СТАТИСТИЧЕСКИЕ МЕТОДЫ И АНАЛИЗ ДАННЫХ

А.В. Ермолаев
(*Москва*)

МЕТОДЫ АНАЛИЗА И ВИЗУАЛИЗАЦИИ СТРУКТУРЫ ДАННЫХ О БЛИЗОСТИ

В статье приводится обзор моделей анализа и визуализации матриц близости. Рассматриваются три класса моделей: пространственные, теоретико-множественные и графы. Осуждаются формальные и содержательные аспекты применения моделей к анализу социологических данных.

Ключевые слова: матрица близости, многомерное шкалирование, классификация, иерархическая классификация, аддитивная кластеризация, качественный факторный анализ, ультраметрические деревья, аддитивные деревья, сети.

Социальные явления характеризуются большим количеством взаимосвязанных переменных, зачастую неподдающихся непосредственному измерению. Этим латентным переменным на практике ставится в соответствие множество эмпирических индикаторов. Тем самым социологические данные имеют сложную структуру и одна из исследовательских задач – компактное ее описание.

Обычно социолог имеет дело с данными типа «объект-признак», но нередко возникает потребность изучения данных иного вида, в частности, таких, которые можно рассматривать как *данные о близости* объектов. Эти данные представляют собой матрицу (или

Андрей Владимирович Ермолаев – аспирант Института социологии РАН. E-mail: ermolaev_av@mail.ru.

несколько матриц), строки и столбцы которой отвечают некоторым интересующим исследователя объектам. На пересечении *i*-й строки и *j*-го столбца этой матрицы стоит оценка того, насколько сходны или различны *i*-й и *j*-й объекты.

Существуют различные способы получения данных о близости (см., например, [1; 2]). Это могут быть как прямые оценки респондентов о сходстве или различии объектов, так и результат агрегирования данных типа «объект-признак». В качестве данных о близости могут рассматриваться корреляционные матрицы, построенные для объектов или признаков, социометрические данные, таблицы мобильности и т.п.

Существует целый ряд методов анализа данных о близости. В литературе наиболее часто встречается разделение этих моделей на *пространственные модели*, *теоретико-множественные модели* и *графы*.

Пространственные модели включают в себя широкий класс методов многомерного шкалирования. Они позволяют представить объекты как точки в некотором многомерном пространстве, образованном *непрерывными* латентными характеристиками. Обычно предполагается, что размерность пространства невелика (2 или 3), что позволяет представить структуру данных графически в простой, интуитивно понятной форме. Чем ближе расположены точки в пространстве, тем более схожи соответствующие объекты и наоборот.

В рамках *теоретико-множественного подхода* предполагается, что объекты характеризуются набором *дискретных* признаков, т.е. обладающих счетным множеством значений. Обычно это множество конечно, и каждое значение можно обозначить натуральным числом или буквой (дискретным признаком является, например, национальность; «1» означает «русский», «2» – «француз» и т.д.). Каждому значению признака отвечает определенная характеристика объектов («быть русским», «быть французом» и т.д.), которую можно рассматривать как класс или множество

объектов, обладающих этой характеристикой. Так, очевидно, характеристика «русский» задает определенный класс людей – тех, которые имеют указанную национальность.

Графы являются еще одним мощным инструментов анализа и визуализации структуры данных о близости. Граф представляет собой диаграмму, состоящую из точек (вершин) и соединяющих их линий (ребер). Вершины графа соответствуют объектам, а ребра отражают наиболее существенные связи между объектами, что позволяет наглядно представить структуру данных.

Отметим, что теоретико-множественный подход и графы в значительной степени пересекаются. Например, ряд теоретико-множественных моделей позволяет применять графы для визуализации результатов, а модели, использующие теорию графов, в свою очередь можно интерпретировать в терминах теоретико-множественного подхода.

Пространственные модели

В основе модели многомерного шкалирования лежит понятие *метрического пространства*. Предполагается, что существует пространство невысокой размерности, в котором объекты можно представить как точки. Оси пространства соответствуют характеристикам объектов, а проекции точек на оси пространства – значениям характеристик. В этом пространстве вводится метрика d_{ij}, т.е. функция, которая каждой паре точек i и j ставит в соответствие некоторое число, которое называется расстоянием между точками. Функция расстояния должна обладать рядом свойств, которые называются *метрическими аксиомами*:

$$0 \le d_{ii} \le d_{ij}$$
$$d_{ij} = d_{ji}$$
$$d_{ij} + d_{jk} \ge d_{ik}$$

Задачу, решаемую с помощью метода многомерного шкалирования, можно сформулировать следующим образом. Пусть задана

матрица различий $\{\delta_{ij}\}$ между N объектами, где элемент матрицы δ_{ij} соответствует оценке различия между объектами i и j. Необходимо найти такие оценки координат точек в некотором пространстве заданной размерности R, чтобы расстояния между точками $\{d_{ij}\}$ наилучшим образом соответствовали $\{\delta_{ij}\}$ по некоторому критерию, например, сумме квадратов ошибок или проценту объясненной моделью дисперсии исходных данных.

Выбор метрики

Вид метрики d_{ij} определяет механизм агрегирования индивидуальных значений характеристик объектов в оценку различия между ними и наряду с предположением о существовании пространства характеристик объектов является важным элементом модели.

Наибольшее распространение получила степенная метрика Минковского, которая имеет следующий вид:

$$d_{ij} = \left[\sum_{r=1}^{R} \left| x_{ir} - x_{jr} \right|^{p} \right]^{\frac{1}{p}}.$$

Изменяя параметр p, можно получить семейство метрик с различными свойствами. При $p = 1$ метрика Минковского принимает следующий вид:

$$d_{ij} = \sum_{r=1}^{R} \left| x_{ir} - x_{jr} \right|.$$

Эта функция расстояния известна в литературе как *прямоугольная метрика*. Аттнив [3] продемонстрировал, что прямоугольная метрика адекватна в тех случаях, когда объекты описываются небольшим количеством признаков, хорошо различимых респондентами, и объекты могут сравниваться независимо по каждому из этих признаков.

При $p = 2$ метрика Минковского принимает вид хорошо известной функции *расстояния Евклида*:

$$d_{ij} = \sqrt{\sum_{r=1}^{R} (x_{ir} - x_{jr})^2}.$$

В отличие от прямоугольной метрики Евклидово расстояние адекватно в ситуации, когда респонденты не в состоянии четко выделить какие-то конкретные характеристики объектов и сравнивают их как единое целое [4].

По мере увеличения параметра p характеристики, по которым объекты различаются наиболее значимо, начинают приобретать больший вес в интегральной оценке близости [5]. Предельным случаем является *метрика доминирования*, получаемая при $p = \infty$:

$$d_{ij} = \max \left| x_{ir} - x_{jr} \right|.$$

Подобную метрику следует применять в случае, если при сравнении двух объектов основное внимание уделяется той характеристике, по которой объекты сильнее всего различаются [6].

В литературе также встречаются работы с использованием других метрик. Например, пространства с метрикой Римана [7] хорошо зарекомендовали себя при моделировании процессов зрительного восприятия [8]. Кокс и Кокс [9] предложили модель, в которой объекты размещаются на поверхности сферы.

При выборе той или иной метрики необходимо руководствоваться предположениями о природе объектов и процессе, стоящим за оценками близости между ними. Если же предположения о механизме агрегирования характеристик объектов в общую оценку различия между ними отсутствуют, то Евклидова метрика является лучшим выбором. Кроме того, она допускает непосредственную интерпретацию и интуитивно понятна [10].

Метрическое многомерное шкалирование

Формальная задача поиска координат точек в Евклидовом пространстве при известной матрице расстояний между ними была решена в работе Юнга и Хаусхолдера [11]. Торгерсон [12] использовал предложенный ими метод для решения задачи шкалирова-

ния, когда оценки различий предполагаются равными расстояниям в Евклидовом пространстве небольшой размерности:

$$\delta_{ij} = d_{ij},$$

где $d_{ij}^2 = \sum_{r=1}^{R} (x_{ir} - x_{jr})^2$.

Такое предположение является очень жестким, поскольку опирается на то, что оценки различий получены по шкале отношений. Менее жесткой является следующая модель, в которой различия соответствуют расстояниям с точности до некоторой аддитивной константы:

$$\delta_{ij} = d_{ij} + c.$$

В данном случае предполагается, что оценки различий сделаны по интервальной шкале, в которой начало шкалы не задано. Одними из первых проблему оценки аддитивной константы затронули Мессик и Абельсон [13]. Аналитическое решение проблемы было предложено Лингосом [14], а также Каиллиезом [15].

Метрический алгоритм Торгерсона сейчас используется в неметрических алгоритмах многомерного шкалирования, основанных на итеративных процедурах, для оценки стартовых значений параметров модели.

Неметрическое многомерное шкалирование

Настоящим прорывом в области методов шкалирования стало появление модели *неметрического многомерного шкалирования*, в соответствии с которой предполагается, что исходные данные $\{\delta_{ij}\}$ измерены в ранговой шкале. Поскольку задан только порядок близостей, то задача сводится к поиску таких координат, чтобы расстояния $\{d_{ij}\}$ сохранили тот же порядок, что и в исходных данных, т.е.

$$d_{ij} \cong f(\delta_{ij}),$$

где f – монотонная функция (т.е. такая, что $\delta_{ij} < \delta_{kl} \Rightarrow f(\delta_{ij}) \le f(\delta_{kl})$ для всех i, j, k, l).

Первое решение этой проблемы была дано Шепардом [16; 17]. В его алгоритме точки сначала размещались в пространстве размерностью $(N-1)$, а затем на каждом шаге конфигурация проектировалась с наименьшими потерями в пространство меньшей размерности.

Краскал [18; 19] предложил иной, более формальный, алгоритм неметрического многомерного шкалирования. В отличие от подхода Шепарда алгоритм Краскала размещал точки в пространстве заданной размерности таким образом, чтобы расстояния между ними наилучшим образом соответствовали исходным данным. Это достигалось путем минимизации функции соответствия модели исходным данным. Она названа Краскалом STRESS (STandardize REsidual Sum of Squares) и имеет следующий вид:

$$\mathrm{STRESS} = \left[\frac{\sum_{i,j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{ij} d_{ij}^2} \right]^{1/2},$$

где $\{\hat{d}_{ij}\}$ – последовательность чисел, наиболее близкая расстояниям $\{d_{ij}\}$ при условии, что порядок $\{\hat{d}_{ij}\}$ соответствует порядку $\{\delta_{ij}\}$. Для нахождения $\{\hat{d}_{ij}\}$ Краскалом была разработана процедура, названная им монотонной регрессией.

Практически одновременно с работами Краскала свой вариант решения задачи неметрического многомерного шкалирования предложил Гуттман [20], назвав его анализом пространств наименьшей размерности (Smallest Space Analysis). Алгоритм поиска решения, разработанный Гуттманом, в целом похож на алгоритм Краскала за исключением ряда технических деталей.

Теоретическое и эмпирическое сравнение этих алгоритмов было проведено Лингосом и Роскамом [21].

Интересное решение проблемы предложил Джонсон [22]. В его подходе используется функция соответствия, которая не требует вычисления $\{\hat{d}_{ij}\}$.

Такане и др. [23] представил функцию, названную SSTRESS (Squared STRESS), отличающуюся от функции STRESS Краскала тем, что в ней применяются квадраты расстояний. Использование квадратов расстояния в функции SSTRESS позволило значительно повысить эффективность процедур поиска решений. Однако данный подход приводит к искажению оценок близостей, поскольку большие расстояния приобретают больший вес.

Обобщения модели многомерного шкалирования

Рассмотренные методы многомерного шкалирования предполагают наличие одной матрицы сходства/различия. Однако люди могут по-разному оценивать близости. Встает вопрос, как действовать в том случае, если имеется несколько разных матриц сходства/различия, данных субъектами. Конечно, полученные матрицы могут быть агрегированы в одну общую. Однако при этом теряется значительная часть информации об индивидуальных особенностях восприятия.

Революционная модель *индивидуального многомерного шкалирования* была предложена Блоксомом [24], а также Кэрроллом и Чангом [25], которая позволяет наряду с общей картиной получать оценки индивидуальных особенностей. В рамках модели предполагается, что характеристики объектов, по которым оцениваются близости объектов, неодинаково важны для разных респондентов, т.е. функция расстояния имеет следующий вид:

$$d_{ijk}^2 = \sum_{r=1}^{R} w_{kr}(x_{ir} - x_{jr})^2,$$

где w_{kr} – вес или важность латентной характеристики r для респондента k.

Метод индивидуального многомерного шкалирования дает возможность получить общее пространство объектов, а также пространство индивидуальных весов. Общее пространство позволяет рассмотреть структуру данных в целом. Индивидуальные

веса показывают, насколько различаются индивидуальные данные, и разрешают реконструировать индивидуальные пространства восприятия из общего.

Такер [26], а также Кэрролл и Чанг [27] предложили *обобщенную модель* индивидуального многомерного шкалирования. В этой модели индивидуальные пространства определяются не только весами, но и ориентацией осей.

Теоретико-множественные модели

При анализе матриц близости исследователь должен внимательно относиться к выбору метода решения задачи. Так, обдумывая вопрос о возможности использования того или иного пространственного метода, он должен учитывать, что в ряде случаев предположение о непрерывности характеристик, лежащих в основе суждений о сходстве или различии, может быть неадекватным. Например, о близости можно судить по таким характеристикам, как пол, национальность, политическая ориентация и т.п., которые естественно моделировать с помощью дискретных признаков.

Модели Тверски

Тверски [28] предположил модель, в которой сходство s_{ij} между объектами i и j можно определить как функцию от количеств общих и различных признаков объектов.

Пусть ($\mathbf{f}_i = \{f_{im}\}$) множество дискретных признаков, характеризующих объект i, такое, что $f_{im} = 1$, если объект i обладает признаком m, и $f_{im} = 0$ в противном случае. Обозначим ($\mathbf{f}_i \cap \mathbf{f}_j$) количество признаков, присущих обоим объектам, или их общие признаки, а ($\mathbf{f}_i - \mathbf{f}_j$) количество признаков, которыми обладает объект i и не обладает объект j.

Было предложено два варианта определения близости: *контрастная модель* (contrast model):

$$s_{ij} \cong \theta g(\mathbf{f}_i \cap \mathbf{f}_j) - \alpha g(\mathbf{f}_i - \mathbf{f}_j) - \beta g(\mathbf{f}_j - \mathbf{f}_i)$$

и *модель отношений* (ratio model):

$$s_{ij} \cong \frac{g(\mathbf{f}_i \cap \mathbf{f}_j)}{g(\mathbf{f}_i \cap \mathbf{f}_j) - \alpha g(\mathbf{f}_i - \mathbf{f}_j) - \beta g(\mathbf{f}_j - \mathbf{f}_i)},$$

где g – некоторая монотонная функция, θ, α и β – неотрицательные параметры.

Интерпретация модели Тверски проста: чем большим количеством общих признаков и меньшим количеством различных признаков характеризуется пара объектов, тем больше сходство между ними.

Модель Тверски обобщает и подводит теоретическую базу под другие модели. Например, при $\alpha = \beta = 0$ первая из моделей обращается в модель общих признаков, лежащую в основе метода аддитивной кластеризации [29], а при $\theta = 0$ и $\alpha = \beta$ – в модель различных признаков [16].

Иерархическая кластеризация

В самом простом случае предполагается, что признаки не пересекаются, т.е. один и тот же объект не может обладать двумя признаками одновременно. Эта модель соответствует классификации объектов на некоторое количество непересекающихся однородных классов. Под однородностью в общем смысле понимается, что объекты, принадлежащие одному классу, должны быть более схожи, чем объекты из разных классов. Однако такой подход не всегда отвечает реальности.

Признаки могут образовывать иерархическую структуру. В этом случае любые два признака либо не пересекаются, либо один из них является подмножеством другого, т.е. объекты, обладающие первым признаком, обязательно обладают и вторым. Например, имеет место следующее включение признаков друг в друга: «животные» ⊃ «млекопитающие» ⊃ «приматы» ⊃ «шимпанзе» (каждый шимпанзе является приматом, каждый примат – млекопитающим и т.д.). Иерархическую структуру можно рассматривать как

несколько вложенных друг в друга простых разбиений. Так, для каждого уровня иерархии можно получить соответствующее простое разбиение объектов на непересекающиеся классы.

В модели *иерархической классификации* предполагается, что объекты можно разнести на непересекающиеся классы несколькими способами, причем полученные в результате простые разбиения иерархически организованы, т.е. любые два класса либо не пересекаются, либо один из них является подмножеством другого.

Наиболее распространен агломеративный алгоритм иерархической кластеризации, который заключается в следующем. На первом шаге каждый объект рассматривается как тривиальный кластер. Затем находится пара наиболее близких объектов. Эти объекты объединяются и рассматриваются как новый объект. Вычисляются расстояния между новым объектом и всеми остальными объектами. Процедура повторяется снова до тех пор, пока все объекты не будут объединены в один кластер.

Основное различие между алгоритмами иерархического кластерного анализа заключается в выборе функции расстояния между кластером и объектом или между двумя кластерами. В литературе предложено множество функций расстояния между кластерами [30]. Различные функции расстояния между кластерами приводят к разным результатам, и это является проблемой использования методов иерархического кластерного анализа.

Процесс иерархической кластеризации естественным образом можно представить с помощью дендрограммы, которая является простым способом визуализации структуры данных о близости. Этот факт связывает иерархическую кластеризацию с моделями графов, которые рассматриваются в следующем разделе статьи.

Аддитивная кластеризация

Шепард и Арабье [29] предложили модель *аддитивной кластеризации*, где сходство между двумя объектами определяется как сумма весов их общих признаков:

$$s_{ij} \cong \sum_{m=1}^{M} w_m f_{im} f_{jm} + c.$$

Как уже отмечалось выше, эта модель является частным случаем контрастной модели Тверски, где $\alpha = \beta = 0$, а g – аддитивная функция.

Для оценки параметров модели Шепард и Арабье предложили алгоритм ADCLUS (ADditive CLUStering) [29], идея которого заключается в поиске всех возможных подмножеств объектов, с последующей оценкой их весов и отбрасыванием тех подмножеств, веса которых ниже некоторого порогового значения.

Более простой и эффективный способ решения задачи поиска пересекающихся признаков и их весов был предложен Миркиным и Трофимовым, известный в отечественной литературе как *качественный факторный анализ* [31]. Данный алгоритм, названный методом последовательного исчерпания матрицы близости, последовательно определяет подмножества объектов и соответствующие веса [32].

Арабье и Кэрролл [33] предложили алгоритм MAPCLUS (MAthematical Programming CLUStering). Этот алгоритм размещал объекты на заранее заданное количество пересекающихся подмножеств и оценивал их веса таким образом, чтобы полученные оценки близости наилучшим образом соответствовали исходным данным.

Кэрролл и Арабье [34] обобщили модель ADCLUS для случая несколько матриц близости одновременно, INDCLUS. В этой модели, аналогично модели индивидуального многомерного шкалирования, объекты обладают одинаковыми общими признаками, а индивидуальные различия моделируются с помощью разных значений весов для каждой матрицы k:

$$s_{ijk} \cong \sum_{m=1}^{M} w_{km} f_{im} f_{jm} + c_k .$$

Аддитивная кластеризация не накладывает ограничений на структуру признаков и является наиболее общей моделью с точки зрения предположения о структуре данных. Отсутствие ограничений на моделируемую структуру признаков приводит к тому, что для одной и той же матрицы близости может быть получено несколько решений, эквивалентных с точки зрения формальных показателей качества модели.

Графы

Граф состоит из множества вершин и множества отрезков, соединяющих эти вершины, которые называются ребрами графа. Графы удобно отображать с помощью диаграмм, в которых вершины представлены точками, а ребра, соединяющие две вершины, – линиями между соответствующими точками.

Связанным называют граф, если для каждой пары вершин существует путь, по которому можно попасть из одной вершины в другую. Если каждому ребру приписано некоторое неотрицательное число, то такой граф называется взвешенным, а само число – длиной ребра. Расстоянием между двумя вершинами графа называется сумма длин ребер, образующих путь из одной вершины в другую.

При использовании графов для представления структуры данных о близости, вершины графа соответствуют объектам. В качестве расстояния между объектами на графе обычно применяется *геодезическое* расстояние, т.е. длина кратчайшего пути из одной вершины в другую. Это расстояние является метрикой, поскольку удовлетворяет метрическим аксиомам.

Деревья

Особым видом графов являются *деревья*. Деревьями называются связанные графы, не содержащие циклов. Отсутствие циклов означает, что для каждой пары точек – вершин графа –

существует единственный путь, по которому можно попасть из одной точки в другую. Одним из особых видов деревьев, которые получили широкое распространение в социальных науках, являются графы, в которых кроме вершин, соответствующих объектам, присутствуют и другие вершины, которые можно интерпретировать как классы объектов.

Как отмечалось выше, процедура иерархической кластеризации может быть представлена в виде дендрограмы или дерева. Джонсон [35] один из первых, кто рассмотрел процесс иерархической кластеризации как отображение исходной матрицы близости в пространство с метрикой, удовлетворяющей так называемому ультраметрическому неравенству[1]:

$$d_{ij} \leq \max(d_{ik}, d_{kj}).$$

Из этого неравенства следует, что расстояния от всех объектов одного кластера до объектов другого кластера равны. Причем межклассовые расстояния всегда больше внутриклассовых расстояний. Например, пусть объекты i и j принадлежат одному классу, а объект k – другому классу, тогда расстояния между ними должны удовлетворять следующему условию:

$$d_{ij} \leq d_{ik} = d_{kj}.$$

Ситуация, когда

$$d_{ij} = d_{ik} = d_{kj}$$

означает, что все три объекта i, j и k принадлежат одному классу.

С содержательной точки зрения это условие означает, что модель ультраметрического дерева не учитывает внутриклассовые расстояния. Это ограничение достаточно жесткое, поэтому ультраметрическое дерево не всегда адекватно репрезентирует структуру исходных данных. В связи с этим ряд авторов предложили менее жесткую модель, которая была названа моделью *аддитивного дерева* [36].

[1] В связи с этим деревья, удовлетворяющие этому неравенству, называют *ультраметрическими деревьями*.

Модель аддитивного дерева отличается тем, что в этом дереве каждому ребру может быть поставлено в соответствие любое неотрицательное число. Метрика, соответствующая этой модели, удовлетворяет аддитивному неравенству [37], т.е. расстояния между любыми объектами i, j, k и l удовлетворяют соотношению:

$$(d_{ij} + d_{kl}) \leq \max[(d_{ik} + d_{jl}),(d_{jk} + d_{il})].$$

Ультраметрическое неравенство удовлетворяет аддитивному, поэтому иерархическое дерево является частным случаем аддитивного дерева [37]. Рассмотрим четыре объекта i, j, k и l, тогда расстояния между ними удовлетворяют следующему соотношению

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{jk} + d_{il}.$$

Модель аддитивного дерева накладывает менее жесткие ограничения на структуру данных. Так, в отличие от ультраметрического дерева, в модели аддитивного дерева среднее расстояние между объектами из одного класса должно быть меньше, чем среднее расстояние от них до объекта из другого кластера. Аддитивное дерево адекватно представляет данные с более общей структурой, чем ультраметрическое дерево.

Наибольшую популярность получил алгоритм ADDTREE (ADDitive TREE), предложенный Саттасом и Тверски [36]. Алгоритм состоит из двух основных этапов: (1) поиска топологии дерева; (2) оценки длин ребер дерева. Картер разработал программу, реализующую алгоритм ADDTREE [38], а также эффективный с вычислительной точки зрения алгоритм GTREE [39].

Кэрролл и Пружански [40] использовали метод чередующихся наименьших квадратов для построения модели аддитивного дерева.

Десоет [41] предложил метод для нахождения матрицы расстояний $\{d_{ij}\}$, наилучшим образом соответствующей матрице близости и удовлетворяющей ультраметрическому или аддитивному неравенству. После того, как матрица $\{d_{ij}\}$ получена, на ее

основе может быть построено аддитивное дерево. Кроме того, этот алгоритм может быть использован для анализа матриц близости с пропущенными данными [42].

Обобщенные модели деревьев

Одно из направлений обобщения моделей деревьев представлено моделями, которые дают возможность обрабатывать несколько матриц близости, аналогично модели индивидуального многомерного шкалирования. Кэрролл и др. [43] предложили модель, позволяющую выявить общую структуру данных, а также получить оценки индивидуальных различий за счет введения индивидуальных весов для ребер дерева.

Другое направление связано с обобщением моделей деревьев для анализа данных с более сложной структурой. Кэрролл и Пружански [40] представили *модель множественных деревьев*, исходя из предположения, что в основе данных о близости может лежать не одна иерархическая структура, а несколько. Модель имеет следующий вид:

$$\delta_{ij} \cong \sum_{k=1}^{K} d_{ij}^{k},$$

где δ_{ij} – различие между объектами i и j, d_{ij}^{k} – расстояния для дерева k.

Картер и Тверски [44] предложили *расширенную модель аддитивного дерева* EXTREE. В ней иерархическая структура дополнена небольшим количеством неиерархических признаков, которые позволяют моделировать пересечения между иерархически организованными признаками.

Схожую с содержательной точки зрения модель предложили Макаренков и Легендре [45], где пересечения представлены как дополнительные ребра графа, соединяющие вершины дерева. Лапонте [46] приводит примеры других методов для анализа данных с подобной структурой.

Сетевые модели

Во многих случаях модели деревьев хорошо описывают исходные данные. Однако иерархическая структура только один из возможных и достаточно жесткий тип структур. Не все структуры могут быть адекватно представлены с помощью моделей деревьев. Все рассмотренное выше является попыткой обобщить иерархические модели для анализа более сложных структур данных, но эти модели все же предполагают, что в основе данных лежит иерархическая структура. *Сетевые модели* позволяют моделировать данные о близости с помощью графов общего вида, которые, в отличие от моделей деревьев, допускают наличие циклов.

Любая матрица близости, удовлетворяющая метрическим аксиомам, может быть описана с помощью полного графа, т.е. графа, в котором каждая пара вершин соединена ребром. Это не представляет особого интереса с точки зрения решения задач анализа и визуализации данных о близости, поскольку это аналогично размещению N точек в пространстве размерности $(N - 1)$. Как и в других методах, здесь одна из задач – получение наиболее простой модели, которая наилучшим образом описывает исходные данные.

В последние годы одним из интенсивно развивающихся направлений являются сетевые модели, основанные на геодезической метрике [47; 48]. В вычислении геодезического расстояния используются только те ребра, которые образуют кратчайший путь из одной вершины в другую. Следовательно, исключение остальных ребер не приведет к изменению расстояний между объектом. Задача алгоритмов сводится к тому, чтобы максимально упростить граф, исключив эти несущественные с точки зрения расстояния ребра. Основным критерием в данном случае, как и в других моделях, является соответствие модельных расстояний исходным данным.

Гибридные модели

Описанные выше модели основаны на различных предположениях о структуре данных, и многие авторы склонны рассматривать эти модели не как альтернативные, а как дополняющие друг друга, мотивируя это тем, что изучаемые явления, как правило, гораздо более сложны, чем предполагает каждая из моделей в отдельности [49]. Поэтому в последние годы был предложен ряд походов, включающих в себя достоинства пространственных и непространственных моделей.

Рассматривая методы многомерного шкалирования, Торгерсон [50] отметил, что возможны ситуации, когда близость между объектами определяется как непрерывными факторами, так и дискретными характеристиками. Если объекты принадлежат к q непересекающимся однородным классам, то их можно изобразить как q точек в $(q-1)$-мерном пространстве. Непрерывные факторы в этом случае будут представлены дополнительными осями. Основываясь на этой идее, Дегерман [51] одним из первых предложил процедуру, которая поворачивает исходное R-мерное пространство таким образом, чтобы первые $(q-1)$ осей наилучшим образом отражали межгрупповые различия для q групп, где $q \le R$. Эта задача аналогична задаче дискриминантного анализа. Оставшиеся $(R-q)$ осей соответствуют подпространству непрерывных факторов.

Кэрролл и Пружански [40] обобщили модель множественных деревьев следующим образом:

$$\delta_{ij} \cong \sum_{k=1}^{K} d_{ij}^{\,k} + d_{ij}^{\,E},$$

где $d_{ij}^{\,k}$ – расстояние для дерева k, $d_{ij}^{\,E}$ – расстояние в Евклидовом пространстве между этими же точками.

Наварро и Ли [52] предложил модель, объединяющую модели многомерного шкалирования и аддитивной кластеризации:

$$s_{ij} \cong \left(\sum_{m=1}^{M} w_m f_{im} f_{jm} \right) - \left(\sum_{r=1}^{R} (x_{ir} - x_{jr})^p \right)^{\frac{1}{p}} + c.$$

Чатурведи и Кэрролл [53] на основе моделей INDSCAL и INDCLUS предложили обобщение модели для нескольких матриц близости:

$$s_{ijk} \cong \left(\sum_{m=1}^{M} w_{km} f_{ik} f_{jk} \right) - \left(\sum_{r=1}^{R} w_{kr} (x_{ir} - x_{jr})^p \right)^{\frac{1}{p}} + c_k.$$

***

Описанные методы анализа и визуализации структуры данных о близости между объектами основаны на различных предположениях о структуре данных и нередко могут приводить на практике к различным результатам. Как было показано, это связано с тем, что каждый из рассмотренных методов акцентирует различные аспекты структуры данных.

ЛИТЕРАТУРА

1. *Coxon A.P.M.* The User's Guide to Multidimensional Scaling. London: Heinemann Educational Books Ltd, 1982.

2. *Дэйвисон М.* Многомерное шкалирование: Методы наглядного представления данных. М.: Финансы и статистика, 1982.

3. *Attneave F.* Dimensions of Similarity // American Journal of Psychology. 1950. No. 63. P. 546–554.

4. *Shepard R.N.* Representation of Structure in Similarity Data: Problems and Prospects // Psychometrika. 1974. Vol. 39. P. 373–421.

5. *Coombs C.H.* A Theory of Data. N.Y.: Wiley, 1964.

6. *Shepard R.N.* Attention and the Metric Structure of the Stimulus Space // Journal of Mathematical Psychology. 1964. Vol. 1. P. 54–87.

7. *Pieszko H.* Multidimensional Scaling in Riemannian Space // Journal of Mathematical Psychology. 1975. Vol. 14. P. 449–477.

8. *Терехина А.Ю.* Анализ данных методами многомерного шкалирования. М.: Наука, 1986.

9. *Cox T.F., Cox M.A.A.* Multidimensional Scaling on a Sphere // Communicational Statistics. 1991. No. 20. P. 2943–2953.

10. *Torgerson W.S.* Theory and Methods of Scaling. N.Y.: Wiley, 1958.

11. *Young F.W., Householder A.S.* Discussion of a Set of Points in Terms of Their Mutual Distances // Psychometrika. 1938. Vol. 3. P. 19–22.

12. *Torgerson W.S.* Multidimensional Scaling: Theory and Method // Psychometrika. 1952. Vol. 17.

13. *Messick S.M., Abelson R.P.* The Additive Constant Problem in Multidimensional Scaling // Psychometrika. 1956. Vol. 21. P. 1–15.

14. *Lingoes J.C.* Some Boundary Conditions for Monotone Analysis of Symmetric Matrices // Psychometrika. 1971. Vol. 36. P. 195–203.

15. *Cailliez F.* The Analytical Solution of Additive Constant Problem // Psychometrika. 1983. Vol. 48. P. 305–308.

16. *Restle F.* A Metric and an Ordering on Sets // Psychometrika. 1959. Vol. 24. P. 207–220.

17. *Shepard R.N.* The Analysis of Proximity Data with Unknown Distance Function // Psychometrika. 1962. Vol. 27.

18. *Kruskal J.B.* Multidimensional Scaling by Optimizing Goodness of Fit to a Non-metric Hypothesis // Psychometrika. 1964. Vol. 29.

19. *Kruskal J.B.* Non-metric Multidimensional Scaling: a Numerical Method // Psychometrika. 1964. Vol. 29.

20. *Guttman L.* A General Nonmetric Technique for Finding the Smallest Coordinate Space for a Configuration of Points // Psychometrika. 1968. Vol. 33. P. 469–504.

21. *Lingoes J.C., Roskam E.E.* A Mathematical and Empirical Analysis of Two Multidimensional Scaling Algorithms // Psychometrika. 1973. Vol. 38.

22. *Johnson R.M.* Pairwise Nonmetric Multidimensional Scaling // Psychometrika. 1973. Vol. 38. P. 11–18.

23. *Takane Y., Young F.W., DeLeeuw J.* Non-metric Individual Difference Multidimensional Scaling: an Alternating Least Squares Method with Optimal Scaling Features // Psychometrika. 1977. Vol. 42. P. 7–67.

24. *Bloxom B.* An Alternative Method of Fitting a Model of Individual Differences in Multidimensional Scaling // Psychometrika. 1974. Vol. 39. P. 365–367.

25. *Carroll J.D., Chang J.J.* Analysis of Individual Differences in Multidimensional Scaling via N-way Generalization of «Eckart–Young» Decomposition // Psychometrika. 1970. Vol. 35. P. 283–319.

26. *Tucker L.R.* Relation Between Multidimensional Scaling and Three-mode Factor Analysis // Psychometrika. 1972. Vol. 37. P. 3–27.

27. *Carroll J.D.* Individual Differences and Multidimensional Scaling // Multidimensional Scaling: Theory and Applications in the Behavioral Sciences / Ed. by R.N. Shepard, A.K. Romney, S.B. Nerlove. N.Y.: Seminar Press, 1972.

28. *Tversky A.* Features of Similarity // Psychological Review. 1977. No. 84(4). P. 227–252.

29. *Shepard R.N., Arabie P.* Additive Clustering Representation of Similarities as Combination of Discrete Overlapping Properties // Psychological Review. 1979. No. 86(2). P. 87–123.

30. *Дюран Б., Оделл П.* Кластерный анализ. М.: Статистика, 1977.

31. *Куперштох В.Л., Миркин Б.Г., Трофимов В.А.* Метод наименьших квадратов в анализе качественных признаков // Проблемы анализа дискретной информации. Новосибирск: ИЭиОПП СО АН СССР, 1976. С. 4–23. Вып. 2.

32. *Миркин Б.Г.* Группировки в социально-экономических исследованиях. М.: Финансы и статистика, 1985.

33. *Arabie P., Carroll J.D.* MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model // Psychometrika. 1980. Vol. 45(2). P. 211–235.

34. *Carroll J.D., Arabie P.* INDCLUS: An Individual Difference Generalization of the ADCLUS Model and the MAPCLUS Algorithm // Psychometrika. 1982. Vol. 48. P. 157–169.

35. *Johnson S.C.* Hierarchical Clustering Schemes // Psychometrika. 1967. Vol. 32(3). P. 241–254.

36. *Sattath S., Tversky A.* Additive Similarity Trees // Psychometrika. 1977. Vol. 42(3). P. 319–345.

37. *Buneman P.* The Recovery of Trees from Measures of Dissimilarity // Mathematics in the Archeological and Historical Sciences. Edinburgh, UK: Edinburgh University Press, 1971.

38. *Corter J.E.* ADDTREE/P: A PASCAL Program for Fitting Additive Trees Based on Sattath and Tversky's ADDTREE Algorithm // Behavioral Research Methods & Instrumentation. 1982. No. 14. P. 353–354.

39. *Corter J.E.* An Efficient Metric Combinatorial Algorithm for Fitting Additive Trees // Multivariate Behavioral Research. 1998. No. 33(2). P. 249–271.

40. *Carroll J.D., Pruzansky S.* Discrete and Hybrid Models for Scaling // Similarity and Choice. Bern: Hans Huber, 1980.

41. *De Soete G.* A Least Squares Algorithm for Fitting Additive Trees to Proximity Data // Psychometrika. 1983. Vol. 48(4). P. 621–626.

42. *De Soete G.* Additive-tree Representation of Incomplete Dissimilarity Data // Quality and Quantity. 1984. No. 18. P. 387–393.

43. *Carroll J.D., Clark L., DeSarbo W.S.* The Representation of Three-way Proximities Data by Single and Multiple Tree Structure Models // Journal of Classification. 1984. Vol. 1. P. 25–74.

44. *Corter J.E., Tversky A.* Extended Similarity Trees // Psychometrika. 1986. Vol. 51(3). P. 429–451.

45. *Makarenkov V., Legendre P.* Improving the Additive Tree Representation of a Dissimilarity Matrix Using Reticulations // Data Analysis, Classification, and Related Methods. Berlin: Springer, 2000. P. 35–46.

46. *Lapointe F.-J.* How to Account for Reticulation Events in Phylogenic Analysis: a Comparison of Distance-based Methods // Journal of Classification. 2000. Vol. 17. P. 175–184.

47. *Hutchinson J.W.* NETSCAL: A Network Scaling Algorithm for Nonsymmetric Proximity Data // Psychometrika. 1989. Vol. 54. P. 25–51.

48. *Klauer K.C., Carroll J.D.* A Mathematical Programming Approach to Fitting General Graph // Journal of Classification. 1989. Vol. 6. P. 247–270.

49. *Краскал Дж.* Взаимосвязь между многомерным шкалированием и кластер-анализом // Классификация и кластер. М.: Мир, 1980.

50. *Torgerson W.S.* Multidimensional Scaling of Similarity // Psychometrika. 1965. Vol. 30(4).

51. *Degerman R.* Multidimensional Analysis of Complex Structure: Mixture of Class and Quantitative Variation // Psychometrika. 1970. Vol. 35(4). P. 475–491.

52. *Navarro D.J., Lee M.D.* Combining Dimensions and Features in Similarity-based Representations // Advances in Neural Information Processing Systems / Ed. by S. Becker, S. Thrun, K. Obermayer. N.Y.: MIT Press, 2003. P. 67–74. No. 15.

53. *Carroll J.D., Chaturvedi A.* A General Approach to Clustering and Multidimensional Scaling of Two-way, Three-way, or Higher-way Data // Geometric Representation of Perceptual Phenomena / Ed. by R.D. Luce, M. D'Zmura, D.D. Hoffman, G. Iverson, A.K. Romney. Mahwah, NJ: Erbaum, 1995. P. 295–318.