить алгоритмы на формализованные категории.

Другие варианты классификации методов кластерного анализа можно найти в ряде работ [5, 36, 40].

Далее представлено описание и приведены примеры использования наиболее распространенных алгоритмов, применяемых в кластерном анализе.

Алгоритм *k*-средних

Подход предложен Г. Штейнгаузом в 1956 г. [70] и С. Ллойдом в 1957 г. [49] и на сегодняшний день является одним из наиболее часто используемых. Алгоритм является разделительным, с количеством кластеров, задаваемым пользователем, и жестким распределением объектов по ним. Основной особенностью алгоритма является распределение точек-объектов анализируемых данных по k (заранее заданное число) кластерам, на основании их близости к центроидам. В качестве центроида данного кластера берется объект, имеющий значения признаков, усредненных по всем объектам, принадлежащим данному кластеру (рис. 3).

Описание алгоритма

В процессе выполнения алгоритма циклически (итеративно) осуществляются следующие действия:

1) каждый объект приписывается ближайшему (обычно, в смысле Евклидовой метрики) центроиду, в результате чего образуются кластеры;

2) для каждого кластера заново рассчитывается центроид;

3) процедура повторяется.

Таким образом, в работе алгоритма циклически осуществляется изменение координат центроидов и перераспределение объектов между ними. Остановка алгоритма происходит в случае отсутствия изменения координат центроидов, при изменении их ниже заданного порогового значения, либо после заданного числа повторений процедуры. При работе алгоритма решается задача минимизации суммы квадратов расстояний (СКР) между точками и центроидами. Перед началом циклической процедуры происходит инициализация алгоритма множеством анализируемых объектов, а также начальными координатами для k центроидов.

Преимущества алгоритма:

- хорошая производительность при достаточно большом объеме данных;
- в случае «хорошей» разделенности и геометрии кластеров в структуре исходных данных, алгоритм сходится за небольшое число итераций.

Недостатки:

- исследователь должен сделать как можно более правдоподобное предположение о числе кластеров в структуре данных до начала работы алгоритма;
- медленная работа на больших объемах данных;
- хорошо разделяются только сферические кластеры, т.е. кластеры с равномерным разбросом значений по признакам (следствие минимизации суммы квадратов расстояний);
- зачастую алгоритм достигает только локальный минимум СКР, т.е. разбиение не отражает реальную структуру данных (что соответствует достижению глобального минимума).

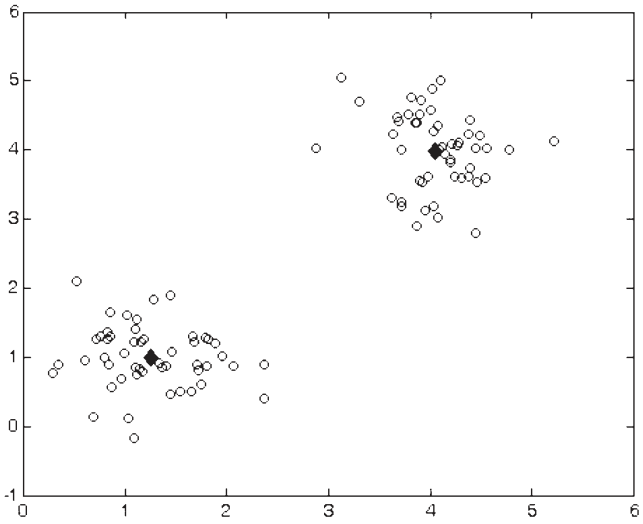

Рис. 3. Центроиды (темные ромбы) двух кластеров, образованных данными, имеющими двумерное нормальное распределение (светлые кружки). Координаты центроида определяются как среднее соответствующих координат объектов кластера.

С последним тесно связана проблема инициализации центроидов [54]. Их «попадание» в один и тот же кластер в начале работы алгоритма ведет к его разбиению, в то время как истинный кластер, в который «не попал» ни один центроид, не выделяется. Стабильность разбиения на кластеры может подтверждаться путем многократного повторения процедуры. Кроме того, для минимизации влияния инициализации можно воспользоваться рядом усовершенствований метода k-средних. Например, используется следующий алгоритм для распределения центроидов [42]:

1. Первым из k центроидов назначается наиболее центральный из кластеризуемых объектов;
2. Следующими центроидами назначаются объекты, вокруг которых группируется максимальное число других объектов.

Также используют подход с равномерным распределением центроидов [37]:

1. Первым из k центроидов назначается выбранный случайно объект или точка, являющаяся центром выборки;
2. Следующими центроидами назначаются объекты, максимально удаленные от выбранных на предыдущем шаге.

В литературе предлагается и рассматривается ряд других усовершенствований процедуры инициализации [28, 31, 34, 75].

Применение алгоритма *k*-средних

Кластеризация временных профилей экспрессии генов Saccharomycescerevisae в клеточном цикле [73]

В работе анализировались данные экспрессии для 3000 генов [20], полученные методом ДНК-микрочипов [64]. После предварительной обработки экспериментальных данных, каждый ген описывался вектором значений уровней экспрессии во времени (15 значений). Для нахождения глобального максимума функции СКР использовалось 200 повторов работы алгоритма, что обеспечивало стабильно воспроизводимый результат кластеризации. Инициализация алгоритма осуществлялась с помощью равномерного распределения центроидов. Валидацию результата проводили путем выявления функционально близких генов в одном кластере, причем статистически значимое обогащение функциональными категориями наблюдалось для 15 кластеров из 30 и совпадало с наиболее «плотными» кластерами. Кроме того, для генов 12 кластеров были обнаружены общие регуляторные мотивы (всего 18 мотивов), причем наличие семи из них было подтверждалось экспериментально до исследования. Таким образом, кластерный анализ позволил выявить наличие функциональных связей между генами, которые подтверждались как существующими данными об их функциях, так и общими механизмами регуляции.

Изучение зависимости экспрессии генов в печени крыс от пола и возраста [47]

В исследовании изучались данные экспрессии генов крыс 9 возрастных групп обоих полов, полученные методом ДНК-микрочипов. После предобработки, матрица данных была образована 7951 объектом (точки на микрочипах, соответствующие 3770 дифференциально экспрессируемым генам), каждый из которых характеризовался уровнями экспрессии в 9 временных точках (возраст от 2 до 104 недель). Значение k устанавливалось равным 30, что соответствовало наименьшему числу кластеров, при котором коэффициент корреляции любых двух точек внутри каждого из них составлял не менее 0,7. Многие кластеры представляли собой профили экспрессии генов, которые могли быть интерпретированы в физиологическом контексте: зависимость уровней экспрессии от стадии развития организма. Некоторые кластеры обладали негативной корреляцией друг с другом (гены-«антагонисты»).

Выявление типов сигнальных путей в мононуклеарных периферических клетках крови, активируемых в период обострения бронхиальной астмы [18]

Активация сигнальных путей определялась по изменению уровней экспрессии генов в мононуклеарных периферических клетках крови (МПКК) из образцов, забранных у пациентов в период обострения по сравнению со спокойным периодом. Изучались 166 образцов МПКК, полученных от лиц, страдающих бронхиальной астмой II, III и IV ступеней. В отличие от приведенных выше исследований, кластеризация проводилась по образцам, а не по временным профилям генов. Таким образом, больные разделялись на кластеры по профилям экспрессии генов МПКК в стадии обострения заболевания. Значение k было выбрано равным 3 в результате оценки качества разделения на кластеры при различных k с помощью silhouette-статистики (см. приложение). С той же целью использовались повторные кластеризации с искусственным добавлением шума к данным [53]. В результате проведенного анализа 3 полученных кластера были охарактеризованы по особенностям профилей экспрессии генов в МПКК: 1-й — с активацией генов сигнальных путей врожденного иммунитета (сигнальные пути TLR и рецепторов интерферона); 2-й — с активацией генов антиген-зависимых путей (сигнальные пути Т- и В-клеточных рецепторов, а также рецепторов ИЛ-4); 3-й — без ярко выраженных особенностей. Таким образом, авторами была показана гетерогенность молекулярных механизмов протекания заболевания в стадии обострения.

Кластеризация генов для последующей реконструкции динамической сетевой модели их взаимодействия в условиях протекания инфекции вирусом гриппа H1N1 [24]

Исследовались уровни экспрессии генов в легочной ткани мышей в различных временных точках после инфицирования вирусом гриппа H1N1 методом ДНК-микрочипов. В данном исследовании кластерный анализ проводился на этапе подготовки данных для последующего построения с их помощью компьютерной модели сети взаимодействия генов. При этом моделирование было бы не возможно без снижения размерности данных ввиду их большого объема. Исследователями было принято решение об объединении групп генов в кластеры с последующим их представлением в модели центроидами или отдельными представителями соответствующих кластеров. Данные после предобработки анализировались с помощью алгоритма k-средних, с использованием расстояния Евклида (учитывает, прежде всего, сонаправленность изменений экспрессии). Устойчивость разбиения на кластеры достигалась повторением процедуры более 100 раз, начальные центроиды выбирались случайным образом из кластеризуемых объектов при каждом повторе. Оптимальное число кластеров определяли осуществляя пробные кластеризации со значением k от 10 до 80 и оценивая качество получаемых кластеров по индексу Дана (см. приложение), а также их обогащенности функционально близкими генами.

Иерархические алгоритмы кластерного анализа

Иерархические алгоритмы (или алгоритмы таксономии) образуют семейство крайне популярных методик кластерного анализа: по данным исследования библиографических баз данных иерархические методики оказались наиболее используемы в публикациях за 2003 г. [43]. Результатом проведения иерархической кластеризации является дендрограмма — древовидная структура, объединяющая кластеризуемые объекты (рис. 4 А,Б), и образующая систему вложенных разбиений [3].

Алгоритмы иерархического кластерного анализа дают наглядное представление о структуре кластеров и отношении близости объектов в них. К прочим преимуществам иерархических методов относятся [16]:

- гибкость в выборе степени разбиения данных (возможен выбор любого числа кластеров, на которые разбиваются данные);
- возможность использования различных мер сходства, расстояний, применимых к различным типам данных.

Недостатки, свойственные иерархическому подходу:

- окончательный выбор числа кластеров не очевиден, и не может базироваться лишь на результате работы алгоритма;

• отсутствует возможность перераспределения объектов, уже объединенных в кластеры.

Иерархические методы делятся на агломеративные и дивизимные [28]. Двигаясь с нижнего уровня вверх, агломеративные процедуры объединяют объекты во все более крупные кластеры, заканчивая формированием одного корневого кластера, объединяющего все объекты. В противоположность, дивизимные алгоритмы осуществляют разделение единого кластера на все более мелкие, вплоть до помещения каждого объекта в индивидуальный кластер. Более распространенными являются агломеративные процедуры.

Агломеративные алгоритмы

Простейшая процедура агломеративного кластерного анализа:

1) каждый объект помещается в отдельный кластер;

2) строится матрица расстояний или мер близости между кластерами;

3) в матрице находятся два ближайших кластера (при первой итерации — объекта), которые объединяются в один;

4) алгоритм повторяется, начиная с п.2, до тех пор, пока все объекты не окажутся в одном кластере.

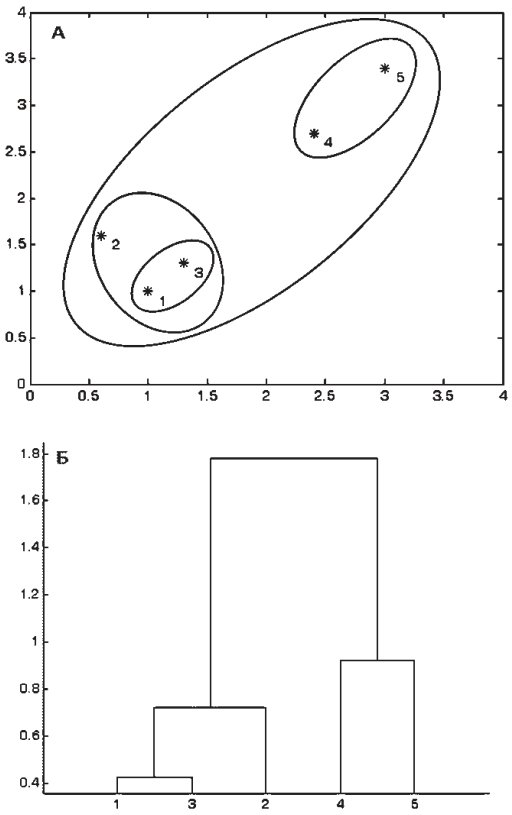

Рис. 4. Построение дендрограммы при иерархической кластеризации: А — данные в двумерном пространстве, состоящие из 5 объектов (звездочки), овалами отражено распределение данных по кластерам разных уровней иерархии; Б — те же кластеры приведены в виде дендрограммы

Как правило, в алгоритме иерархической кластеризации в п.2 и 3 используется один из трех способов определения расстояния между кластерами [30, 40]:

• одиночной связи (single-linkage) [67] — за расстояние между кластерами принимается расстояние между двумя ближайшими объектами этих кластеров;

• полной связи (complete-linkage) [52] — за расстояние между кластерами принимается расстояние между двумя наиболее удаленными объектами этих кластеров;

• средней связи (average-linkage) [69] — расстояние между кластерами определяется как среднее расстояний от объектов одного кластера до объектов другого.

Кроме того, выбор кластеров для объединения может осуществляться с помощью критерия Уорда (Ward'smethod), использующего целевую функцию, наиболее часто — СКР от точек кластеров до их центроидов. На каждом этапе работы алгоритма выбираются такие два кластера, что при их объединении прирост СКР минимален [63].

Применение алгоритмов иерархической кластеризации

Кластеризация и визуальное представление полногеномных профилей экспрессии генов [27]

В исследовании использовались данные исследования экспрессии генов Saccharomycescerevisae в различных условиях: при диауксии, в цикле митотического деления, при споруляции, а также при температурном и восстановительном стрессе. Кластерный анализ осуществлялся с помощью агломеративного алгоритма средней связи. В качестве метрики использовалась модификация коэффициента корреляции. Применение иерархического алгоритма позволило сформировать графическую структуру — дендрограмму, отражающую степень близости генов в смысле их коэкспрессии в различных условиях. Для обеспечения лучшего визуального восприятия, профили экспрессии генов были представлены в виде «листьев» дендрограммы, которые образованы цветными ячейками, соответствующими уровню экспрессии гена при данном условии. Гены со сходными функциями оказывались близки на дендрограмме. В частности, это касалось группы генов, кодирующих рибосомальные белки, чья экспрессия подавлялась при стрессе и взаимно коррелировала в течение митотического цикла.

Анализ распределения уровней экспрессии типов рецепторов, связанных с G-белками, в различных тканях мышей [60]

Исследователями анализировались уровни экспрессии 353 рецепторов, связанных с G-белками в 41-й ткани взрослых мышей с помощью метода количественной ПЦР в реальном времени. Данные кластеризовались с помощью алгоритма, использованного

в описанной выше работе [27]. При этом дендрограммы строились как в отношении тканей, так и рецепторов. Выявлены кластеры, образованными тканями иммунной и кроветворной систем (селезенка, костный мозг, тимус), а также центральной нервной системы (кора мозга, гипоталамус, мозжечок, ствол мозга). Многие другие ткани относящиеся к одним системам органов также демонстрировали на дендрограмме близость набора экспрессируемых рецепторов. На базе полученных результатов авторами были выдвинуты гипотезы о ранее неизвестных функциях некоторых рецепторов.

Выявление фенотипов с высоким риском смертности среди пациентов с хронической обструктивной болезнью легких [19]

Исходными данными для анализа являлись значения клинических показателей 527 пациентов, страдающих ХОБЛ, полученных в рамках проспективного исследования. Среди показателей были как количественные (ИМТ, ФЖЕЛ и др.), так и качественные (номинальные и ординальные данные о сопутствующих заболеваниях, наличии бронхоэктазов, степени утолщения стенок бронхиол и др.). Предобработка данных для кластерного анализа заключалась в преобразовании количественных данных с помощью анализа главных компонент (выделение двух переменных вместо 7, часть из которых демонстрировали взаимную корреляцию), а также построении независимых количественных переменных на основе категориальных данных с помощью метода множественного анализа соответствий (14 переменных). Кластерный анализ осуществляли с использованием критерия Уорда. На основании визуального изучения полученной дендрограммы авторы сделали предположение о наибольшей вероятности разбиений на 3 и 5 кластеров. Анализ смертности для кластеров данных разбиений показал, что между тремя кластерами обнаруживается значительная разница в смертности. В то же время дополнительное разбиение этих кластеров до пяти формирует кластеры со схожими значениями смертности. Таким образом, в исследовании удалось выделить 3 группы пациентов с различными профилями клинических показателей, а также различным прогнозом смертности, используя для кластеризации признаки различной природы.

Адаптивные алгоритмы кластерного анализа

Адаптивные алгоритмы обладают способностью изменять свою работу в зависимости от поступающих данных. Распространенными классом таких алгоритмов являются искусственные нейронные сети (ИНС), в работе которых прослеживаются аналогии с биологическими процессами.

Кластеризация с использованием искусственных нейронных сетей

Искусственные нейронные сети — класс математических моделей, которые возникли и развивались в результате изучения нейрофизиологических процессов, а также попыток их моделирования. В классической работе (1943 г.) У. МакКалока (нейрофизиолог) и В. Питтса (математик-логик) описаны элементарный нейрон, работающий по принципу «все или ничего», а также вычислительные возможности ансамблей таких нейронов [44]. Ф. Розенблатт предложил принципиальную схему и реализацию в виде вычислительной машины нейронной сети (рис. 5), названной *персептрон*. Она позволяла распознавать образы (буквы латинского алфавита), обучаясь на тренировочной выборке данных, моделируя тем самым работу головного мозга [10, 61].

В 1969 г. вышла книга М. Мински и С. Пейперта [56], содержавшая выводы о невозможности решения ряда задач с помощью персептронов, в связи с отсутствием надежного алгоритма обучения на тренировочных выборках. Это привело к серьезному охлаждению интереса к ИНС, вплоть до 80-х годов, когда преодоление указанной проблемы послужило толчком к применению ИНС в широком спектре задач «обучения с учителем»

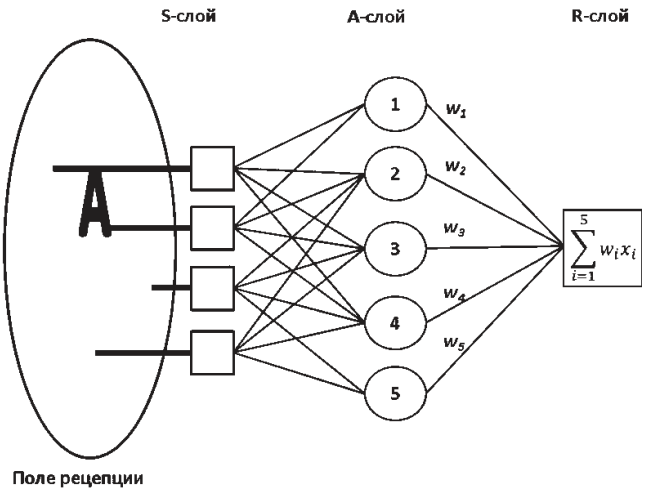

Рис. 5. Персептрон Розенблатта.
S-слой образован рецепторными элементами, активирующимися при условии их совпадения с участком изображения; *A*-слой образован ассоциативными элементами, каждый из которых соединен с частью *S*-элементов и активируется при получении порогового числа сигналов активации последних; при активации элемента слоя *S* или *A* элементу следующего слоя передается сигнал x_i, равный 1, в противном случае передается 0. *R*-слой представлен единственным элементом, выдающим результат 1, если сумма произведений сигналов от предыдущих элементов на соответствующие им веса w_i $\left(\sum_{i=1}^{5} w_i x_i\right)$ больше порогового значения, и –1, если это не так.
При обучении персептрона происходит подбор весов w_i по определенному правилу, что позволяет ему адаптироваться к поступающим данным для решения конкретной задачи.

[1, 63]. В этот период появились работы Т. Кохонена, описывающие модель ИНС, обучавшихся без учителя (сети Кохонена) и применявшихся автором, в частности, для решения задач кластеризации данных [44, 45]. В настоящее время ИНС находят широкое применение как при решении задач кластеризации и классификации [4], так и ряда других задач, включающих, например, анализ временных рядов [65].

Самоорганизующиеся карты (Кохонена)

Одной из реализаций искусственных нейронных сетей Кохонена являются самоорганизующиеся карты (self-organizing maps, SOM). Сеть Кохонена, предназначенная для обработки n-мерных данных, включает в себя слой нейронов, каждый из которых характеризуется n-мерным вектором весов: $\overline{w}_i = (w_{i1}, ..., w_{in})$, где i — номер нейрона (рис. 6). Для образования самоорганизующейся карты каждому нейрону также приписываются векторы $\overline{m}_i$ пространства более низкой размерности (как правило, двумерного: $\overline{m}_i = (m_{i1}, m_{i2})$). Векторы этого пространства являются упорядоченными относительно друг друга в прямоугольную, либо гексагональную решетку [46]. Таким образом, нейроны сети представлены в двух пространствах: первое совпадает с пространством анализируемых объектов, а второе служит для отображения последних в упорядоченную структуру (далее — W и M соответственно). Работа сети делится на два этапа: «обучение» и «картирование». На этапе обучения в ответ на предъявляемые данные осуществляется координированная (согласно их соседству в решетке) подстройка весов нейронов, в результате чего нейроны равномерно распределяются по анализируемой выборке, сохраняя упорядоченность друг с другом. Алгоритм обучения самоорганизующейся карты Кохонена [13]:

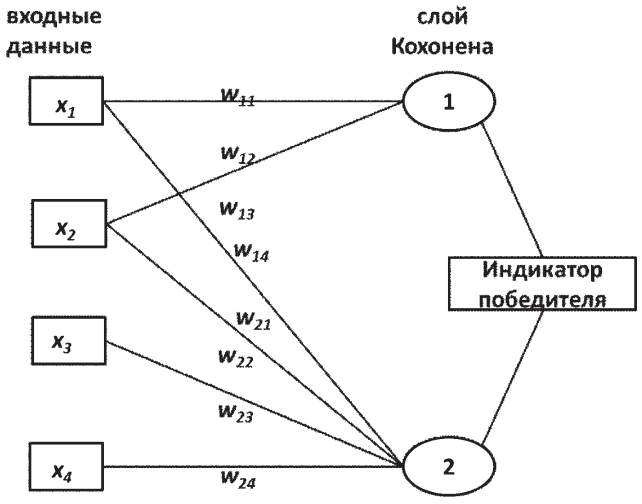

Рис. 6. Сеть Кохонена.
Представлена схема ИНС Кохонена, состоящая из двух нейронов и предназначенная для обработки данных 4-мерного пространства (4 числовых признака). В процессе работы сети вектора весов нейронов $\overline{w}_j = (w_{j1}, ..., w_{j4})$ сравниваются с векторами признаков обрабатываемых объектов $\overline{x} = (x_1, ..., x_4)$, в результате чего выбирается самый близкий к данному объекту по значениям весов нейрон-победитель.

1. Инициализация значений весов нейронов в пространстве W (например, небольшими случайными значениями);

2. Формирование обучающей выборки из кластеризуемых объектов;

3. Для каждого объекта $\overline{x}$ обучающей выборки:

- вычисление расстояния Евклида между объектом $\overline{x}$ и векторами весов нейронов $\overline{w}_i$;
- определение ближайшего к объекту нейрона — победителя;
- подстройка весов нейронов сети по формуле:

$$\overline{w}_i^{t+1} = \overline{w}_i^t + \eta(t) \times h_{i,j(\overline{x})}(t) \times (\overline{x} - \overline{w}_i^t),$$

где:

t — момент дискретного времени предъявления вектора данных $\overline{x}$;

i — порядковый номер нейрона в решетке;

$\overline{w}_i^t$ и $\overline{w}_i^{t+1}$ — векторы весов в момент предъявления вектора данных $\overline{x}$ и после их подстройки.

Функции $\eta(t)$ и $h_{i,j(\overline{x})}(t)$ зависят от времени выполнения алгоритма и убывают с течением времени. Первая из них определяет скорость обучения сети и может быть определена, например, так: $\eta^{(t)} = \eta_0 \times (1 - t/T)$. В этом выражении η_0 — число в диапазоне от 0 до 1, а T, как видно из соотношения, определяет момент времени, в который обучение останавливается [25]. Функция окрестности $h_{i,j(\overline{x})}(t)$, где $j(\overline{x})$ — номер нейрона-победителя, определяет степень вовлечения в процесс подстройки весов его соседей по решетке в пространстве M. Как правило, она имеет вид либо ступенчатый (для нейронов определенной окрестности 1, для всех остальных 0), либо Гауссовой функции от расстояния между нейроном-победителем $j(\overline{x})$ и остальными нейронами i. Можно видеть, что если функция $h_{i,j(\overline{x})}(t)$ выбирается так, что в подстройке весов на каждом шаге будет участвовать только нейрон-победитель, то алгоритм становится похожим на k-средних. Результатом использования функции окрестности является то, что нейроны соседствующие в пространстве M, сохраняют отношение соседства между нейронами в пространстве W [46].

Описанный алгоритм самоорганизующихся карт основан на нескольких принципах, в которых прослеживается влияние нейрофизиологических моделей. Отображение сложного пространства признаков входных объектов на упорядоченную плоскую структуру-решетку подобно связям анализаторов (моторных, зрительных и других) с соответствующими участками коры. Принцип конкуренции лежит в основе выбора нейрона-победителя. Принцип кооперации реализуется с помощью функции окрестности, опре-

деляющей число нейронов-соседей, подстраивающих свои веса для сближения с входным объектом. Принцип синаптического усиления также связан с подстройкой весов нейрона для сближения с объектом тестовой выборки. В результате нее при поступлении в ИНС аналогичного объекта данный нейрон окажется к нему ближе, таким образом, его распознавание будет более эффективно [13].

На этапе картирования осуществляется распределение анализируемых объектов по кластерам, соответствующим наиболее близким нейронам. Наличие представления каждого нейрона в составе решетки в пространстве M позволяет эффективно визуализировать структуру исходных данных. Так возможна визуализация проанализированных данных в виде набора карт, каждая из которых соответствует одному признаку данных. Карты состоят из ячеек, каждая из которых соответствует представлению нейрона в пространстве низкой размерности M. Ячейки закрашиваются согласно цветовой шкале признака, в зависимости от его значения у нейрона, соответствующего данной ячейке (например, в серой шкале — чем темнее ячейка, тем выше значение признака). Таким образом, просматривая карту за картой, можно получить представление о том, какие признаки отличают кластеры внутри данных друг от друга. Другой вариант визуализации предполагает раскраску карты соответственно расстояниям между векторами, закрашивая более темным ячейки, соответствующие более плотно расположенным векторам [46]. Возможны и другие варианты визуализаций.

Применение самоорганизующихся карт

Выявление наиболее существенных факторов, характерных для лиц с суицидальным поведением [48]

Набор данных образован сведениями о 8699 лицах, характеризуемых 606 признаками, собранными в ходе исследованияроли различных факторов в развитии суицидального поведения. Признаки включали социодемографические данные, результаты ответов на вопросы специализированных опросников, сведения о перенесенном в детстве насилии и другие. Использовалась самоорганизующаяся карта с гексагональной решеткой размера 16×12. Среди признаков присутствовал индикатор, сигнализирующий о факте суицидального поведения. Данный признак использовался для раскрашивания карты с целью выявления кластеров, в который вошло наибольшее число лиц, совершавших попытки суицида. В отобранных таким образом кластерах проводился дальнейший анализ для определения признаков, в наибольшей степени выраженных у лиц совершавших суицидальные попытки.

Классификация пациентов, страдающих сколиотической болезнью [58]

Анализировались данные 1776 пациентов со сколиотической болезнью, в качестве признаков использовались восемь значений углов Кобба (определяются по рентгенограммам), характеризующих искривление позвоночника. Нейроны в пространстве M формировали гексагональную решетку. Для сравнения результатов кластеризации с результатами стандартной клинической классификации, строилась карта, которая раскрашивались соответственно преобладающему среди пациентов кластера типу классификационных категорий по Ленке (рис. 7).

Также авторами показано, что после обобщения профилей пациентов с помощью самоорганизующейся карты точность прогнозирования областей проведения хирургической коррекции существенно возрастала. Таким образом, авторам удалось разработать систему клинической диагностики, потенциально обладающую высокой клинической значимостью.

Определение числа кластеров

Две тесно связанные проблемы кластерного анализа — это контроль качества группировки данных в кластеры и определение оптимального числа последних. Число кластеров k, на которые разбиваются входные данные, зачастую является параметром, который требуется задавать алгоритму перед началом работы. Существует грубая эмпирическая оценка числа кластеров: $k = \sqrt{\frac{n}{2}}$, где n — объектов [38]. В работе Milligan и Cooper 1985 г. проведено сравнительное исследование эффективности 30 процедур, многие из которых используют различные

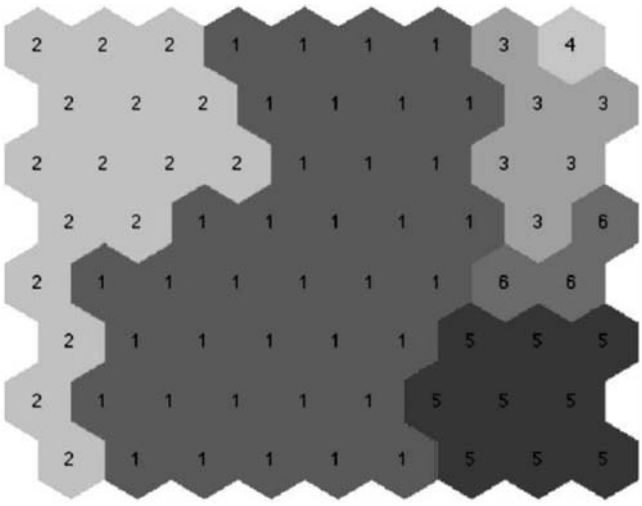

Рис. 7. Гексагональная карта Кохонена, раскрашенная в градациях серого соответственно преобладающей категории клинической классификации Ленке среди пациентов кластеров. Пациенты, относящиеся к одной клинической категории, относятся к близким кластерам. Воспроизведено из [58].

характеристики, вычисляемые для последовательно осуществляемых разбиений с различными k [55]. Более поздняя методика с использованием gap-статистики предлагается в работе [74].

Оценка «качества» кластеризации

Silhouette-статистика

Метод silhouette-статистики [62] относится к техникам анализа «качества» кластеризации объектов. Входными данными для метода являются: матрица расстояний (или сходства) между объектами и данные о распределении объектов по кластерам. Для каждого объекта i рассчитывается показатель:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

где:
$a(i)$ — среднее расстояние от объекта i до объектов кластера к которому он принадлежит;
$b(i)$ — среднее расстояние от объекта i до объектов ближайшего к нему кластера, к которому он не принадлежит.

Показатель s обладает следующим свойством:

$$-1 \le s(i) \le 1.$$

Близость $s(i)$ к 1 свидетельствует о «правильном» отнесении объекта i к его кластеру. Близость к 0 говорит о затруднительности выбора двух соседних кластеров для объекта. В случае же близости к –1 можно говорить о том, что объект должен быть отнесен к ближайшему соседнему кластеру.

Графическое представление, предложенное авторами методики, представляет собой столбчатую диаграмму (в оригинале столбцы ориентированы вдоль абсциссы), где столбцы соответствуют объектам и пропорциональны значениям $s(i$. Столбцы-объекты сгруппированы по кластерам и расположены внутри кластеров в порядке уменьшения $s(i)$. Также на диаграмме для каждого объекта отмечается кластер принадлежности и ближайший соседний кластер. Таким образом, по диаграмме можно делать выводы о качестве распределения объектов по кластерам, а также принимать решение об изменении положения первых относительно вторых.

Максимизация среднего показателя метода Silhouette позволяет рассчитывать на хорошую разделенность кластеров и правильное распределение объектов между ними.

Индекс Данна

Метод Данна [26], аналогично методу силуэта, определяет уровень «качества» процедуры кластеризации. В отличие от последнего, в качестве показателя используется отношение минимального междукластерного расстояния к максимальному внутрикластерному (или диаметру кластера), причем обе эти характеристики могут быть определены различными способами (Δ_i — внутрикластерное расстояние или диаметр; δ_i — межкластерное расстояние; $d(x,y)$ — расстояние между точками (примерами) одного кластера):

1) максимальное внутрикластерное расстояние;
2) среднее расстояние между примерами;
3) среднее расстояние от центра кластера до точки.

Аналогичные варианты оценки δ_i включают: расстояние между двумя ближайшими точками кластеров, среднее расстояние между точками кластеров, расстояние между центроидами кластеров. Ранняя версия индекса Данна предполагала использование первого варианта Δ_i и, соответственно, расстояние между двумя ближайшими точками как межкластерное расстояние (так же и в 5 k-средних). Обобщенные индексы Данна предполагают различные реализации характеристики расстояния или диаметров, в том числе (2) и (3).

Заключение

Кластерный анализ является группой многопараметрических методик анализа, направленных на выявление внутренней структуры данных. Сфера его применения в медико-биологических науках достаточно широка — от обработки данных современных высокопроизводительных молекулярных методик исследований до анализа медико-социальных проблем в области общественного здоровья. Учет преимуществ и недостатков перечисленных методов кластерного анализа необходим для повышения эффективности обработки больших массивов данных.

Список литературы

1. ***Барцев С.И., Охонин В.А.*** Адаптивные сети обработки информации. Препринт N 59Б // 1986.

2. ***Бериков В.С., Лбов Г.С.*** Современные тенденции в кластерном анализе // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». — АНО «Международное образование и наука», 2008.

3. ***Воронцов К.В.*** Лекции по алгоритмам кластеризации и многомерногошкалирования // 2007.

4. ***Ежов А.А., Шумский С.А.*** Нейрокомпьютинг и его применения в экономике и бизнесе. — М.: МИФИ, 1998.

5. ***Ким Д.-О., Мьюллер Ч.У., Клекка У.Р.*** Факторный, дискриминантный и кластерный анализ. — М.: Финансы и статистика, 1989. — 215 с.

6. ***Колмогоров А.Н., Фомин С.В.*** Элементы теории функций и функционального анализа. — М.: Физматлит, 2006.

7. ***Левенштейн В.И.*** Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академий Наук СССР. — 1965. — Т. 163, №4. — С. 845—848.

8. ***Леонов В.П., Гарганеева Н.П.*** Возможности биометрического анализа взаимосвязи соматических показателей и систематики психических расстройств // Сибирский медицинский журнал. — 2001. — №2. — С. 25—32.

9. ***Мандель И.Д.*** Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.

10. ***Назаров М.Г.*** Курс социально-экономической статистики. — М.: Финстатинформ, ЮНИТИ-ДАНА, 2000. — 771 с.

11. ***Похачевский А.Л.*** Временной анализ распределения кардиоинтервалов при нагрузочном тестировании // Патологическая физиология и экспериментальная терапия. — 2011. — №2. — С. 34—40.

12. ***Розенблатт Ф.*** Принципы нейродинамики. Перцептроны и теория механизмов мозга. — М.: Мир, 1965. — 478 с.

13. ***Хайкин С.*** Нейронные сети: полный курс. — М.: Вильямс, 2008.

14. ***Шипунов А.Б.*** Основы теории систематики. — Открытый лицей ВЗМШ, «Книжный дом Университет», 1999. — 56 с.

15. ***Aldenderfer M.S.*** Cluster analysis. — Beverly Hills: Sage Publications, 1984. — 88 p.

16. ***Berkhin P.*** A Survey of Clustering Data Mining Techniques // Grouping Multidimensional Data / J. Kogan, C. Nicholas, M. Teboulle. — Berlin/Heidelberg: Springer-Verlag. — P. 25—71.

17. ***Bhattacharjee A., Richards W.G., Staunton J., Li C., Monti S., Vasa P., Ladd C., Beheshti J., Bueno R., Gillette M., Loda M., Weber G., Mark E.J., Lander E.S., Wong W., Johnson B.E., Golub T.R., Sugarbaker D.J., Meyerson M.*** Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses // Proc. Natl. Acad. Sci. U.S.A. — 2001. — Vol. 98, №24. — P. 13790—13795.

18. ***Bjornsdottir U.S., Holgate S.T., Reddy P.S., Hill A.A., McKee C.M., Csimma C.I., Weaver A.A., Legault H.M., Small C.G., Ramsey R.C., Ellis D.K., Burke C.M., Thompson P.J., Howarth P.H., Wardlaw A.J., Bardin P.G., Bernstein D.I., Irving L.B., Chupp G.L., Bensch G.W., Bensch G.W., Stahlman J.E., Karetzky M., Baker J.W., Miller R.L., Goodman B.H., Raible D.G., Goldman S.J., Miller D.K., Ryan J.L., Dorner A.J., Immermann F.W., O'Toole M.*** Pathways Activated during Human Asthma Exacerbation as Revealed by Gene Expression Patterns in Blood // PLoS ONE. — 2011. — Vol. 6, №7. — e21902.

19. ***Burgel P.R., Paillasseur J.L., Peene B., Dusser D., Roche N., Coolen J., Troosters T., Decramer M., Janssens W.*** Two Distinct Chronic Obstructive Pulmonary Disease (COPD) Phenotypes Are Associated with High Risk of Mortality // PLoS ONE. — 2012. — Vol. 7, №12. — e51048.

20. ***Cho R.J., Campbell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D.J., Davis R.W.*** A genome-wide transcriptional analysis of the mitotic cell cycle // Mol. Cell. — 1998. — Vol. 2, №1. — P. 65—73.

21. ***Clark M.C. Hall, L.O. Goldgof, D.B. Clarke, Laurence P. Velthuizen, R.P. Silbiger, M.S.*** MRI segmentation using fuzzy clustering techniques // IEEE Engineering in Medicine and Biology Magazine. — 1994. — Vol. 13, №5. — P. 730—742.

22. ***Czekanowski J.*** Zur differential Diagnose der Neandertalgruppe // Korrespbl. Dtsch. Ges. Anthropol. — 1909. — №40. — P. 44—47.

23. ***Deza E., Deza M.*** Dictionary of distances. — Amsterdam, the Netherlands.

24. ***Dimitrakopoulou K., Dimitrakopoulou K., Tsimpouris C., Papadopoulos G., Pommerenke C., Wilk E., Sgarbas K.N., Schughart K., Bezerianos A.*** Dynamic gene network reconstruction from gene expression data in mice after influenza A (H1N1) infection // Journal of Clinical Bioinformatics. — 2011. — Vol. 1, №1. — P. 27.

25. ***Du K.-L.*** Clustering: A neural network approach // Neural Networks. — 2010. — Vol. 23, №1. — P. 89—107.

26. ***Dunn J.C.*** A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters // Journal of Cybernetics. — 1973. — Vol. 3, №3. — P. 32—57.

27. ***Eisen M.B., Spellman P.T., Brown P.O., Botstein D.*** Cluster analysis and display of genome-wide expression patterns // Proc. Natl. Acad. Sci. U.S.A. — 1998. — Vol. 95, №25. — P. 14863—14868.

28. ***El Agha M., M. Ashour W.*** Efficient and Fast Initialization Algorithm for K-means Clustering // International Journal of Intelligent Systems and Applications. — 2012. — Vol. 4, №1. — P. 21—31.

29. ***Estivill-Castro V.*** Why so many clustering algorithms // ACM SIGKDD Explorations Newsletter. — 2002. — Vol. 4, №1. — P. 65—75.

30. ***Ferreira L., Hitchcock D.B.*** A Comparison of Hierarchical Methods for Clustering Functional Data // Communications in Statistics — Simulation and Computation. — 2009. — Vol. 38, №9. — P. 1925—1949.

31. ***Forgy E.W.*** Cluster analysis of multivariate data: efficiency versus interpretability of classifications // Biometrics. — 1965. — Vol. 21. — P. 768—769.

32. ***Gentleman R.*** Bioinformatics and computational biology solutions using R and Bioconductor. — New York: Springer Science + Business Media, 2005.

33. ***Gevaert O., Daemen A., De Moor B., Libbrecht L.*** A taxonomy of epithelial human cancer and their metastases // BMC Med. Genomics. — 2009. — Vol. 2. — P. 69.

34. ***Hamerly G., Elkan C.*** Alternatives to the k-means algorithm that find better clusterings. — ACM Press, 2002. — P. 600.

35. ***Hamming R.W.*** Error detecting and error correcting codes // Bell System Tech. J. — 1950. — Vol.2, №29. — P.147—160.

36. ***Han J., Kamber M., Pei J.*** Data mining: concepts and techniques. 1st Ed. — Amsterdam, 2001.

37. ***Han J., Kamber M., Pei J.*** Data mining: concepts and techniques. 2nd Ed. — Morgan Kaufmann Publishers, 2006.

38. ***Hartigan J.A.*** Clustering algorithms. — New York: Wiley, 1975.

39. ***Holmberg K., Nord C.E.*** Numerical taxonomy and laboratory identification of Actinomyces and Arachnia and some related bacteria // J. Gen. Microbiol. — 1975. — Vol. 91, №1. — P. 17—44.

40. ***Jain A.K., Murty M.N., Flynn P.J.*** Data clustering: a review // ACM Computing Surveys. — 1999. — Vol. 31, №3. — P. 264—323.

41. ***Jianchang Mao, Jain A.K.*** A self-organizing network for hyperellipsoidal clustering (HEC) // IEEE Transactions on Neural Networks. — 1996. — Vol. 7, №1. — P. 16—29.

42. ***Kaufman L.*** Finding groups in data: an introduction to cluster analysis. — New York: Wiley, 1990. — 342 p.

43. ***Kettenring J.R.*** The Practice of Cluster Analysis // Journal of Classification. — 2006. — Vol. 23, №1. — P. 3—30.

44. ***Kohonen T.*** Self-organization and associative memory. — Berlin.

45. ***Kohonen T.*** Self-organized formation of topologically correct feature maps // Biological Cybernetics. — 1982. — Vol. 43, №1. — P. 59—69.

46. ***Kohonen T.*** Self-organizing maps. — Berlin.

47. ***Kwekel J.C., Desai V.G., Moland C.L., Branham W.S., Fuscoe J.C.*** Age and sex dependent changes in liver gene expression during the life cycle of the rat // BMC Genomics. — 2010. — Vol. 11, №1. — P. 675.

48. ***Leiva-Murillo J.M., Lopez-Castroman J., Baca-Garcia E.*** EURECA Consortium. Characterization of Suicidal Behaviour with Self-Organizing Maps // Computational and Mathematical Methods in Medicine. — 2013. — Vol. 2013. — P. 1—9.

49. ***Lloyd S.*** Least squares quantization in PCM // IEEE Transactions on Information Theory. — 1982. — Vol. 28, №2. — P. 129—137.

50. ***Martin R., Riley P.S., Hollis D.G., Weaver R.E., Krichevsky M.I.*** Characterization of some groups of gram-negative nonfermentative bacteria by the carbon source alkalinization technique // J. Clin. Microbiol. — 1981. — Vol. 14, №1. — P. 39—47.

51. ***McCulloch W., Pitts W.*** A Logical Calculus of Ideas Immanent in Nervous Activity // 1943. — №5. — P. 115—133.

52. ***McQuitty L.L.*** Hierarchical Linkage Analysis for the Isolation of Types // Educational and Psychological Measurement. — 1960. — Vol. 20, №1. — P. 55—67.

53. ***McShane L.M., Radmacher M.D., Freidlin B., Yu R., Li M.C., Simon R.*** Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data // Bioinformatics. — 2002. — Vol. 18, №11. — P. 1462—1469.

54. ***Milligan G.W.*** An examination of the effect of six types of error perturbation on fifteen clustering algorithms // Psychometrika. — 1980. — Vol. 45, №3. — P. 325—342.

55. ***Milligan G.W., Cooper M.C.*** An examination of procedures for determining the number of clusters in a data set // Psychometrika. — 1985. — Vol. 50, №2. — P. 159—179.

56. ***Minsky M.L., Papert S.*** Perceptrons. — Oxford, England: M.I.T. Press, 1969.

57. ***Moore W.C., Meyers D.A., Wenzel S.E., Teague W.G., Li H., Li X., D'Agostino R.Jr., Castro M., Curran-Everett D., Fitzpatrick A.M., Gaston B., Jarjour N.N., Sorkness R., Calhoun W.J., Chung K.F., Comhair S.A., Dweik R.A., Israel E., Peters S.P., Busse W.W., Erzurum S.C., Bleecker E.R.*** Identification of Asthma Phenotypes Using Cluster Analysis in the Severe Asthma Research Program // American Journal of Respiratory and Critical Care Medicine. — 2009. — Vol. 181, №4. — P. 315—323.

58. ***Phan P., Mezghani N., Wai E.K., de Guise J., Labelle H.*** Artificial neural networks assessing adolescent idiopathic scoliosis: comparison with Lenke classification // The Spine Journal. — 2013.

59. ***Pilcher C.D., Wong J.K., Pillai S.K.*** Inferring HIV Transmission Dynamics from Phylogenetic Sequence Relationships // PLoS Medicine. — 2008. — Vol. 5, №3. — e69.

60. ***Regard J.B., Sato I.T., Coughlin S.R.*** Anatomical Profiling of G Protein-Coupled Receptor Expression // Cell. — 2008. — Vol. 135, №3. — P. 561—571.

61. ***Rosenblatt F.*** The perceptron: a probabilistic model for information storage and organization in the brain // Psychol. Rev. — 1958. — Vol. 65, №6. — P. 386—408.

62. ***Rousseeuw P.J.*** Silhouettes: A graphical aid to the interpretation and validation of cluster analysis // Journal of Computational and Applied Mathematics. — 1987. — Vol. 20. — P. 53—65.

63. ***Rumelhart D.E., Hinton G.E., Williams R.J.*** Learning representations by back-propagating errors // Nature. — 1986. — Vol. 323, №6088. — P. 533—536.

64. ***Schena M., Shalon D., Davis R.W., Brown P.O.*** Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray // Science. — 1995. — Vol. 270, №5235. — P. 467—470.

65. ***Shugai J., Guzhva A., Dolenko S., Persiantsev I.*** An algorithm for construction of a hierarchical neural network complex for time series analysis and its application for studying sun-earth relations // 8th International Conference «Pattern Recognition and Image Analysis: New Information Technologies» (PRIA-8-2007): Conference Proceedings (Yoshkar-Ola). — 2007. — Vol. 2. — P. 355—358.

66. ***Silver M.*** Scales of Measurement and Cluster Analysis: An Application Concerning Market Segments in the Babyfood Market // The Statistician. — 1995. — Vol. 44, №1. — P. 101.

67. ***Sneath P.H.*** The application of computers to taxonomy // J. Gen. Microbiol. — 1957. — Vol. 17, №1. — P. 201—226.

68. ***Sneath P.H.A.*** Numerical taxonomy.

69. ***Sokal R.R., Michener C.D.*** A Statistical Method for Evaluating Systematic Relationships. — University of Kansas, 1958.

70. ***Steinhaus H.*** Sur la division des corps materiels en parties // Bull. Acad. Polon. Sci. — 1956. — Vol. III, №4. — P. 801—804.

71. ***Stevens S.S.*** On the Theory of Scales of Measurement // Science. 1946. — Vol. 103, №2684. — P. 677—680.

72. ***Tan P.-N., Steinbach M., Kumar V.*** Introduction to data mining. — Boston: Pearson Addison Wesley, 2005.

73. ***Tavazoie S., Hughes J.D., Campbell M.J., Cho R.J., Church G.M.*** Systematic determination of genetic network architecture // Nat. Genet. — 1999. — Vol. 22, №3. — P. 281—285.

74. ***Tibshirani R., Walther G., Hastie T.*** Estimating the number of clusters in a data set via the gap statistic // Journal of the Royal Statistical Society: Series B (Statistical Methodology). — 2001. — Vol. 63, №2. — P. 411—423.

75. ***Ting Su, Dy J.*** A deterministic method for initializing K-means clustering // IEEE Comput. Soc. — P. 784—786.

76. ***Trion R.G.*** Cluster analysis. — London: Ann Arbor Edwards Bros, 1939.

Поступила 12.10.13

Сведения об авторах:

Московцев Алексей Александрович, канд. мед. наук, вед. науч. сотр. лаб. тромболизиса и тромбогенеза ФГБУ «НИИОПП» РАМН, доцент каф. общей патологии и патофизиологии ФГБОУ ДПО «РМАПО» Минздрава РФ

Доленко Сергей Анатольевич, канд. физ.-мат. наук, старш. науч. сотр. отд. оперативного космического мониторинга НИИЯФ им. Д.В. Скобельцына

Савина Галина Дмитриевна, старш. преп. каф. общей патологии и патофизиологии ФГБОУ ДПО «РМАПО» Минздрава РФ