МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Е.П. ЧУРАКОВ

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Рязань 2014

Министерство образования и науки Российской Федерации

Рязанский государственный радиотехнический университет

Е.П. ЧУРАКОВ

Многомерные статистические методы

Часть 2

Учебное пособие

Рязань 2014

УДК 519.237(075.8)
ББК 22.172я73

Многомерные статистические методы: учеб. пособие / Е.П. Чураков, Рязан. гос. радиотехн. ун-т; Рязань, 2014. ч.2 48с.

Вторая часть пособия является естественным продолжением курса «Многомерные статистические методы», первая часть которого под таким же названием опубликована ранее [13]. Основное внимание во второй части пособия уделено методам кластерного анализа и многомерного шкалирования. Приводится ряд конкретных исследований, проведенных с использованием этих подходов. Предназначено для студентов специальности 080116 “Математические методы в экономике” и направлений 010400 “Прикладная математика и информатика” с профилем подготовки “Математическое и информационное обеспечение экономической деятельности” и 080500 “Бизнес-информатика”.

Библиогр.: 23 назв.

Многомерный анализ, векторные случайные величины, кластерный анализ, многомерное шкалирование

Печатается по решению редакционно-издательского совета Рязанского государственного радиотехнического университета.

Рецензент: кафедра ЭиММ Рязанского государственного радиотехнического университета (зав. кафедрой В.К. Клочко)

Ч у р а к о в Евгений Павлович

Многомерные статистические методы

Редактор М.Е.Цветкова
Корректор С.В. Макушина
Подписано в печать18.07.14. Формат бумаги 60х84 1/16.
Бумага писчая. Печать трафаретная. Усл. печ. л 3,0.
Тираж 50 экз. Заказ
Рязанский государственный радиотехнический университет.
390005, Рязань, ул. Гагарина, 59/1.
Редакционно-издательский центр РГРТУ.

© Рязанский государственный
радиотехнический университет, 2014

4. Основы кластерного анализа

4.1. Назначение кластерного анализа

Кластерный анализ в настоящее время представляет собой хорошо развитую методологию решения задач классификации данных и обладает разнообразными подходами и обширным алгоритмическим обеспечением. Достаточно цельное изложение существа кластерного анализа содержится в работах [1] – [5]. Кратко изложим наиболее фундаментальные положения классификационных проблем, решаемых средствами кластерного анализа.

Существо кластерного анализа заключается в следующем. Пусть некоторое экономическое (для определенности) явление характеризуется k признаками $X_1, X_2, ..., X_k$, т.е. многомерной величиной $\mathbf{X} = [X_1, X_2, ..., X_k]^T$. Например: $X_1 -$ уровень среднегодовой выработки на одного работника; $X_2 -$ уровень фондоотдачи; $X_3 -$ размер оборотных производственных средств; $X_4 -$ размер затрат на выпуск единицы товарной продукции; $X_5 -$ численность промышленно-производственного персонала; $X_6 -$ рентабельность продукции; $X_7 -$ уровень энерговооруженности труда и т.п. Далее полагаем, что эти k признаков были зарегистрированы на n объектах (например, на n предприятиях), т.е. каждый признак измерен n раз. В итоге формируется матрица $\overline{\mathbf{X}}$ экспериментальных данных

$$\overline{\mathbf{X}} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{bmatrix} \in R^{k \times n} \qquad (4.1)$$

где x_{ij} - значение i-го признака на j-м объекте (при j-м измерении), $i = \overline{1,k},\ j = \overline{1,n}.$ Произвольный j-й столбец этой матрицы, таким образом, представляет собой значение вектора $\mathbf{X}$ на j-м объекте. Так как все компоненты вектора $\mathbf{X}$ могут иметь различные смысловое содержание и единицы измерения, элементы матрицы $\overline{\mathbf{X}}$ обычно центрируются и нормируются, т. е. осуществляется переход к стандартизованной форме представления данных. Для этого, что уже отмеча-

лось, вычисляются эмпирическое среднее каждого признака, несмещенная оценка его дисперсии и среднеквадратичное отклонение, соответственно:

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}\,, \quad s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \overline{x_i}\right)^2, \quad s_i = \sqrt{s_i^2}\,, \qquad i = \overline{1,k}. \tag{4.2}$$

С использованием этих величин строится матрица

$$\mathbf{Y} = \begin{bmatrix} \dfrac{x_{11} - \overline{x_1}}{s_1} & \dfrac{x_{12} - \overline{x_1}}{s_1} & \cdots & \dfrac{x_{1n} - \overline{x_1}}{s_1} \\ \dfrac{x_{21} - \overline{x_2}}{s_2} & \dfrac{x_{22} - \overline{x_2}}{s_2} & \cdots & \dfrac{x_{2n} - \overline{x_2}}{s_2} \\ \cdots & \cdots & \cdots & \cdots \\ \dfrac{x_{k1} - \overline{x_k}}{s_k} & \dfrac{x_{k2} - \overline{x_k}}{s_k} & \cdots & \dfrac{x_{kn} - \overline{x_k}}{s_k} \end{bmatrix}. \tag{4.3}$$

Каждый столбец этой матрицы, следовательно, является стандартизованным вектором $\mathbf{X}$, зарегистрированным на соответствующем объекте. Пусть $\mathbf{Y}_i \in R^k$, $i = \overline{1,n}$, i-й столбец матрицы $\mathbf{Y}$ (i-й вектор). В пространстве R^k эти векторы формируют n точек, каждая из которых отображает положение соответствующего ей объекта в пространстве R^k. Множество этих точек, как и матрицу, обозначим символом $\mathbf{Y}$. *Задача кластерного анализа* заключается в том, чтобы все множество $\mathbf{Y}$, состоящее из n точек в пространстве R^k, разбить на $m < n$ однородных в некотором смысле групп (кластеров, таксонов, образов), где m в общем случае неизвестно, таким образом, чтобы каждая точка $\mathbf{Y}_i$ была отнесена только к одному кластеру и чтобы точки, отнесенные к разным кластерам, опять же, в определенном смысле, разнились. Под кластером, таким образом, понимается полученная в результате разбиения группа объектов (векторов, точек), обладающих определенной общностью. Методы построения таких групп и называются *кластерным анализом* (иногда численной таксономией или распознаванием образов с самообучением).

4.2. Расстояние между объектами

Одним из начальных понятий кластерного анализа является “расстояние” между двумя произвольными элементами множества $\overline{\mathbf{Y}}$. Это понятие может носить различный содержательный смысл, но обычно основывается на аксиоматике метрического пространства и удовле-

творяет соответствующим аксиомам метрики (например, [6, 7]). Наиболее часто используют следующие определения расстояний между элементами $\mathbf{Y}_i, \mathbf{Y}_j \in \overline{\mathbf{Y}}$.

1. Гельдерово (или Минковского) расстояние (l_q – расстояние):

$$\rho_q(\mathbf{Y}_i, \mathbf{Y}_j) = \left(\sum_{s=1}^{k} \left| y_{si} - y_{sj} \right|^q\right)^{\frac{1}{q}}, \quad q \geq 1.$$

Здесь и далее y_{si} – s-й компонент вектора $\mathbf{Y}_i$..

2. Если пространство R^k является евклидовым, т.е. в нем определено скалярное произведение $(\mathbf{Y}_i, \mathbf{Y}_j)$ векторов $\mathbf{Y}_i,\ \mathbf{Y}_j \in \overline{\mathbf{Y}}$, то обычно используют евклидово расстояние:

$$\rho_e(\mathbf{Y}_i, \mathbf{Y}_j) = \sqrt{(\mathbf{Y}_i - \mathbf{Y}_j, \mathbf{Y}_i - \mathbf{Y}_j)} = \left\| \mathbf{Y}_i - \mathbf{Y}_j \right\| = \sqrt{(\mathbf{Y}_i - \mathbf{Y}_J)^T (\mathbf{Y}_i - \mathbf{Y}_j)},$$

где здесь и далее символ $\|\mathbf{Y}\|$ означает норму вектора $\mathbf{Y}$. Евклидово расстояние является частным случаем гельдерова при $q = 2.$

3. Расстояние Михаланобиса (обобщенное евклидово расстояние):

$$\rho_M(\mathbf{Y}_i, \mathbf{Y}_j) = \sqrt{(\mathbf{Y}_i - \mathbf{Y}_J)^T \mathbf{W}^{-1} (\mathbf{Y}_i - \mathbf{Y}_j)},$$

где $\mathbf{W}$ – некоторая симметрическая положительно определенная матрица. В случае $\mathbf{W} = \mathbf{E}$: $\mathbf{E}$ – единичная матрица, расстояние Михаланобиса совпадает с евклидовым.

4. “Взвешенное” евклидово расстояние следует из предыдущего, если принять $\mathbf{W}^{-1} = diag[w_{ii}]$, т.е. если матрица $\mathbf{W}^{-1}$ является диагональной с элементами главной диагонали w_{ii}, $i = \overline{1,k}.$

5. При $q = \infty$ гельдерово расстояние порождает так называемое супремум расстояние (l_∞ – расстояние):

$$l_\infty(\mathbf{Y}_i,\ \mathbf{Y}_J) = \sup_{s = \overline{1,k}} \left[\left| y_{si} - y_{sj} \right| \right].$$

6. При $q = 1$ получают так называемое $city - block$ расстояние:

$$l_1(Y_i, Y_j) = \sum_s \left| y_{si} - y_{sj} \right|.$$

Наиболее распространенным в кластерном анализе является евклидово расстояние. Это объясняется как наглядностью геометрического смысла евклидова расстояния, так и математическим "комфортом" работы с ним. Его целесообразно применять, если компоненты произвольного вектора $\mathbf{Y}_i$ однородны по своему физическому смыслу и все они одинаково важны при решении вопроса об отнесении вектора $\mathbf{Y}_i$ к определенному кластеру. Если последние условия не выполняются, переходят к взвешенному расстоянию или расстоянию Михаланобиса с более гибкими возможностями уравновешивания значимости и содержания различных компонентов векторов $\mathbf{Y}_i$, $\mathbf{Y}_j$.

4.3. Расстояние между кластерами

Вторым важным понятием кластерного анализа является расстояние между кластерами. Это понятие вводится также не единственным образом и имеет различный содержательный смысл. Снова ограничим рассмотрение наиболее характерными определениями.

Предположим, что на множестве $\mathbf{Y}$ выделены два кластера – $\mathbf{S}_l$ и $\mathbf{S}_m$, причем $\mathbf{S}_l \subset \mathbf{Y}$, $\mathbf{S}_m \subset \mathbf{Y}$, $\mathbf{S}_l \cap \mathbf{S}_m = \otimes$, $\otimes -$ пустое множество. Далее допускаем, что кластер $\mathbf{S}_l$ содержит n_l объектов (точек), а кластер $\mathbf{S}_m - n_m$ объектов. Пусть

$$\overline{\mathbf{Y}_l} = \frac{1}{n_l} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \mathbf{Y}_i, \qquad \overline{\mathbf{Y}_m} = \frac{1}{n_m} \sum_{\mathbf{Y}_j \in \mathbf{S}_m} \mathbf{Y}_j -$$

арифметические средние векторов, формирующих кластеры $\mathbf{S}_l$ и $\mathbf{S}_m$ соответственно ("центры тяжестей" кластеров). Наиболее используемыми в кластерном анализе оказываются следующие виды расстояний между кластерами.

1. Минимальное локальное расстояние по принципу "ближайшего соседа":

$$d_{\min}(\mathbf{S}_l, \mathbf{S}_m) = \min_{\mathbf{Y}_i \in \mathbf{S}_l,\, \mathbf{Y}_j \in \mathbf{S}_m} \rho(\mathbf{Y}_i, \mathbf{Y}_j).$$

В этом случае на множествах $\mathbf{S}_l, \mathbf{S}_m$ ищутся две минимально удаленные друг от друга точки и расстояние между кластерами отождествляется с расстоянием между этими точками.

2. Максимальное локальное расстояние по принципу "дальнего соседа":

$$d_{\max}(\mathbf{S}_l,\mathbf{S}_m)=\max_{\mathbf{Y}_i\in\mathbf{S}_l,\,\mathbf{Y}_j\in\mathbf{S}_m}\rho(\mathbf{Y}_i,\mathbf{Y}_j).$$

Это определение аналогично предыдущему, но основано на наиболее удаленных друг от друга точках.

3. Расстояние по "центрам тяжести":

$$d_{\text{цц}}(\mathbf{S}_l,\mathbf{S}_m)=\rho(\overline{\mathbf{Y}_l},\overline{\mathbf{Y}_m}).$$

4. Среднее расстояние между кластерами (расстояние по принципу "средней связи"):

$$d_{\text{сс}}(\mathbf{S}_l,\mathbf{S}_m)=\frac{1}{n_l n_m}\sum_{\mathbf{Y}_i\in\mathbf{S}_l}\sum_{\mathbf{Y}_j\in\mathbf{S}_m}\rho(\mathbf{Y}_i,\mathbf{Y}_j).$$

5. Статистическое расстояние между кластерами:

$$d_{\text{сцц}}(\mathbf{S}_l,\mathbf{S}_m)=\frac{n_l n_m}{n_l+n_m}(\overline{\mathbf{Y}_l}-\overline{\mathbf{Y}_m})^T(\overline{\mathbf{Y}_l}-\overline{\mathbf{Y}_m}).$$

Во всех этих определениях расстояние между точками $\mathbf{Y}_i$ и $\mathbf{Y}_j$ понимается в одном из рассмотренных выше принципов.

4.4. Индивидуальные свойства кластеров

Третий вид характеристик, используемых в кластерном анализе, направлен на отображение индивидуальных свойств отдельных кластеров или всего множества $\mathbf{Y}$ в целом. Так как существо этих характеристик сохраняется и при рассмотрении отдельного кластера, и всего множества $\mathbf{Y}$, рассмотрим их применительно к $\mathbf{Y}$.

1. Величина

$$r_d=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\rho(\mathbf{Y}_i,\mathbf{Y}_j),\quad \mathbf{Y}_i,\mathbf{Y}_j\in\mathbf{Y},$$

называется *общим* расстоянием множества $\mathbf{Y}$.

2. Величина

$$\overline{r_d}=\frac{1}{N}r_d,\ N=(n^2-4)/2$$

называется *средним* расстоянием множества $\mathbf{Y}$.

В этих определениях традиционно принимается $\rho(\mathbf{Y}_i,\mathbf{Y}_j)=\rho(\mathbf{Y}_j,\mathbf{Y}_i)$. и $\rho(\mathbf{Y}_i,\mathbf{Y}_i)=0.$

3. Квадратная матрица размерностью n

$$\overline{\mathbf{R}_\mathbf{Y}}=\sum_{i=1}^{n}(\mathbf{Y}_i-\overline{\mathbf{Y}})(\mathbf{Y}_i-\overline{\mathbf{Y}})^T,\quad \overline{\mathbf{Y}}=\frac{1}{n}\sum_{1}^{n}\mathbf{Y}_i$$

называется *матрицей рассеивания* множества $\mathbf{Y}$ (иногда матрицей суммы квадратов).

4. След матрицы $\overline{\mathbf{R}_\mathbf{Y}}$ называют *статистическим рассеянием* множества $\mathbf{Y}$ (суммой квадратов отклонений, дисперсий)

$$r_{\text{ñò}} = Sp\overline{\mathbf{R}_\mathbf{Y}} = Sp\left(\sum_{i=1}^{n}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)^T\right) = \sum_{i=1}^{n}\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right)^T\left(\mathbf{Y}_i - \overline{\mathbf{Y}}\right).$$

5. Определитель $\left|\mathbf{R}_\mathbf{Y}\right|$ матрицы $\mathbf{R}_\mathbf{Y}$ называют *статистическим рассеянием, соответствующим определителю*, и обозначают

$$r_d = \left|\mathbf{R}_\mathbf{Y}\right|.$$

4.5. Целевые функции классификации

Разбиение множества $\mathbf{Y}$ на отдельные классы может быть осуществлено различными способами. Многие из них описаны в различных документах по применению пакетов прикладных программ (например, [8], [9]). Снова ограничим рассмотрение изложением некоторых достаточно общих принципов.

Чтобы один способ предпочесть остальным, в рассмотрение вводят некую целевую функцию (показатель качества классификации), значение которой явилось бы количественной оценкой степени совершенствования того или иного метода классификации. Тогда поиск наилучшего варианта разбиения множества на отдельные классы сводится к поиску экстремума этой целевой функции. Разбиение, при котором целевая функция достигает экстремального значения, и будет решением задачи классификации.

В настоящее время отсутствует строгая формализованная система рекомендаций по выбору целевой функции. Этот выбор подчиняется интуитивно-эмпирическим соображениями и в каждом частном случае осуществляется индивидуально. Достаточно распространенными оказываются следующие целевые функции.

Пусть на множестве $\mathbf{Y}$ выделены кластеры $\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_m$; $\mathbf{S}_i \in \mathbf{Y}, i = \overline{1, m}$. Тогда качество классификации можно характеризовать следующими функциями.

1. Сумма внутри классовых дисперсий:

$$J_1 = \sum_{l=1}^{m}\sum_{\mathbf{Y}_i \in \mathbf{S}_l} \rho^2(\mathbf{Y}_i, \overline{\mathbf{Y}_l}).$$

Если расстояние $\rho\left(\mathbf{Y}_i, \overline{\mathbf{Y}_l}\right)$ понимается в евклидовом смысле, то внутренняя сумма в J_1 есть не что иное, как статистическое рассеяние $r_{\text{ñò}}(l)$ l-го кластера (см. соответствующее определение), и J_1 представляет собой сумму статистических рассеяний всех m кластеров:

$$J_1 = \sum_{l=1}^{m} r_{\text{ñò}}\ (l).$$

2. Сумма квадратов попарных внутриклассовых расстояний между объектами:

$$J_2 = \sum_{l=1}^{m} \sum_{\mathbf{Y}_i, \mathbf{Y}_j \in \mathbf{S}_{li}} \rho^2(\mathbf{Y}_i, \mathbf{Y}_j).$$

Или:

$$J_3 = \sum_{l=1}^{m} \frac{1}{n_l} \sum_{\mathbf{Y}_i, \mathbf{Y}_j \in \mathbf{S}_{lj}} \rho^2(\mathbf{Y}_i, \mathbf{Y}_j).$$

В большинстве практических задач целевые функции J_1, J_2, J_3 приводят к одним и тем же разбиениям множества $\mathbf{Y}$ на m классов.

3. Обобщенная внутриклассовая дисперсия.

Пусть

$$\mathbf{R}_l = \sum_{\mathbf{Y}_i \in \mathbf{S}_l} (\mathbf{Y}_i - \overline{\mathbf{Y}_l})(\mathbf{Y}_i - \overline{\mathbf{Y}_l})^T$$

вляется матрицей рассеяния l-го кластера. Тогда рассеяние всех m кластеров опишется матрицей $\sum_{l=1}^{m} \mathbf{R}_l$. В качестве целевой функции принимается:

$$J_4 = \left| \sum_{l=1}^{m} \mathbf{R}_l \right| = \left| \sum_{l=1}^{m} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} (\mathbf{Y}_i - \overline{\mathbf{Y}_l})(\mathbf{Y}_i - \overline{\mathbf{Y}_l})^T \right| \cdot$$

Из сопоставления с J_1 функцию J_4 можно назвать обобщенным статистическим рассеянием, соответствующим определителю. Эту функцию рекомендуется использовать в случаях, когда возникают предположения о том, что наблюдения $\mathbf{Y}_i$ принадлежат пространству размерности, меньшей k.

4. Целевые функции J_1, J_2, J_3, J_4 используют при априори известном числе m классов, на которое разбивается множество $\mathbf{Y}$. В практических задачах это число обычно неизвестно и подлежит определению в процессе проведения классификации. В подобных случаях целевая функция должна быть ориентирована на это обстоятельство и надлежащим образом его учитывать. Одним из приемов формирования целевой функции при неизвестном числе m является использование аддитивной структуры, приводящей к построению целевой функции в виде суммы двух слагаемых, одно из которых является невозрастающей функцией числа классов m и характеризует, как правило, внутриклассовый разброс наблюдений, а второе является неубывающей функцией числа классов m. В частности, предлагается [10, 11]:

$$J_5 = b_1 + b_2,$$

$$b_1 = \sum_{l=1}^{m} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \rho(\mathbf{Y}_i, \overline{\mathbf{Y}_l}), \qquad b_2 = cm,$$

где $c > 0-$ некоторая константа, определяющая потери при возрастании числа классов. Наилучшее разбиение (выбор числа классов и разнесение точек $\mathbf{Y}_i$ по класссам) соответствует минимуму целевой функции J_5.

4.6. Основные алгоритмы кластерного анализа

Прежде чем рассматривать основные алгоритмы кластерного анализа, еще раз обратим внимание читателя на существование большего числа статистических пакетов прикладных программ, средствами которых успешно решаются многие задачи обработки экспериментальных наблюдений, в том числе и сводящиеся к кластерной методологии. Среди пакетов подобной направленности отметим такие распространенные у нас в стране, как STATISTICA, SPSS, STADIA, STATGRAPHICS и др. Опыт и техника применения этих пакетов для решения задач кластеризации известны из многочисленных публикаций, часть которых отражена в Internet [4]. Не ставя перед собой цели детального погружения в технологию практической работы с этими пакетами, отметим только одну алгоритмическую особенность, свойственную многим из них. Эта особенность проявляется в специфике организации вычислений, направленных на решение задач кластеризации. Используемые при этом методы принято делить на *ирархические* и *неиерархические*. Среди алгоритмов первой группы выделяются *агломеративные* методы и *дивизимные* алгоритмы.

Особенность агломеративных алгоритмов проявляется в том, что на начальной стадии работы алгоритма все n наблюдаемых точек $\mathbf{Y}_j \in R^k,\ j = \overline{1,n},$ принимаются за центры n начальных кластеров. В последующем каждый из построенных таким образом кластеров объединяется с ближайшим к нему уже существующим кластером в рамках используемого понятия близости, что приводит к уменьшению числа кластеров и увеличению числа объектов, их формирующих. Этот процесс продолжают до тех пор, пока все n точек не окажутся объединенными в один кластер. Результаты такого последовательного построения кластеров удобно сопровождать графическими отображениями, что осуществляется использованием так называемых *дендрогрмм*, на которых в направлении оси абсцисс откладываются номера объединяемых кластеров, а в направлении оси ординат – результаты кластеризации. Каждый этап такого эволюционного построения кла-

стеров можно сопровождать количественной оценкой эффективности соответствующих результатов с последующим выбором наилучшего результата.

Логика дивизимных алгоритмов носит противоположный характер: на начальном этапе все наблюдаемые точки $\mathbf{Y}_j$, $j=\overline{1,n}$, рассматриваются как образующие первый начальный кластер. На последующих шагах алгоритма предыдущий кластер расщепляется и образуется кластер, содержащий меньшее количество точек.

Использование неиерархических способов кластерного анализа предполагает, что при разбиении множества $\mathbf{Y}$ на отдельные классы возможны два основополагающих подхода. При первом из них решается оптимизационная задача, связанная с использованием целевой функции, отражающей качество решения задачи классификации. Второй подход использует идею построения кластеров, "накрывающих" места наибольшей концентрации точек наблюдения.

Реализация *первого подхода*, в первую очередь, связана с так называемым алгоритмом "последовательного переноса точек из класса в класс". Существо алгоритма заключается в следующем. Проводится некоторое начальное разбиение множества $\mathbf{Y}$ на классы $\mathbf{S}^0=\left\{\mathbf{S}_1^0,\mathbf{S}_2^0,...,\mathbf{S}_k^0\right\}$. При этом руководствуются какими-либо априорными соображениями или результатами предварительных исследований. Для этого начального разбиения вычисляют значение целевой функции, выбранной в определенном виде при известном или неизвестном числе кластеров m. Затем каждое из наблюдений $\mathbf{Y}_i$ перемещается в каждый из m кластеров или рассматривается как самостоятельный кластер в случае неизвестного m и оставляется в том положении, которое соответствует наилучшему значению целевой функции (наибольшему или наименьшему в зависимости от ее содержательного наполнения). Когда подобные перемещения перестанут приводить к улучшению значений целевой функции, работа алгоритма прекращается.

Реализация второго подхода к проведению кластерного анализа может быть осуществлена, например, с помощью алгоритмов, объединенных общим названием "*Форель*" [12] и характеризующихся достаточно прозрачной логикой применяемого аппарата, что удачно их выделяет на фоне конкурирующих альтернатив.

Пусть множество $\mathbf{Y}=\left\{\mathbf{Y}_1,\mathbf{Y}_2,...,\mathbf{Y}_n\right\}$ наблюдений необходимо разбить на неизвестное число m классов. Обозначим:

$$\overline{\mathbf{Y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i ,$$

т.е. $\overline{\mathbf{Y}}$ – среднее значение всех наблюдений. Через R_0 обозначим радиус минимальной гиперсферы с центром в точке $\overline{\mathbf{Y}}$, содержащей все точки множества $\mathbf{Y}$,

$$R_0 = \max_{\mathbf{Y}_i \in \mathbf{Y}} \left(\rho\left(\mathbf{Y}_i, \overline{\mathbf{Y}} \right) \right).$$

Далее зададим произвольный радиус $R < R_0$ и из любой точки $\mathbf{Y}_i$, принимаемой за центр и обозначаемой как $\overline{\mathbf{Y}_!}$, построим гиперсферу $\mathbf{C}_1$ радиусом R. В эту гиперсферу попадает ряд точек из $\mathbf{Y}$, "центр тяжести" которых обозначим как $\overline{\mathbf{Y}_2}$:

$$\overline{Y_2} = \frac{1}{n_1}\sum_{\mathbf{Y}_i \in \mathbf{C}_1} \mathbf{Y}_i ,$$

где n_1 – количество точек, попавших в гиперсферу $\mathbf{C}_1$. Из $\overline{\mathbf{Y}_2}$ как из центра строим новую гиперсферу $\mathbf{C}_2$ радиусом R и аналогичным образом находим "центр тяжести" $\overline{\mathbf{Y}_3}$ попавших в гиперсферу $\mathbf{C}_2$ точек. Далее процедура построения гиперсфер $\mathbf{C}_j$, $j = 3, 4$ и точек $\overline{Y_j}, j = 4,5,...$ продолжается до тех пор, пока точки $\overline{\mathbf{Y}_J}$ не перестанут меняться. Попавшие в "остановившуюся" гиперсферу точки принимаются за первый кластер $\mathbf{S}_1$. Для точек, не попавших в $\mathbf{S}_1$, вновь применяется подобная процедура, следствием чего будет еще один кластер. Подобное построение кластеров продолжается до тех пор, пока все точки множества $\mathbf{Y}$ не окажутся распределенными по кластерам.

Полученная классификация может быть уточнена, если ее повторить для ряда значений радиуса R, положив, например, $R^{(\nu)} = R_0 - \nu\nabla,\ \nu = 1,2,3,...,\ \nabla = const.$ Если при этом обнаружится, что число кластеров при ряде последовательных значений $R^{(\nu)}$ радиуса гиперсферы одно и то же и резко возрастает на следующем шаге, то это является основанием для выбора числа кластеров.

Значительное число алгоритмов кластерного анализа направлено на минимизацию внутригрупповых сумм квадратов (отклонений). Они выражаются в терминах евклидовых расстояний и называются *методами минимизации дисперсий* [11]. Рассмотрим ряд характерных из этих подходов.

1 . Метод полных связей

Существо метода заключатся в объединении в один кластер тех элементов множества $\mathbf{Y}$, расстояние между которыми не превышает некоторого порога r. Величина r определяет максимально допустимый диаметр образующего кластер подмножества множества $\mathbf{Y}$.

2. Метод локального максимального расстояния

Объекты последовательно группируются по правилу: вначале каждый объект рассматривается как одноточечный кластер; два кластера объединяются в один, если максимальное расстояние между точками одного кластера и точками другого минимально по сравнению с расстояниями до всех остальных кластеров. Процедура состоит из $m-1$ шагов.

3. Метод внутриклассовых дисперсий

В качестве характеристик отдельного кластера используется сумма квадратов расстояний между каждым объектом кластера и "центром тяжести" кластера, т.е. величина

$$r_l = \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \left(\mathbf{Y}_i - \overline{\mathbf{Y}_l}\right)^T \left(\mathbf{Y}_i - \overline{\mathbf{Y}_l}\right),\ \overline{\mathbf{Y}_l} = \frac{1}{n_l} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \mathbf{Y}_i ,$$

где n_l – количество элементов, содержащихся в кластере $\mathbf{S}_l$. На каждом шаге метода объединяются вместе такие два кластера, которые приводят к минимальному увеличению суммы квадратов отклонений (дисперсий). Можно показать, что увеличение Δr_{lm} при объединении кластеров $\mathbf{S}_l$ и $\mathbf{S}_m$ в один равно:

$$\Delta r_{lm} = \frac{n_l n_m}{n_l + n_m} \left(\overline{\mathbf{Y}_l} - \overline{\mathbf{Y}_m}\right)^T \left(\overline{\mathbf{Y}_l} - \overline{\mathbf{Y}_m}\right).$$

Таким образом, метод направлен на объединение близко расположенных кластеров.

4. Центроидный метод

Расстояние между кластерами $\mathbf{S}_l$ и $\mathbf{S}_m$ определяется по центрам:

$$d_{lm}^2 \left(\mathbf{S}_l, \mathbf{S}_m\right) = \left(\overline{\mathbf{Y}_l} - \overline{\mathbf{Y}_m}\right)^T \left(\overline{\mathbf{Y}_l} - \overline{\mathbf{Y}_m}\right).$$

На каждом шаге объединяются кластеры, для которых величина d_{lm}^2 окажется минимальной.

5. Двухгрупповой метод

Пусть кластер $\mathbf{S}_l$ содержит объекты $Y_1, Y_2, ..., Y_{n_1}$, а объект $\mathbf{Y}^* \notin \mathbf{S}_l$. Обозначим, как обычно, через $\overline{\mathbf{Y}_l}$ центр кластера $\mathbf{S}_l$. Тогда средний квадрат расстояний от объекта $\mathbf{Y}^*$ до всех объектов кластера $\mathbf{S}_l$ будет равен:

$$d_{\text{ñð}}^2\left(\mathbf{S}_l, \mathbf{Y}^*\right) = \frac{1}{n_l} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \left(\mathbf{Y}_i - \mathbf{Y}^*\right)^T \left(\mathbf{Y}_i - \mathbf{Y}^*\right) =$$

$$= \frac{1}{n_l} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \left(\mathbf{Y}_i - \overline{\mathbf{Y}_l}\right)^T \left(\mathbf{Y}_i - \overline{\mathbf{Y}_l}\right) + \left(\overline{\mathbf{Y}_l} - \mathbf{Y}^*\right)^T \left(\overline{\mathbf{Y}_l} - \mathbf{Y}^*\right).$$

Первое слагаемое в этом выражении определяет внутригрупповую дисперсию, а второе представляет собой квадрат расстояния от объекта $\mathbf{Y}^* \notin \mathbf{S}_l$ до центра l -го кластера. Объект $\mathbf{Y}^*$ в процессе последовательной кластеризации объединяется с тем кластером $\mathbf{S}_l$, до которого квадрат расстояния $d_{\text{ñð}}^2$ окажется наименьшим. Если два “конкурирующих” и претендующих на присоединение элемента $\mathbf{Y}^*$ кластера имеют соизмеримые внутригрупповые дисперсии, то элемент $\mathbf{Y}^*$ присоединяют к тому кластеру, расстояние до центра которого окажется меньшим. Для кластеров с различными дисперсиями, но одинаково удаленными от $\mathbf{Y}^*$, предпочтение отдается кластеру с меньшей внутригрупповой дисперсией .

6. Метод групповых средних

Средний квадрат расстояний между двумя кластерами $\mathbf{S}_l$ и $\mathbf{S}_m$, как и выше, определяется выражением:

$$d_{\text{ñð}}^2\left(\mathbf{S}_l, \mathbf{S}_m\right) = \frac{1}{n_l n_m} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \sum_{\mathbf{Y}_j \in \mathbf{S}_m} \left(\mathbf{Y}_i - \mathbf{Y}_j\right)^T \left(\mathbf{Y}_i - \mathbf{Y}_j\right) =$$

$$= \frac{1}{n_l} \sum_{\mathbf{Y}_i \in \mathbf{S}_l} \left(\mathbf{Y}_i - \overline{\mathbf{Y}_\mathbf{l}}\right)^T \left(\mathbf{Y}_i - \overline{\mathbf{Y}_l}\right) + \frac{1}{n_m} \sum_{\mathbf{Y}_j \in \mathbf{S}_m} \left(\mathbf{Y}_j - \overline{\mathbf{Y}_m}\right)^T \left(\mathbf{Y}_j - \overline{\mathbf{Y}_m}\right) +$$

$$+ \left(\overline{\mathbf{Y}}_l - \overline{\mathbf{Y}_m}\right)^T \left(\overline{\mathbf{Y}_l} - \overline{\mathbf{Y}_m}\right).$$

Отсюда следует, что средний квадрат расстояний между двумя класте-

рами равняется сумме внутригрупповых дисперсий и квадрата расстояний между центрами кластеров. При последовательной кластеризации объединяются кластеры с минимальной величиной $d^2_{ñð}$, т.е. в первую очередь объединяются кластеры с наибольшей плотностью точек и наименее удаленные друг от друга по центрам.

7. Метод случайных начальных точек

На множестве $\mathbf{Y}$ случайным образом выбирается некоторая точка $\mathbf{Y}_i$, которая объявляется центром кластера. Все точки, удаленные от $\mathbf{Y}_i$ на расстояние, не превышающее некоторого порога r, объединяются в кластер $\mathbf{S}_1$. Среди оставшихся точек из подмножества $\mathbf{Y} \setminus \mathbf{S}_1$ снова случайным образом выбирается некоторая точка, объявляемая центром второго кластера, и весь процесс повторяется до формирования второго кластера. Затем аналогичным образом создается третий кластер и т.д. до полного разбиения множества $\mathbf{Y}$ на отдельные группы. Подход допускает многочисленные модификации [1,2].

Изложенные методы обладают определенной общностью и часто приводят к близким результатам. Выбор конкретного из них во многом определяется частными особенностями решаемой задачи. При этом самостоятельные наработки полезно совмещать с возможностями стандартных пакетов прикладных программ.

4.7. Примеры компонентного и кластерного анализа

В заключение данного раздела рассмотрим два примера. Первый из них посвящен компонентному анализу и основан на результатах второй главы первой части настоящего пособия [13]. Численные значения параметров, используемые в примере, заимствованы в [11]. Результаты проведенного компонентного анализа используются во втором примере, в котором изучается один из вариантов выполнения кластерного анализа, использующий технику метода “Форель” из [12]. Кластерный анализ осуществляется по двум из пяти главных компонентов, найденных при рассмотрении первого примера.

Итак, пусть некая экономическая система характеризуется пятимерным вектором $\tilde{\mathbf{O}} \in R^5$ экспериментальных наблюдений, проведенных на пятидесяти объектах. Численные значения наблюдений, представленных в натуральной и в стандартизованной формах, таковы.

Пример 1 - компонентный анализ

$N := 50$

$X6_1 := (0.4 \;\; 0.26 \;\; 0.4 \;\; 0.5 \;\; 0.4 \;\; 0.19 \;\; 0.250 \;\; 0.44 \;\; 0.17 \;\; 0.39)^T$

$X6_2 := (0.33 \;\; 0.25 \;\; 0.32 \;\; 0.02 \;\; 0.06 \;\; 0.15 \;\; 0.08 \;\; 0.2 \;\; 0.2 \;\; 0.3)^T$

$X6_3 := (0.24 \;\; 0.1 \;\; 0.11 \;\; 0.47 \;\; 0.53 \;\; 0.34 \;\; 0.2 \;\; 0.24 \;\; 0.54 \;\; 0.4)^T$

$X6_4 := (0.2 \;\; 0.54 \;\; 0.42 \;\; 0.27 \;\; 0.37 \;\; 0.38 \;\; 0.35 \;\; 0.42 \;\; 0.32 \;\; 0.33)^T$

$X6_5 := (0.29 \;\; 0.3 \;\; 0.56 \;\; 0.42 \;\; 0.26 \;\; 0.16 \;\; 0.45 \;\; 0.31 \;\; 0.08 \;\; 0.68)^T$

$X8_1 := (1.23 \;\; 1.04 \;\; 1.8 \;\; 0.43 \;\; 0.88 \;\; 0.57 \;\; 1.72 \;\; 1.7 \;\; 0.84 \;\; 0.6)^T$

$X8_2 := (0.82 \;\; 0.84 \;\; 0.67 \;\; 1.04 \;\; 0.66 \;\; 0.86 \;\; 0.79 \;\; 0.34 \;\; 1.6 \;\; 1.46)^T$

$X8_3 := (1.27 \;\; 1.58 \;\; 0.68 \;\; 0.86 \;\; 1.98 \;\; 0.33 \;\; 0.45 \;\; 0.74 \;\; 0.03 \;\; 0.99)^T$

$X8_4 := (0.24 \;\; 0.57 \;\; 1.22 \;\; 0.68 \;\; 1 \;\; 0.81 \;\; 1.27 \;\; 1.14 \;\; 1.89 \;\; 0.67)'$

$X8_5 := (0.96 \;\; 0.67 \;\; 0.98 \;\; 1.16 \;\; 0.54 \;\; 1.23 \;\; 0.78 \;\; 1.16 \;\; 4.44 \;\; 1.06)'$

$X11_1 := (26006 \;\; 23935 \;\; 22589 \;\; 21220 \;\; 7394 \;\; 11586 \;\; 26609 \;\; 7801 \;\; 11587 \;\; 9475)^T$

$X11_2 := (10811 \;\; 6371 \;\; 26761 \;\; 4210 \;\; 3557 \;\; 14148 \;\; 9872 \;\; 5975 \;\; 16662 \;\; 9166)^T$

$X11_3 := (15118 \;\; 11429 \;\; 6462 \;\; 24628 \;\; 49727 \;\; 11470 \;\; 19448 \;\; 18963 \;\; 9185 \;\; 1$

$X11_4 := (6265 \;\; 8810 \;\; 17659 \;\; 10342 \;\; 8901 \;\; 8402 \;\; 32625 \;\; 31160 \;\; 46461$

$X11_5 := (6391 \;\; 11115 \;\; 6555 \;\; 11085 \;\; 9484 \;\; 3967 \;\; 15283 \;\; 20874 \;\; 19418 \;\; 3351)^T$

$X12_1 := (167.69 \;\; 186.1 \;\; 220.45 \;\; 169.3 \;\; 39.53 \;\; 40.41 \;\; 102.96 \;\; 37.02 \;\; 45.74 \;\; 40.07)^T$

$X12_2 := (45.44 \;\; 41.08 \;\; 135.14 \;\; 42.39 \;\; 37.39 \;\; 101.78 \;\; 47.55 \;\; 32.61 \;\; 103.25 \;\; 38.95$

$X12_3 := (81.32 \;\; 67.26 \;\; 59.92 \;\; 107.34 \;\; 512.6 \;\; 53.81 \;\; 80.83 \;\; 59.42 \;\; 36.96 \;\; 91.43)^T$

$X12_4 := (17.16 \;\; 27.29 \;\; 184.33 \;\; 58.42 \;\; 59.4 \;\; 49.63 \;\; 391.27 \;\; 258.62 \;\; 75.66 \;\; 123.68)^T$

$X12_5 := (37.21 \;\; 53.37 \;\; 32.87 \;\; 45.63 \;\; 48.41 \;\; 13.58 \;\; 63.99 \;\; 104.55 \;\; 222.11 \;\; 25.76$

$X17_1 := (17.72 \;\; 18.39 \;\; 26.46 \;\; 22.37 \;\; 28.13 \;\; 17.55 \;\; 21.92 \;\; 19.52 \;\; 23.99 \;\; 21.76)^T$

$$\mathbf{X17_2} := (25.68 \quad 18.13 \quad 25.74 \quad 21.21 \quad 22.97 \quad 16.38 \quad 13.21 \quad 14.48 \quad 13.38 \quad 13.69)^{\mathbf{T}}$$

$$\mathbf{X17_3} := (16.66 \quad 15.06 \quad 20.09 \quad 15.98 \quad 18.27 \quad 14.42 \quad 22.76 \quad 15.41 \quad 19.35 \quad 16.83)^{\mathbf{T}}$$

$$\mathbf{X17_4} := (30.53 \quad 17.98 \quad 22.09 \quad 18.29 \quad 26.05 \quad 26.2 \quad 17.26 \quad 18.83 \quad 19.7 \quad 16.87)^{\mathbf{T}}$$

$$\mathbf{X17_5} := (14.63 \quad 22.17 \quad 22.62 \quad 26.44 \quad 22.26 \quad 19.13 \quad 18.28 \quad 28.23 \quad 12.39 \quad 11.64)^{\mathbf{T}}$$

Построим объединенные векторы:

$$\mathbf{X6} := \mathbf{stack}(\mathbf{X6_1}, \mathbf{X6_2}, \mathbf{X6_3}, \mathbf{X6_4}, \mathbf{X6_5})$$

$$\mathbf{X8} := \mathbf{stack}(\mathbf{X8_1}, \mathbf{X8_2}, \mathbf{X8_3}, \mathbf{X8_4}, \mathbf{X8_5})$$

$$\mathbf{X11} := \mathbf{stack}(\mathbf{X11_1}, \mathbf{X11_2}, \mathbf{X11_3}, \mathbf{X11_4}, \mathbf{X11_5})$$

$$\mathbf{X12} := \mathbf{stack}(\mathbf{X12_1}, \mathbf{X12_2}, \mathbf{X12_3}, \mathbf{X12_4}, \mathbf{X12_5})$$

$$\mathbf{X17} := \mathbf{stack}(\mathbf{X17_1}, \mathbf{X17_2}, \mathbf{X17_3}, \mathbf{X17_4}, \mathbf{X17_5})$$

Используя встроенные функции, найдем характеристики:

$$\mathbf{mX6} := \mathbf{mean}(\mathbf{X6}) \qquad \mathbf{DX6} := \mathbf{Var}(\mathbf{X6})$$

$$\mathbf{mX8} := \mathbf{mean}(\mathbf{X8})$$

$$\mathbf{mX8} := \mathbf{mean}(\mathbf{X8}) \qquad \mathbf{DX8} := \mathbf{Var}(\mathbf{X8})$$

$$\mathbf{mX11} := \mathbf{mean}(\mathbf{X11}) \qquad \mathbf{DX11} := \mathbf{Var}(\mathbf{X11})$$

$$\mathbf{mX12} := \mathbf{mean}(\mathbf{X12}) \qquad \mathbf{DX12} := \mathbf{Var}(\mathbf{X12})$$

$$\mathbf{mX17} := \mathbf{mean}(\mathbf{X17}) \qquad \mathbf{DX17} := \mathbf{Var}(\mathbf{X17})$$

$$\mathbf{ORIGIN} := 1$$

$$\mathbf{Z}^{\langle 2 \rangle} := \frac{\mathbf{X6} - \mathbf{mX6}}{\sqrt{\mathbf{DX6}}} \qquad \mathbf{Z}^{\langle 3 \rangle} := \frac{\mathbf{X8} - \mathbf{mX8}}{\sqrt{\mathbf{DX8}}} \qquad \mathbf{Z}^{\langle 4 \rangle} := \frac{\mathbf{X11} - \mathbf{mX11}}{\sqrt{\mathbf{DX11}}}$$

$$\mathbf{Z}^{\langle 5 \rangle} := \frac{\mathbf{X12} - \mathbf{mX12}}{\sqrt{\mathbf{DX12}}} \qquad \mathbf{Z}^{\langle 6 \rangle} := \frac{\mathbf{X17} - \mathbf{mX17}}{\sqrt{\mathbf{DX17}}}$$

Пусть x1, x2, x3, x4, x5 -стандартизованные признаки X6, X8, X11, X12, X17.
Тогда целью компонентного анализа является построение модели,
которая на N объектах приводит к соотношениям:

$$x_{ij} = a_{i1} \cdot f_j + a_{2j} \cdot f_{2j} + a_{3j} \cdot f_{3j} + a_{4j} \cdot f_{4j} + a_{5j} \cdot f_{5j}$$

$$i = 1, 2, \ldots, 5.$$

$$x_i = a_{i1} \cdot f_1 + a_{i2} \cdot f_2 + a_{i3} \cdot f_3 + a_{i4} \cdot f_4 + a_{i5} \cdot f_5$$

$$i = 1..5 \qquad j = 1..N$$

или в матричной форме:

$$X = A \cdot F,$$

где X имеет размерность nxN (n=5), A имеет размерность nxn, F имеет размерность nxN. Определению подлежат матрицы A и F.

Определим **корреляционную матрицу признаков**. Для этого используем имеющиеся данные и для последующего удобства проведем переобозначения:

$$X^{\langle 1 \rangle} := Z^{\langle 2 \rangle} \qquad X^{\langle 2 \rangle} := Z^{\langle 3 \rangle} \qquad X^{\langle 3 \rangle} := Z^{\langle 4 \rangle} \qquad X^{\langle 4 \rangle} := Z^{\langle 5 \rangle} \qquad X^{\langle 5 \rangle} := Z^{\langle 6 \rangle}$$

$\sum X^{\langle 1 \rangle} = 1.599 \times 10^{-14}$ $\sum X^{\langle 2 \rangle} = 1.125 \times 10^{-14}$ и т.п., т.е. столбцы у X центрированы

Найдем корреляционную матрицуа признаков, которая находится так:

$$R := \frac{1}{N - 1} \cdot X^T \cdot X$$

$$R = \begin{pmatrix} 1 & -0.128 & 0.212 & 0.2 & 0.086 \\ -0.128 & 1 & 0.379 & 0.431 & -0.241 \\ 0.212 & 0.379 & 1 & 0.784 & -0.04 \\ 0.2 & 0.431 & 0.784 & 1 & -0.09 \\ 0.086 & -0.241 & -0.04 & -0.09 & 1 \end{pmatrix}$$

Найдем вектор собственных чисел матрицы R:

$$\Lambda := \mathbf{eigenvals}(R)$$

$$\sum \Lambda = 5$$

$$\Lambda = \begin{pmatrix} 2.149 \\ 1.249 \\ 0.858 \\ 0.531 \\ 0.213 \end{pmatrix}$$

Сумма компонентов этого вектора равна n=5, т.е. суммарной дисперсии стандартизованных признаков. Упорядочим компоненты вектора в направлении их убывания:

$$\Lambda 1 := \mathbf{reverse}(\mathbf{sort}(\Lambda))$$

$$\Lambda 1^T = (2.149 \quad 1.249 \quad 0.858 \quad 0.531 \quad 0.213)$$

$\Lambda 1_1 + \Lambda 1_2 + \Lambda 1_3 = 4.256$

$\frac{4.256}{5} = 0.851$, т.е. три первые компоненты объясняют 85,1% суммарной дисперсии исходных признаков.

Находим **матрицу собственных векторов** матрицы R:

$\mathbf{U} := \mathbf{eigenvecs}(\mathbf{R})$

$$\mathbf{U} = \begin{pmatrix} -0.159 & 0.66 & -0.593 & 0.432 & -0.011 \\ -0.447 & -0.434 & 0.189 & 0.757 & -0.057 \\ -0.606 & 0.173 & 0.12 & -0.34 & -0.688 \\ -0.621 & 0.116 & 0.086 & -0.267 & 0.723 \\ 0.152 & 0.577 & 0.768 & 0.23 & 0.031 \end{pmatrix}$$

Столбцы этой матрицы упорядочены в соответствии с компонентами вектора . Их же следует упорядочить в соответствии с компонентами вектора 1, для чего следует переставить третий и четвертый столбцы местами.

$\mathbf{U1}^{\langle 1 \rangle} := \mathbf{U}^{\langle 1 \rangle}$ $\mathbf{U1}^{\langle 2 \rangle} := \mathbf{U}^{\langle 2 \rangle}$ $\mathbf{U1}^{\langle 3 \rangle} := \mathbf{U}^{\langle 4 \rangle}$ $\mathbf{U1}^{\langle 4 \rangle} := \mathbf{U}^{\langle 3 \rangle}$ $\mathbf{U1}^{\langle 5 \rangle} := \mathbf{U}^{\langle 5 \rangle}$

$$\mathbf{U1} = \begin{pmatrix} -0.159 & 0.66 & 0.432 & -0.593 & -0.011 \\ -0.447 & -0.434 & 0.757 & 0.189 & -0.057 \\ -0.606 & 0.173 & -0.34 & 0.12 & -0.688 \\ -0.621 & 0.116 & -0.267 & 0.086 & 0.723 \\ 0.152 & 0.577 & 0.23 & 0.768 & 0.031 \end{pmatrix}$$

Теперь находим **матрицу нагрузок** A:

$\underset{\wedge\wedge\wedge\wedge}{\mathbf{V}} := \mathbf{diag}\left(\sqrt{\Lambda \mathbf{1}}\right)$

$$\mathbf{V} = \begin{pmatrix} 1.466 & 0 & 0 & 0 & 0 \\ 0 & 1.118 & 0 & 0 & 0 \\ 0 & 0 & 0.926 & 0 & 0 \\ 0 & 0 & 0 & 0.729 & 0 \\ 0 & 0 & 0 & 0 & 0.461 \end{pmatrix}$$

$\underset{\wedge\wedge\wedge}{\mathbf{A}} := \mathbf{U1} \cdot \mathbf{V}$

$$\mathbf{A} = \begin{pmatrix} -0.232 & 0.738 & 0.4 & -0.432 & -5.253 \times 10^{-3} \\ -0.655 & -0.485 & 0.701 & 0.138 & -0.026 \\ -0.888 & 0.193 & -0.315 & 0.087 & -0.317 \\ -0.91 & 0.129 & -0.248 & 0.063 & 0.334 \\ 0.222 & 0.645 & 0.213 & 0.56 & 0.015 \end{pmatrix}$$

$\mathbf{A}^{\langle 1 \rangle} \cdot \mathbf{A}^{\langle 2 \rangle} = 0$ $\mathbf{A}^{\langle 2 \rangle} \cdot \mathbf{A}^{\langle 5 \rangle} = 0$ $\left|\left(\mathbf{A}^{\mathbf{T}}\right)^{\langle 3 \rangle}\right| = 1.016$ $\left|\left(\mathbf{A}^{\mathbf{T}}\right)^{\langle 1 \rangle}\right| = 0.973$ $|\mathbf{A}| = 0.51$

$\left(\mathbf{A}^{\mathbf{T}}\right)^{\langle 2\rangle} \cdot \left(\mathbf{A}^{\mathbf{T}}\right)^{\langle 4\rangle} = 0.36$ $\left|\mathbf{A}^{\langle 3\rangle}\right| = 0.926$

Таким образом, у матрицы нагрузок столбцы ортоональны, строки - нормированы.

Наибольшая корреляция (0.91) обнаруживается между первым компонентом и признаком X12-среднегодовой стоимостью ОПФ; вторая по величине корреляция (0.888) связывает первый компонент с X11 со среднегодовой численностью ППП. Тогда первый компонент можно интерпретировать как показатель размера предприятия.

Суммарные нагрузки компонентов:

$\left|\mathbf{A}^{\langle 1\rangle}\right| = 1.466$ $\left|\mathbf{A}^{\langle 2\rangle}\right| = 1.118$ $\left|\mathbf{A}^{\langle 3\rangle}\right| = 0.926$ $\left|\mathbf{A}^{\langle 4\rangle}\right| = 0.729$ $\left|\mathbf{A}^{\langle 5\rangle}\right| = 0.461$

$\sum_{\mathbf{i}=1}^{5} \left|\left(\mathbf{A}^{\langle \mathbf{i}\rangle}\right)\right| = 4.7$ $\sum_{\mathbf{i}=1}^{5}\sum_{\mathbf{j}=1}^{5} \left(\mathbf{A}_{\mathbf{i},\mathbf{j}}\right)^2 = 5$

$\sum_{\mathbf{j}=1}^{5} \left(\mathbf{A}_{1,\mathbf{j}}\right)^2 = 0.946$ $\sum_{\mathbf{j}=1}^{5} \left(\mathbf{A}_{2,\mathbf{j}}\right)^2 = 1.176$ и т.д. $\sum_{\mathbf{j}=1}^{5} \left(\mathbf{A}_{5,\mathbf{j}}\right)^2 = 0.824$

Вклад первого компонента в суммарную дисперсию всех исходных признаков:

$$\sum_{\mathbf{i}=1}^{5} \left(\mathbf{A}_{\mathbf{i},1}\right)^2 = 2.149$$

Вклад второго компонента в суммарную дисперсию всех исходных признаков:

$$\sum_{\mathbf{i}=1}^{5} \left(\mathbf{A}_{\mathbf{i},2}\right)^2 = 1.249$$

Вклад третьего компонента в суммарную дисперсию всех исходных признаков:

$$\sum_{\mathbf{i}=1}^{5} \left(\mathbf{A}_{\mathbf{i},3}\right)^2 = 0.858$$

Вклад четвертого компонента в суммарную дисперсию всех исходных признаков:

$$\sum_{\mathbf{i}=1}^{5} \left(\mathbf{A}_{\mathbf{i},4}\right)^2 = 0.531$$

Вклад пятого компонента в суммарную дисперсию всех исходных признаков:

$$\sum_{i=1}^{5} \left(A_{i,5}\right)^2 = 0.213$$

Общий вклад всех компонентов: $2.149 + 1.249 + 0.858 + 0.531 + 0.213 = 5$

Сопоставим с собственными числами матрицы R:

$$\Lambda 1^{T} = \begin{pmatrix} 2.149 & 1.249 & 0.858 & 0.531 & 0.213 \end{pmatrix}$$

.

Матрица глвных компонентов принимает значение:

$$\underset{\wedge\wedge\wedge}{F} := A^{-1} \cdot X^{T}$$

$F =$

	1	2	3	4	5	6
1	-0.985	-0.775	-1.148	-0.395	0.749	0.634
2	0.255	-0.139	0.91	1.586	1.208	-0.592
3	-0.198	-0.828	0.948	-0.447	1.002	-0.788
4	-0.619	0.236	1.628	-0.501	1.176	-0.133
5	-0.488	0.193	0.916	0.448	0.354	…

Главные компоненты должны быть центрированы, нормированы в смысле дисперсии и взаимно не коррелированы. Убедимся в этом:

$$\mathbf{mean}\left[\left(F^{T}\right)^{\langle 1\rangle}\right] = 0 \qquad \mathbf{mean}\left[\left(F^{T}\right)^{\langle 2\rangle}\right] = 0 \qquad \mathbf{mean}\left[\left(F^{T}\right)^{\langle 3\rangle}\right] = 0$$

$$\mathbf{mean}\left[\left(F^{T}\right)^{\langle 4\rangle}\right] = 0 \qquad \mathbf{mean}\left[\left(F^{T}\right)^{\langle 5\rangle}\right] = 0$$

$$\mathbf{Var}\left[\left(F^{T}\right)^{\langle 1\rangle}\right] = 1 \qquad \mathbf{Var}\left[\left(F^{T}\right)^{\langle 2\rangle}\right] = 1 \qquad \mathbf{Var}\left[\left(F^{T}\right)^{\langle 3\rangle}\right] = 0.619$$

$$\mathbf{Var}\left[\left(F^{T}\right)^{\langle 4\rangle}\right] = 1.617 \qquad \mathbf{Var}\left[\left(F^{T}\right)^{\langle 5\rangle}\right] = 1$$

$$\mathbf{corr}\left[\left(F^{T}\right)^{\langle 1\rangle}, \left(F^{T}\right)^{\langle 2\rangle}\right] = 0 \qquad \mathbf{corr}\left[\left(F^{T}\right)^{\langle 1\rangle}, \left(F^{T}\right)^{\langle 3\rangle}\right] = 0$$

и так далее. $\mathbf{corr}\left[\left(F^{T}\right)^{\langle 3\rangle}, \left(F^{T}\right)^{\langle 5\rangle}\right] = 0$

Дополнительно убеждаемся:

$$\frac{1}{N-1}\cdot F\cdot F^{T}=\begin{pmatrix}1 & 0 & 0 & 0 & 1.388\times 10^{-15}\\ 0 & 1 & 0 & 0 & 0\\ 0 & 0 & 0.619 & 0 & 0\\ 0 & 0 & 0 & 1.617 & 0\\ 1.332\times 10^{-15} & 0 & 0 & 0 & 1\end{pmatrix}$$

$$A\cdot A^{T}=\begin{pmatrix}0.946 & 0.016 & 0.187 & 0.179 & 0.267\\ 0.016 & 1.176 & 0.287 & 0.36 & -0.232\\ 0.187 & 0.287 & 1.033 & 0.81 & -0.096\\ 0.179 & 0.36 & 0.81 & 1.021 & -0.132\\ 0.267 & -0.232 & -0.096 & -0.132 & 0.824\end{pmatrix}$$

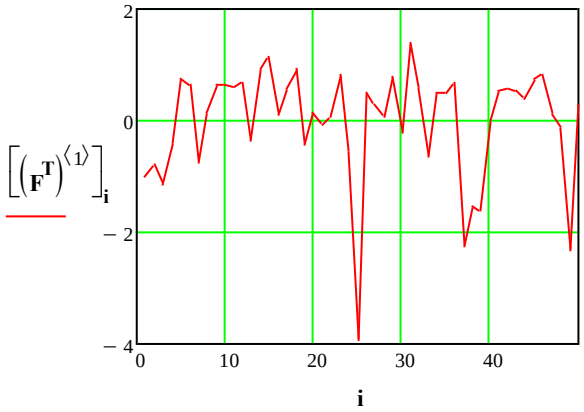
$$R=\begin{pmatrix}1 & -0.128 & 0.212 & 0.2 & 0.086\\ -0.128 & 1 & 0.379 & 0.431 & -0.241\\ 0.212 & 0.379 & 1 & 0.784 & -0.04\\ 0.2 & 0.431 & 0.784 & 1 & -0.09\\ 0.086 & -0.241 & -0.04 & -0.09 & 1\end{pmatrix}$$

т.е $A(A)^{T}:=R$

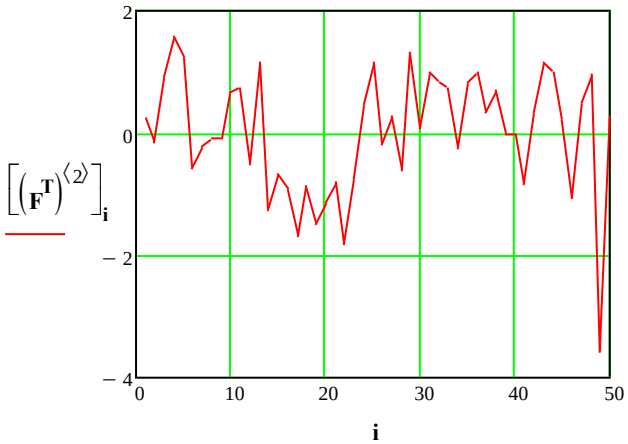
Значения компонентов на объектах:

$i:=1..50$

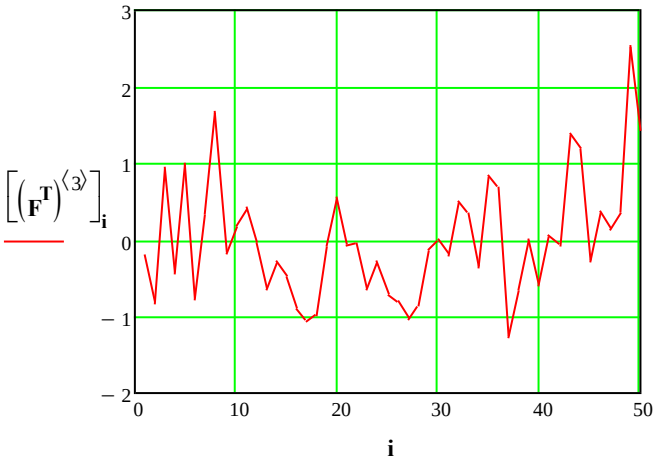
Первый компонент



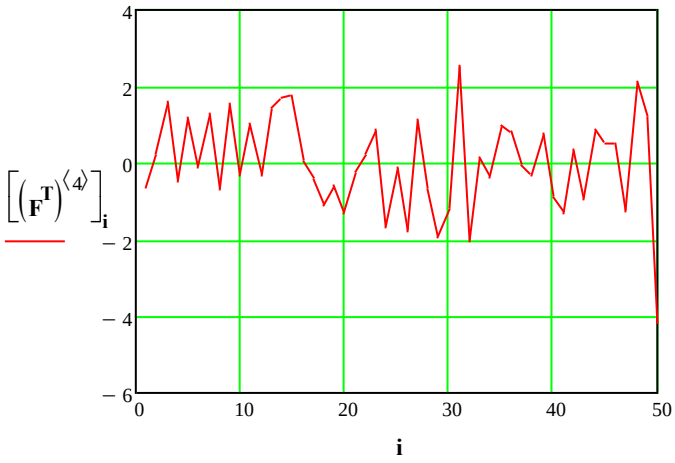
Второй компонент



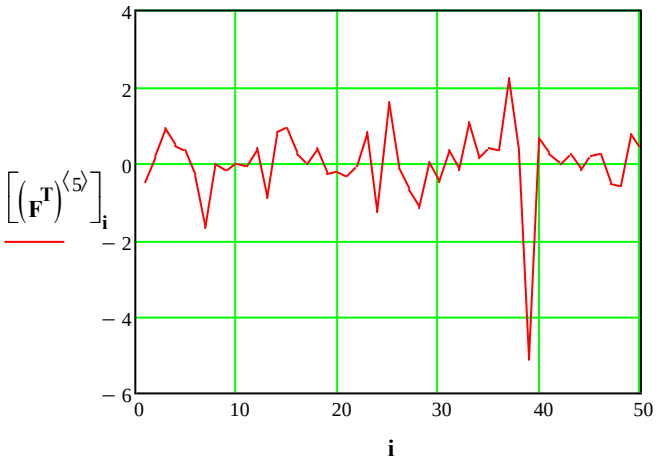
Третий компонент



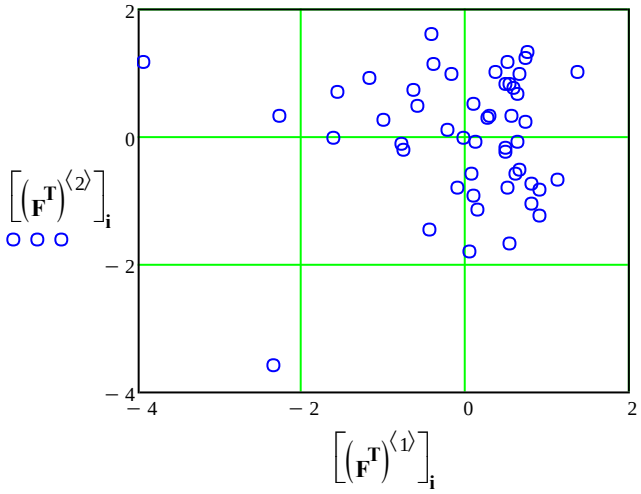
Четвертый компонент



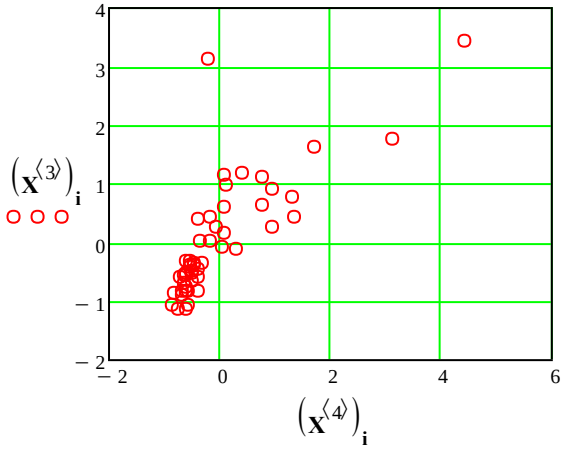
Пятый компонент



Проведем классификацию объектов по двум компонентам, воспользовавшись графическим отображением результатов компонентного анализа.



Для сравнения отобразим эти же объекты в плоскости каких-либо двух других признаков и проведем их кластеризацию.



Пример 2 - кластерный анализ

Построим матрицу расстояний между объектами и найдем объекты, расстояние между которыми минимально.

$$Q := Z^T \qquad N := 50 \qquad i := 1..N \qquad j := 1..N \qquad \mathbf{rows}(Q) = 6$$

$$\rho_{i,j} := \left|Q^{\langle i\rangle} - Q^{\langle j\rangle}\right| \qquad \mathbf{rows}(\rho) = 50 \qquad E1 := \mathbf{identity}(\mathbf{rows}(\rho)) \qquad \mathbf{min}(\rho) = 0$$

Так как $\min\rho(i,i) = 0$, для поиска минимально удаленных объектов удобно изменить диагональные элементы матрицы расстояний, положив $\rho := \rho + \max(\rho)\cdot E1$;

при этом $\min(\rho) = 0.32$ $c := \min(\rho)$.

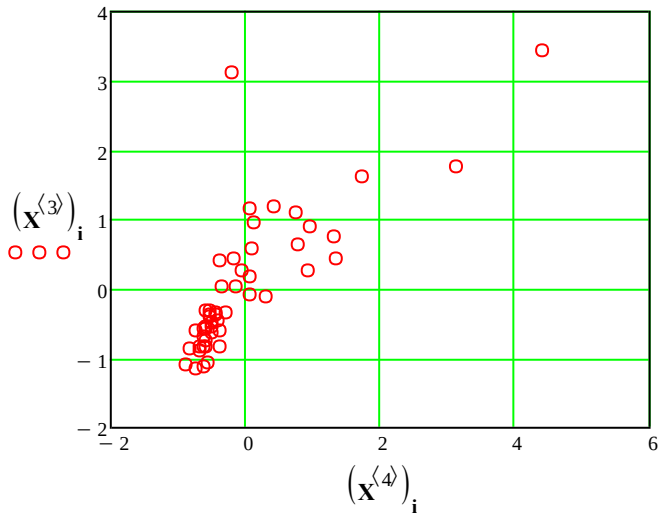
Найдем номера наименее удаленных объектов:

$g :=$ for $m \in 1..\text{rows}(\rho) - 1$
  for $n \in m + 1..\text{cols}(\rho)$
    if $\rho_{m,n} - c = 0$
      $a \leftarrow m$
      $b \leftarrow n$
  $\begin{pmatrix} a \\ b \end{pmatrix}$

$g = \begin{pmatrix} 35 \\ 36 \end{pmatrix}$ $\rho_{11,36} = 0.437$

Кластерный анализ проводится в пространстве R2, т.е. по двум или каким-либо исходным признакам, или первым двум главным компонентам. Чтобы конкретно определиться с подлежащим кластеризации материалом, следует

графически отобразить объекты в плоскости всех возможных пар исходных признаков и первых двух главных компонентов и выбрать для последующей кластеризации ситуацию с наиболее рельефно разделенными объектами. Например:



Кластерный анализ объектов проводится упрощенным методом "Форэль" [1]. В качестве признаков классификации в данном иллюстративном примере выбраны исходные признаки:

- среднегодовая численность ППП $X^{\langle 3 \rangle}$ и
- среднегодовая стоимость ОПФ $X^{\langle 4 \rangle}$. $X^{\langle 4 \rangle}$ $X^{\langle 3 \rangle}$

Существо метода изложено выше.

$\mathbf{Y1} := \mathbf{X}^{\langle 3 \rangle}$ $\mathbf{Y2} := \mathbf{X}^{\langle 4 \rangle}$

Введем координаты каждого объекта в терминах используемых признаков.

$\mathbf{y_i} := \begin{pmatrix} \mathbf{Y1_i} \\ \mathbf{Y2_i} \end{pmatrix}$ Например, $\mathbf{y}_1 = \begin{pmatrix} 1.087 \\ 0.777 \end{pmatrix}$ $\mathbf{y}_{50} = \begin{pmatrix} -1.157 \\ -0.727 \end{pmatrix}$

и найдем среднее значение координат ("центр" объектов):

$$\mathbf{m0} := \frac{1}{\mathbf{N}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{N}} \mathbf{y_i} \qquad \mathbf{m0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Составим вектор расстояний объектов от центра m0:

$$\rho\mathbf{1_i} := \left| \mathbf{y_i} - \mathbf{m0} \right|$$

и найдем максимальное расстояние: $\mathbf{a} := \mathbf{max}(\rho\mathbf{1})$

$\mathbf{a} = 5.607$

Вычислим номер наиболее удаленного объекта:

$$\mathbf{g} := \left| \begin{array}{l} \textbf{for } \mathbf{i} \in 1..\mathbf{N} \\ \quad \mathbf{g} \leftarrow \mathbf{i} \textbf{ if } \rho\mathbf{1_i} - \mathbf{a} = 0 \\ \mathbf{g} \end{array} \right. \qquad \mathbf{g} = 25$$

Построим окружность радиуса r=a/4 с центром в точке m0 и выделим объекты, попадающие в эту окружность. Одновременно найдем "центр" m1 этих точек.

$\mathbf{r} := \frac{\mathbf{a}}{4}$

$$\mathbf{g1} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \textbf{for } \mathbf{i} \in 1..\mathbf{N} \\ \quad \textbf{if } \left| \mathbf{y_i} - \mathbf{m0} \right| < \mathbf{r} \\ \quad\quad \left| \begin{array}{l} \mathbf{g1_m} \leftarrow \mathbf{y_i} \\ \mathbf{m1} \leftarrow \frac{1}{\mathbf{m}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{g1_i} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \begin{pmatrix} \mathbf{m} - 1 \\ \mathbf{m1} \end{pmatrix} \end{array} \right. \qquad \mathbf{g1} = \begin{pmatrix} 44 \\ \{2,1\} \end{pmatrix} \quad \mathbf{g1}_2 = \begin{pmatrix} -0.252 \\ -0.268 \end{pmatrix} \qquad \mathbf{m1} := \mathbf{g1}_2$$

Вычислим расстояние между центрами : $|\mathbf{m0} - \mathbf{m1}| = 0.368$.

Это достаточно большая величина, соизмеримая с расстояниями между объектами. Поэтому переместим центр окружности в точку m1 и повторим операцию относительно этой точки.

.

$$\mathbf{g2} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \textbf{for}\ \ \mathbf{i} \in 1..\mathbf{N} \\ \quad \textbf{if}\ \ \left|\mathbf{y_i} - \mathbf{m1}\right| < \mathbf{r} \\ \quad\quad \left| \begin{array}{l} \mathbf{g2_m} \leftarrow \mathbf{y_i} \\ \mathbf{m2} \leftarrow \frac{1}{\mathbf{m}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{g2_i} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \begin{pmatrix} \mathbf{m} - 1 \\ \mathbf{m2} \end{pmatrix} \end{array} \right. \qquad \mathbf{g2} = \begin{pmatrix} 40 \\ \{2,1\} \end{pmatrix} \qquad \mathbf{g2}_2 = \begin{pmatrix} -0.384 \\ -0.352 \end{pmatrix}$$

$$\mathbf{m2} := \mathbf{g2}_2$$

Снова вычислим расстояние между "центрами": $|\mathbf{m1} - \mathbf{m2}| = 0.156$

"Центр" m2 существенно сместился относительно m1, поэтому повторим операцию относительно m2

$$\mathbf{g3} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \textbf{for}\ \ \mathbf{i} \in 1..\mathbf{N} \\ \quad \textbf{if}\ \ \left|\mathbf{y_i} - \mathbf{m2}\right| < \mathbf{r} \\ \quad\quad \left| \begin{array}{l} \mathbf{g3_m} \leftarrow \mathbf{y_i} \\ \mathbf{m3} \leftarrow \frac{1}{\mathbf{m}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{g3_i} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \begin{pmatrix} \mathbf{m} - 1 \\ \mathbf{m3} \end{pmatrix} \end{array} \right. \qquad \mathbf{g3} = \begin{pmatrix} 37 \\ \{2,1\} \end{pmatrix} \qquad \mathbf{g3}_2 = \begin{pmatrix} -0.464 \\ -0.431 \end{pmatrix}$$

$$\mathbf{m3} := \mathbf{g3}_2$$

Опять вычислим смещение: $|\mathbf{m2} - \mathbf{m3}| = 0.113$

Эта величина по-прежнему достаточно велика, поэтому операция повторяется, но относительно m3 :

$$\mathbf{g4} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \text{for } \mathbf{i} \in 1..\mathbf{N} \\ \quad \text{if } \left|\mathbf{y_i} - \mathbf{m3}\right| < \mathbf{r} \\ \quad\quad \left| \begin{array}{l} \mathbf{g4_m} \leftarrow \mathbf{y_i} \\ \mathbf{m4} \leftarrow \frac{1}{\mathbf{m}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{g4_i} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \begin{pmatrix} \mathbf{m} - 1 \\ \mathbf{m4} \end{pmatrix} \end{array} \right.$$

$$\mathbf{g4} = \begin{pmatrix} 37 \\ \{2,1\} \end{pmatrix} \qquad \mathbf{g4}_2 = \begin{pmatrix} -0.464 \\ -0.431 \end{pmatrix}$$

$$\mathbf{m4} := \mathbf{g4}_2$$

Находим очередное смещение "центра": $\left|\mathbf{m3} - \mathbf{m4}\right| = 0$

Так как точки m3 и m4 совпадают, т.е. положение "центра" стабилизировалось, то точку m3 принимаем за центр первого кластера и к нему относим все 37 объектов, координаты которых попали в окружность радиуса r с центром в точке m3. Найдем номера этих объектов:

$$\mathbf{N1} := \left| \begin{array}{l} \text{for } \mathbf{i} \in 1..\mathbf{N} \\ \quad \mathbf{n_i} \leftarrow \mathbf{i} \quad \text{if } \left|\mathbf{y_i} - \mathbf{m3}\right| < \mathbf{r} \\ \mathbf{n_i} \leftarrow 0 \quad \text{if } \left|\mathbf{y_i} - \mathbf{m3}\right| \geq \mathbf{r} \\ \mathbf{n} \end{array} \right.$$

$\mathbf{N1} =$

	1
1	0
2	0
3	0
4	0
5	5
6	6
7	0
8	8
9	9
10	10
11	11
12	12
13	0
14	14
15	15
16	...

Это объекты (точки) с номерами:
5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50.

Построим новые векторы признаков, в которых сохраним только те их исходные значения, которые не относятся к первому кластеру. Ниже эти векторы сосредоточены в массиве z. Им соответствуют объекты с номерами:
1, 2, 3, 4, 7, 13, 24, 25, 33, 37, 38, 39, 49.

$$\mathbf{zl}_{\mathbf{i}} := \mathbf{if}\left(\mathbf{N1}_{\mathbf{i}} > 0\,,0\,,\mathbf{y}_{\mathbf{i}}\right)$$

Например: $\mathbf{zl}_1 = \begin{pmatrix} 1.087 \\ 0.777 \end{pmatrix}$ $\mathbf{zl}_6 = 0$

Из массива zl выделим только ту его часть, которая не равна 0:

$$\mathbf{z} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \mathbf{for}\ \ \mathbf{i} \in 1..\mathbf{N} \\ \quad \mathbf{if}\ \ \left|\mathbf{zl}_{\mathbf{i}}\right| \neq 0 \\ \quad\quad \left| \begin{array}{l} \mathbf{q}_{\mathbf{m}} \leftarrow \mathbf{zl}_{\mathbf{i}} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \mathbf{q} \end{array} \right.$$

$\mathbf{Nz} := \mathbf{length}(\mathbf{z})$ $\mathbf{Nz} = 13$

$\mathbf{z}_1 = \begin{pmatrix} 1.087 \\ 0.777 \end{pmatrix}$ $\mathbf{z}_6 = \begin{pmatrix} 1.162 \\ 0.432 \end{pmatrix}$

Таким образом, z представляет собой массив координат объектов, не вошедших в первый кластер.

Теперь следует построить второй кластер, подвергнув данные z тем же преобразованиям, что выше проводились над y.

Находим "центр" n1 этих объектов, строим окружность радиуса r с центром в n1, регистрируем объекты, попавшие в эту окружность, и вычисляем "центр" n2 попавших в окружность точек:

$$\mathbf{n1} := \frac{1}{\mathbf{Nz}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{Nz}} \mathbf{z}_{\mathbf{i}} \qquad \mathbf{n1} = \begin{pmatrix} 1.321 \\ 1.228 \end{pmatrix}$$

$$\mathbf{g1} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \mathbf{for}\ \ \mathbf{i} \in 1..\mathbf{Nz} \\ \quad \mathbf{if}\ \ \left|\mathbf{z}_{\mathbf{i}} - \mathbf{n1}\right| < \mathbf{r} \\ \quad\quad \left| \begin{array}{l} \mathbf{g1}_{\mathbf{m}} \leftarrow \mathbf{z}_{\mathbf{i}} \\ \mathbf{n2} \leftarrow \frac{1}{\mathbf{m}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{g1}_{\mathbf{i}} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \begin{pmatrix} \mathbf{m} - 1 \\ \mathbf{n2} \end{pmatrix} \end{array} \right.$$

$\mathbf{g1} = \begin{pmatrix} 10 \\ \{2,1\} \end{pmatrix}$ $\mathbf{n2} := \mathbf{g1}_2$ $\mathbf{n2} = \begin{pmatrix} 0.888 \\ 0.858 \end{pmatrix}$

Так как $|\mathbf{n1} - \mathbf{n2}| = 0.569$, т.е.величина большая, то

перенесем центр окружности в точку n2 и повторим вычисления:

$$\mathbf{g2} := \left| \begin{array}{l} \mathbf{m} \leftarrow 1 \\ \textbf{for}\ \ \mathbf{i} \in 1..\mathbf{Nz} \\ \quad \textbf{if}\ \ \left|\mathbf{z_i} - \mathbf{n2}\right| < \mathbf{r} \\ \quad\quad \left| \begin{array}{l} \mathbf{g2_m} \leftarrow \mathbf{z_i} \\ \mathbf{n3} \leftarrow \frac{1}{\mathbf{m}} \cdot \sum_{\mathbf{i}=1}^{\mathbf{m}} \mathbf{g2_i} \\ \mathbf{m} \leftarrow \mathbf{m} + 1 \end{array} \right. \\ \begin{pmatrix} \mathbf{m} - 1 \\ \mathbf{n3} \end{pmatrix} \end{array} \right. \qquad \mathbf{g2} = \begin{pmatrix} 10 \\ \{2,1\} \end{pmatrix} \quad \mathbf{n3} := \mathbf{g2}_2 \qquad \mathbf{n3} = \begin{pmatrix} 0.888 \\ 0.858 \end{pmatrix}$$

Находим очередное смещение: $|\mathbf{n2} - \mathbf{n3}| = 0$

Так как изменения положения "центра" не произошло, то точку n2 принимаем за "центр" второго кластера, содержащего десять объектов из тринадцати, не вошедших в первый кластер. Найдем номера этих точек в обозначениях массива z:

$$\mathbf{N2} := \left| \begin{array}{l} \textbf{for}\ \ \mathbf{i} \in 1..\mathbf{Nz} \\ \quad \mathbf{n_i} \leftarrow \mathbf{i}\ \ \textbf{if}\ \ \left|\mathbf{z_i} - \mathbf{n2}\right| < \mathbf{r} \\ \mathbf{n_i} \leftarrow 0\ \ \textbf{if}\ \ \left|\mathbf{z_i} - \mathbf{n2}\right| \geq \mathbf{r} \\ \mathbf{n} \end{array} \right.$$

$\mathbf{N2} =$

	1
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	0
9	9
10	0
11	11
12	0
13	13

$\mathbf{N1} =$

	1
1	0
2	0
3	0
4	0
5	5
6	6
7	0
8	8
9	9
10	10
11	11
12	12
13	0
14	14
15	15
16	…

Это точки с номерами:
1, 2, 3, 4, 5, 6, 7, 9, 11, 13.

Установим соответствие номеров этих точек с их же номерами, но в обозначениях исходного массива y. С этой целью удобно рядом расположить массивы N1 и N2. Из N1 в N2 "перекочевали" те точки, которым в N1 сответствуют нули. Это точки 1, 2, 3, 4, 7, 13, 24, 25, 33, 37, 38, 39, 49. В N2 эти точки последовательно пронумерованы от 1 до 13, причем из этих тринадцати точек во торой кластер попали только десять точек с указанными номерами. Теперь легко установить их первоначальные номера: это точки 1, 2, 3, 4, 7, 13, 24, 33, 38, 49. Три точки с номерами 25, 37, 39 не вошли ни в один из построенных кластеров.

Таким образом, в первые два кластера вошли точки:
первый кластер -
5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 30, 31, 32, 34, 35, 36, 40, 41, 42, 43, 44, 45, 46, 47, 48, 50;
второй кластер -
1, 2, 3, 4, 7, 13, 24, 33, 38, 49.
Теперь формируем третий кластер, зафиксировав точки, не вошедшие в первые два кластера, т.е. точки с номерами 25, 37, 39. Проверим, целесообразно ли их объединить в единый третий кластер или лучше принять за "центры" самостоятельных кластеров. С этой целью найдем их общий "центр":

$$\mathbf{l1} := \frac{1}{3}\cdot\left(\mathbf{y}_{25} + \mathbf{y}_{37} + \mathbf{y}_{39}\right) \qquad \mathbf{l1} = \begin{pmatrix} 2.764 \\ 2.459 \end{pmatrix}$$

и рассчитаем расстояния этих точек от "центра" l1:

$$\left|\mathbf{y}_{25} - \mathbf{l1}\right| = 2.083 \qquad \left|\mathbf{y}_{37} - \mathbf{l1}\right| = 1.23 \qquad \left|\mathbf{y}_{39} - \mathbf{l1}\right| = 2.68$$

$$\mathbf{r} = 1.402$$

В окружность радиуса r с центром в точке l1 попадает только точка $\mathbf{y}_{37}$; поэтому ее примем за "центр" третьего кластера и выявим возможность отнесения к этому третьему кластеру точек $\mathbf{y}_{25}$ и $\mathbf{y}_{39}$, вычислив расстояния:

$$\left|\mathbf{y}_{37} - \mathbf{y}_{25}\right| = 2.126 \qquad \left|\mathbf{y}_{37} - \mathbf{y}_{39}\right| = 3.613 \qquad \left|\mathbf{y}_{25} - \mathbf{y}_{39}\right| = 4.64$$

Так как все эти расстояния превышают радиус r, то объединять точки $\mathbf{y}_{25}$, $\mathbf{y}_{37}$, $\mathbf{y}_{39}$ в единый кластер нецелесообразно и каждую из этих точек следует рассматривать как самостоятельный кластер. В результате начальные 50 точек оказываются распределенными между 5 кластерами, из которых первый содержит 37 объектов, второй- 10 и оставшиеся три по одному объекту - $\mathbf{y}_{25}$, $\mathbf{y}_{37}$ и $\mathbf{y}_{39}$.

Графически отобразим объекты, формирующие первый кластер. С этой целью проведем сквозную нумерацию этих объектов:

$\mathbf{v1}_1 := \mathbf{y}_5$ $\mathbf{v1}_2 := \mathbf{y}_6$ $\mathbf{v1}_3 := \mathbf{y}_8$ $\mathbf{v1}_4 := \mathbf{y}_9$ $\mathbf{v1}_5 := \mathbf{y}_{10}$ $\mathbf{v1}_6 := \mathbf{y}_{11}$

$\mathbf{v1}_7 := \mathbf{y}_{12}$ $\mathbf{v1}_8 := \mathbf{y}_{14}$ $\mathbf{v1}_9 := \mathbf{y}_{15}$ $\mathbf{v1}_{10} := \mathbf{y}_{16}$ $\mathbf{v1}_{11} := \mathbf{y}_{17}$ $\mathbf{v1}_{12} := \mathbf{y}_{18}$

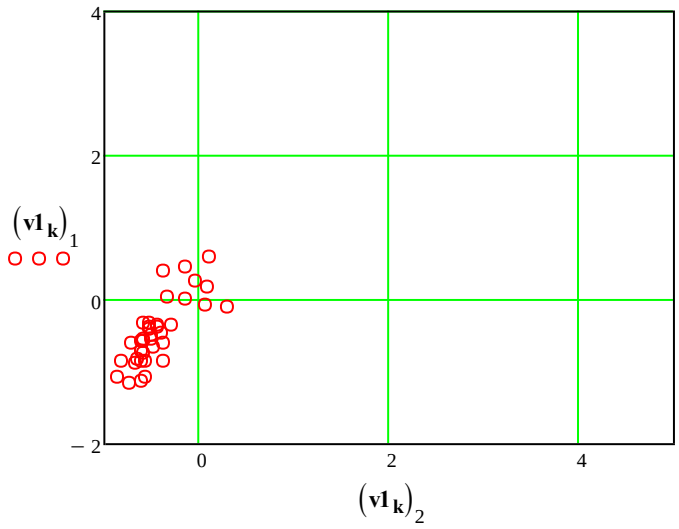
$\mathbf{v1}_{13} := \mathbf{y}_{19}$ $\mathbf{v1}_{14} := \mathbf{y}_{20}$ $\mathbf{v1}_{15} := \mathbf{y}_{21}$ $\mathbf{v1}_{16} := \mathbf{y}_{22}$ $\mathbf{v1}_{17} := \mathbf{y}_{23}$ $\mathbf{v1}_{18} := \mathbf{y}_{26}$

$\mathbf{v1}_{19} := \mathbf{y}_{27}$ $\mathbf{v1}_{20} := \mathbf{y}_{28}$ $\mathbf{v1}_{21} := \mathbf{y}_{29}$ $\mathbf{v1}_{22} := \mathbf{y}_{30}$ $\mathbf{v1}_{23} := \mathbf{y}_{31}$ $\mathbf{v1}_{24} := \mathbf{y}_{32}$

$v1_{25} := y_{34}$ $v1_{26} := y_{35}$ $v1_{27} := y_{36}$ $v1_{28} := y_{40}$ $v1_{29} := y_{41}$ $v1_{30} := y_{42}$

$v1_{31} := y_{43}$ $v1_{32} := y_{44}$ $v1_{33} := y_{45}$ $v1_{34} := y_{46}$ $v1_{35} := y_{47}$ $v1_{36} := y_{48}$

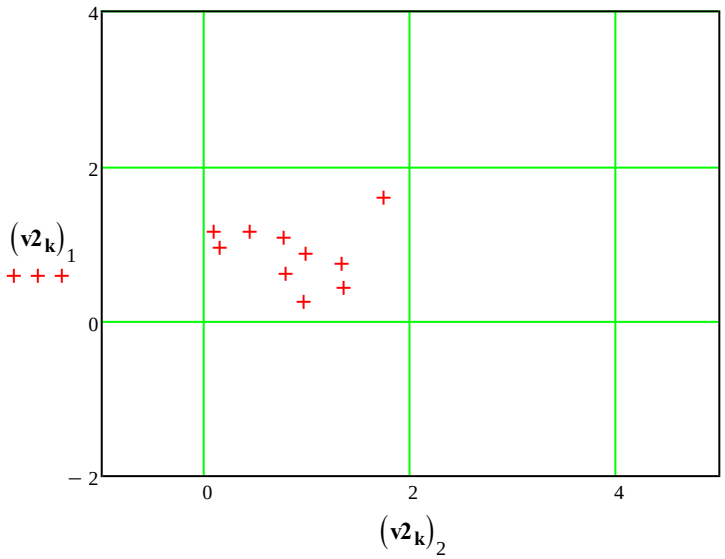
$v1_{37} := y_{50}$ $k := 1..37$



Аналогичным образом отобразим объекты второго кластера:

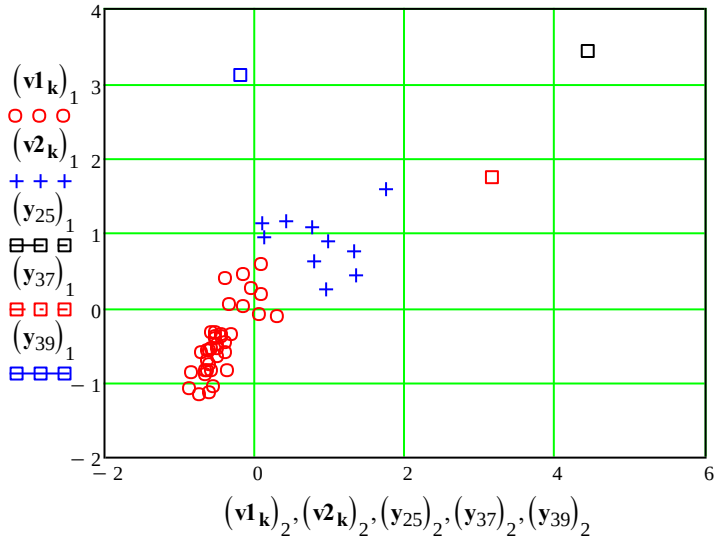
$v2_1 := y_1$ $v2_2 := y_2$ $v2_3 := y_3$ $v2_4 := y_4$ $v2_5 := y_7$ $v2_6 := y_{13}$

$v2_7 := y_{24}$ $v2_8 := y_{33}$ $v2_9 := y_{38}$ $v2_{10} := y_{49}$ $k := 1..10$

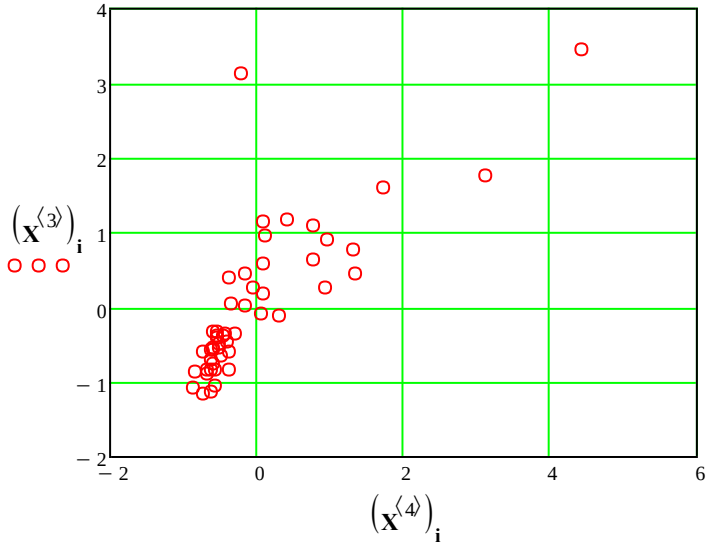


Изобразим все пять кластеров на общем рисунке:

$k := 1..50$



Исходная некласифицированная картина имеет вид:



Из сопоставления этих двух рисунков становится очевидным характер распределения объектов по кластерам.

5. Многомерное шкалирование

5.1. Основное назначение, цели

В первой части пособия [13], рассматривая методы кластерного и факторного анализа, мы отмечали, что эти методы, обладая целым спектром полезных свойств, обладают еще одним замечательным качеством, позволяющим значительно понизить порядок модели представления данных. И это понимается в следующем смысле. Пусть исходные дынные представлены моделью, аналогичной (2.2):

$$\overline{\boldsymbol{X}} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \dots & \dots & \dots & \dots \\ x_{k,1} & x_{k,2} & \dots & x_{k,n} \end{bmatrix}. \qquad (5.1)$$

Здесь $x_{q,j}$ – наблюдаемое значение q-й компоненты вектора $\mathbf{X}$ на j-м объекте $(j = \overline{1,n})$.

Величину k будем воспринимать как *порядок модели* отображения экспериментальных данных. Наряду с матрицей (5.1) рассмотрим матрицу $\mathbf{D}$ расстояний $D_{i,j}$ между k-мерными векторами: $X_i = \begin{bmatrix} x_{1,i} & x_{2,i} & \dots & x_{k,i} \end{bmatrix}^T$ и $X_j = \begin{bmatrix} x_{1,j} & x_{2,j} & \dots & x_{k,j} \end{bmatrix}^T$. Хотя это расстояние может восприниматься различным образом в зависимости от используемой метрики, будем его понимать в евклидовом смысле как в наиболее естественной и физически воспринимаемой форме отображения понятия “расстояние”. Тогда:

$$D_{i,j} = \sqrt{\rho\left(\mathbf{X}_i, \mathbf{X}_j\right)} = \sqrt{\sum_{q=1}^{k} \left(x_{q,i} - x_{q,j}\right)^2}, \qquad (5.2)$$

где $\rho\left(\mathbf{X}_i, \mathbf{X}_j\right)$ – используемая метрика. Расстояния $D_{i,j}$, $i = \overline{1,n}$, $j = \overline{1,n}$, упорядочим в форме n - мерной матрицы расстояний $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} 0 & D_{1,2} & ... & D_{1,n} \\ D_{2,1} & 0 & ... & D_{2,n} \\ ... & ... & ... & ... \\ D_{n,1} & D_{n,2} & ... & 0 \end{bmatrix}. \qquad (5.3)$$

Теперь предположим, что построено новое r - мерное, причем обязательно $r < k$, пространство $\overline{\mathbf{Y}}$, в котором подобным (5.2) образом вычислены расстояния $d_{i,j}$ между элементами $\mathbf{Y}_i, \mathbf{Y}_j$ этого пространства:

$$d_{i,j} = \sqrt{\rho\left(\mathbf{Y}_i, \mathbf{Y}_j\right)} = \sqrt{\sum_{q=1}^{r}\left(y_{q,i} - y_{q,j}\right)^2}, \quad i, j = \overline{1,k},. \qquad (5.4)$$

И построена аналогичная (5.3) матрица $\mathbf{d}$ расстояний:

$$\mathbf{d} = \begin{bmatrix} 0 & d_{1,2} & ... & d_{1,n} \\ d_{2,1} & 0 & ... & d_{2nk} \\ ... & ... & ... & ... \\ d_{n,1} & d_{n,2} & ... & 0 \end{bmatrix}. \qquad (5.5)$$

Будем считать , что цель многомерного шкалирования заключается в построении такого множества $\overline{\mathbf{Y}}$ при обязательном условии $r < k$, при котором элементы матрицы (5.5) в рамках некоторого критерия окажутся близкими к элементам исходной матрицы расстояний (5.3). Таким образом, использование методов многомерного шкалирования позволяет перейти к новым наблюдениям, в общем случае имеющим гипотетический смысл, меньшей размерности, но с сохранением начальной “субординации” наблюдений.

5.2. Немного истории

Вообще говоря, сформулированная цель многомерного шкалирования весьма скромно отражает возможности этого подхода к решению проблем многомерного анализа. Начало этому весьма разветвленному направлению анализа положили работы американских исследователей У.Торгерсона [14, 15], Р. Шепарда [16], Дж.

Краскэла [17]. В последующем идеи этих "пионерских" работ получили обобщение и развитие в отечественных исследованиях (например, [18-21]). Заметим, что эти и другие основополагающие в области многомерного шкалирования работы в основном касались проблематики таких наук, как психология, биология, социология, медицина и в меньшей степени – экономика (например, [22]). Подробный обзор соответствующих публикаций можно обнаружить в [18] и в Интернетовских публикациях. Мы ограничиваемся скромным обзором.

Прикладная направленность "пионерских" работ по многомерному шкалированию наложила свой отпечаток на последующее изложение сути соответствующей задачи. Исходный материал в этих задачах, как правило, сосредоточивается в так называемых *матрицах сходства* или *матрицах различий*. Дать четкое формальное определение этих понятий довольно сложно. Мы будем исходить из интуитивного наполнения этих терминов, понимая под сходством (различием) обследуемых объектов $X_1, X_2, ..., X_n$ некоторые характеристики, являющиеся общими для объектов или указывающими на их индивидуальность. Понятие сходства (различия) объектов "материализуется" в некоторой условной шкале – балльной относительно какого-либо образца или порядковой с указанием рангов объектов. В последующем эти характеристики путем применения различных приемов преобразуются в расстояния. В современном многомерном шкалировании наиболее часто используют матрицу различий, элемент $D_{i,j}$ которой является показателем различия i -го и j -го объектов из состава обследуемых ($(i, j = \overline{1,n})$ и должен обладать свойствами метрики из метрических пространств [7]:

$$D_{i,i} = 0, D_{i,j} = D_{j,i},$$
$$D_{i,j} \le D_{i,k} + D_{k,j}, i, j, k = \overline{1,n}. \qquad (5.6)$$

Так как обследуемые n объектов могут иметь самую различную природу – предприятия, студенческие группы, кандидаты в президенты, университеты и так далее – и при построении матрицы различий нет четких формализованных рекомендаций на ее построение, то ограничения (5.6) могут не выполняться, поэтому поступают так [2].

Обозначим символом $\delta(i, j)$ указанное составителем матрицы различий (респондентом) значение показателя различия i -го и

j -го объектов из числа обследуемых. Тогда для соблюдения условия $D_{i,j} = D_{j,i}$ принимают:

$$D_{i,j} = D_{j,i} = \frac{1}{2}\left[\delta(i,j) + \delta(j,i)\right], \qquad (5.7)$$

т.е полусумме оценок различия i -го и j -го объектов и j -го и i -го объектов.

Если не выполняется последнее условие (5.7) – неравенство треугольника, рекомендации более сложные. В рассмотрение вводится величина

$$c = \max_{i,j,k}\left(\delta(k,j) - \delta(k,i) - \delta(i,j)\right) \qquad (5.8)$$

и полагают:

$$D_{i,j} = \begin{cases} 0, \text{åñëè } i = j, \\ \delta(i,j) + c, \text{èí à÷å,} \end{cases} \quad i,j,k = \overline{1,n}. \qquad (5.9)$$

Справедливо *свойство*: если $\delta(i,j)$ удовлетворяет первым двум условиям из (5.6), то величина (5.9) удовлетворяет всем ограничениям (5.6).

Далее строится матрица Δ^* - так называемая матрица с *двойным центрированием*, у которой среднее значение элементов каждой строки и каждого столбца равняются нулю. Элементы $\delta^*_{i,j}$ этой матрицы вычисляются по формуле

$$\delta^*_{i,j} = -\frac{1}{2}\left(D^2(i,j) - D^2(i,.) - D\delta^2(.,j) + D^2(.,.)\right), \qquad (5.10)$$

где $D^2(i,.)-$ средняя для характеристик различия в j -х столбцах i -й строки, возведенных в квадрат: $D^2(i,.) = \frac{1}{n}\sum_{j=1}^{n} D^2(i,j)$; $D^2(.,j)-$ средняя для характеристик различия в $i-$х строках $j-$го столбца, возведенных в квадрат: $D^2(.,j) = \frac{1}{n}\sum_{i=1}^{n} D^2(i,j)$; средняя для характеристик различия всей матрицы различия, возведенных в квадрат: $D^2(.,.) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} D^2(i,j)$.

Если матрица различий изначально представлена в числовом виде и $v(i,j)-$ количественное различие i -го признака у j -го объекта

($i=\overline{1,k}, j=\overline{1,n}$), то полагают:

$$D(i,j)=\sqrt{\left(\sum_{s=1}^{k}\left(v(s,i)-v(s,j)\right)^2\right)},\quad i,j=\overline{1,n}. \qquad (5.11)$$

Если изначально информация о различиях концентрируется в корреляционной матрице вида:

$$\mathbf{R}_{i,j}=M\{X_i,X_j\}=\begin{bmatrix} 1 & r_{1,2} & \dots & r_{1,n} \\ r_{2,1} & 1 & \dots & r_{2,n} \\ \dots & \dots & \dots & \dots \\ r_{n,1} & r_{n,2} & \dots & 1 \end{bmatrix},$$

то рекомендуется $D_{i,j}=\sqrt{1-r_{i,j}^2}$.

В многомерном шкалировании выдающаяся роль принадлежит *теореме Торгерсона*, являющегося основателем этого направления: если элементы $D(i,j)=\delta(i,j)$ удовлетворяют ограничениям (5.6) и матрица $\mathbf{\Delta}^*$ построена с использованием правила (5.10), то существует такой набор r-мерных($r<k$) точек:

$\mathbf{Y}_i=\begin{bmatrix} y_{i,1} & y_{i,2} & \dots & y_{i,n} \end{bmatrix}^T, i=\overline{1,r},$ которые удовлетворяют условиям

$$\delta_{i,j}^*=\sum_{s=1}^{r} y_{s,i}y_{s,j},\quad i,j=\overline{1,n}, \qquad (5.12)$$

или в матричной форме

$$\mathbf{\Delta}^*=\mathbf{Y}^T\mathbf{Y}, \qquad (5.13)$$

где $r\times n$ матрица $\mathbf{Y}=\begin{bmatrix}\mathbf{Y}_1 & \mathbf{Y}_2 & \dots & \mathbf{Y}_r\end{bmatrix}$.

При доказательстве теоремы дополнительно используется необременительное условие центрированности величин $y_{i,j}$:

$\sum_{j=1}^{n} y_{i,j}=0, i=\overline{0,r}$. По аналогии с (5.11) запишем:

$$D(i,j)=\sqrt{\sum_{s=1}^{r}\left(y_{s,i}-y_{s,j}\right)^2},\quad i,j=\overline{1,n}, \qquad (5.14)$$

и из (5.14) следует:

$$D^2(i,j)=\sum_{i=1}^{r}y_{s,i}^2+\sum_{s=1}^{r}y_{s,j}^2-2\sum_{s=1}^{r}y_{s,i}y_{s,j}\ . \qquad (5.15)$$

Усредняя это равенство по j, получаем:

$$D^2(i,.)=\sum_{s=1}^{r}y_{s,i}^2+\frac{1}{n}\sum_{j=1}^{n}\sum_{s=1}^{r}y_{s,j}^2-\frac{2}{n}\sum_{j=1}^{n}\sum_{s=1}^{r}y_{s,i}y_{s,j}\ .$$

Последнее слагаемое в силу центрированности величин $y_{s,j}$ обращается в нуль, что позволяет записать:

$$D^2(i,.)=\sum_{s=1}^{r}y_{s,i}^2+\sum_{s=1}^{r}y_{s,.}^2\ , \qquad (5.16)$$

где традиционно обозначено:

$$\sum_{s=1}^{r}y_{s,.}^2=\frac{1}{n}\sum_{j=1}^{n}y_{s,j}^2\ .$$

Возвратимся теперь к выражению (5.15) и усредним его по i. Проделав операции, аналогичные только что выполненным, получим

$$D^2(.,j)=\sum_{s=1}^{r}y_{s,j}^2+\sum_{s=1}^{r}y_{s,.}^2\ . \qquad (5.17)$$

Наконец, проводя в (5.15) совместное усреднение по i и j, устанавливаем:

$$D^2(.,.)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{r}y_{s,i}^2+\frac{1}{n}\sum_{i=1}^{n}\sum_{s=1}^{r}y_{s,.j}^2-\frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{s=1}^{r}y_{s,i}y_{s,j}\ .$$

Последнее слагаемое в этом выражении в силу центрированности по-прежнему равно нулю. Поэтому:

$$D^2(.,.)=2\sum_{s=1}^{r}y_{s,.}^2\ . \qquad (5.18)$$

Если теперь в составе (5.10) учесть вычисленные значения слагаемых $D^2(i,j),D^2(i,.),D^2(.,j),D^2(.,.)$, придем к равенству:

$$\delta_{i,j}^{*}=-\frac{1}{2}\sum_{s=0}^{r}-2y_{s,i}y_{s,j}\ ,$$

что совпадает с (5.12), то-есть доказывает справедливость теоремы Торгерсона.

Соотношения (5.10) формируют систему из r уравнений относительно компонентов вектора $\mathbf{Y}$. Разработаны различные вычислительные схемы исследования этой системы (например, [8,9]). Решая эти уравнения, находят положение r-мерных точек, моделирующих положение гипотетических объектов.

5.3. Вычислительные алгоритмы многомерного шкалирования

Наиболее часто при выполнении многомерного шкалирования предполагают, что матрица различий имеет числовую природу, позволяющую различие объектов моделировать в форме евклидовых расстояний между точками в некотором координатном пространстве небольшой размерности, являющимися своеобразными образами объектов в этом пространстве. Решение возникающих при этом задач выполняют в терминах анализа *метрических* пространств.

Если, альтернативно, природа матрицы различий иная и носит, например, ранговый характер, то применяют *неметрические* методы анализа, изобилующие обширным набором вычислительных процедур.

Для практической реализации основополагающих идей многомерного шкалирования в рассмотрение вводят специальную целевую функцию J, называемую *стрессом* и минимизируемую в процессе поиска решения. Эта функция может быть определена различными
способами [6]. Применительно к сформулированному варианту задачи шкалирования зададим ее в виде

$$J = \sum_{i>j} w_{i,j} \left(d_{i,j} - D_{i,j} \right)^2, \qquad j = \overline{1,n}, \tag{5.19}$$

где $D_{i,j}$ – известные начальные расстояния, определяемые, например, аналогичным (5.2) образом, $w_{i,j}$ – веса. Часто рекомендуется $w_{i,j} = D_{i,j}^{\gamma}$, где $\gamma > 0$ при больших расстояниях $D_{i,j}$ и $\gamma < 0$ при малых расстояниях. Обычно ограничиваются $\gamma = -2$. Назначение функции (5.19) заключается в следующем: требуется найти координаты n r-мерных точек и соответствующие им расстояния $d_{i,j}$, определяемые по аналогии с (5.4) и при которых функция (5.19) достигает наименьшего значения.

Функция (5.19) зависит в общем случае от очень большого числа аргументов, что практически делает невозможным ее минимизацию традиционными средствами. Поэтому нам представляется более рациональным поиск оптимального размещения каждой отдельной точки при фиксированных положениях остальных. В связи с этим намерением необходимость использовать двойную индексацию переменных в (5.19) становится неактуальной и можно записать:

$$J(\mathbf{Y}) = \sum_{i \in I} w_i (d_I(\mathbf{Y}) - D_i(\mathbf{Y}))^2, \tag{5.20}$$

где I – множеество индексов всех объектов с известными расстояниями $D_i(Y) = \rho(\mathbf{Y}, \mathbf{Y}_i)$. Минимизацию функции (5.20) будем осуществлять методом Ньютона – Рафсона [7], записав соответствующий алгоритм в виде

$$\hat{\mathbf{Y}}^{k+1} = \hat{\mathbf{Y}}^k - q\nabla^{-2}(\hat{\mathbf{Y}}^k)\nabla(\hat{\mathbf{Y}}^k), k = 0,1,2,\ldots \quad . \tag{5.21}$$

Здесь $\hat{\mathbf{Y}}^k$ – оценка вектора $\mathbf{Y}$ после k-й итерации алгоритма (5.21), $\nabla(\hat{\mathbf{Y}}^k)$ – градиент функции $J(\mathbf{Y})$, вычисленный в точке $\hat{Y}^k$, $\nabla^2(\hat{\mathbf{Y}}^k)$ – вычисленная в той же точке матрица вторых производных (гессиан) функции (5.21), $q > 0$ – параметр, регулирующий скорость сходимости итеративной процедуры (5.21).

Займемся поиском величин, формирующих алгоритм (5.21). По аналогии с (5.4) можем записать относительно некоторой опорной точки y_q^0 :

$$d_i^2 = \sum_{q=1}^{r} (y_q - y_q^0)^2, \quad i = \overline{1,k}. \tag{5.22}$$

Продифференцировав это выражение по y_j, установим:

$$\frac{\partial d_i}{\partial y_j} = \frac{y_j - y_j^0}{d_i}, \qquad j = \overline{1,r}. \tag{5.23}$$

Рассмотрим теперь градиент функции (5.21) $\nabla(\mathbf{Y})$. Как известно, этот вектор представляет собой совокупность частных производных функции (5.20) по всем компонентам вектора $\mathbf{Y}$. Ориентируясь на это определение, найдем

$$\frac{\partial J(\mathbf{Y})}{\partial y_j} = 2\sum_{i\in I} w_i (d_i(\mathbf{Y}) - D_i(\mathbf{Y})) \frac{\partial d_i(\mathbf{Y})}{\partial y_j} =$$

$$2\sum_{i\in I} w_i \left(1 - \frac{D_i(\mathbf{Y})}{d_i(\mathbf{Y})}\right)\left(y_j - y_j^0\right). \tag{5.24}$$

Аналогичным образом находим элементы матрицы вторых производных (гессиана) $\mathbf{\Lambda}^2(\mathbf{Y})$:

$$\frac{\partial^2 J(\mathbf{Y})}{\partial y_j^2} = 2\sum_{i\in I} w_i \left(-D_i(\mathbf{Y}) \frac{\partial d_i^{-1}(\mathbf{Y})}{\partial y_j}\left(y_j - y_j^0\right) + 1 - \frac{D_i(\mathbf{Y})}{d_i(\mathbf{Y})}\right) =$$

$$= 2\sum_{i\in I} w_i \left(D_i(\mathbf{Y}) \frac{\left(y_j - y_j^0\right)^2}{d_i^3} + 1 - \frac{D_1(\mathbf{Y})}{d_i}\right). \tag{5.25}$$

$$\frac{\partial^2 J(\mathbf{Y})}{\partial y_j \partial y_k} = 2\sum_{i\in I} w_i \frac{D_i(\mathbf{Y})}{d_i}\left(\frac{y_j - y_j^0}{d_i}\right)\left(\frac{y_k - y_k^0}{d_i}\right). \qquad (5.26)$$

Вычислив функции (5.24) – (5.26) в явном виде и подставив их в (5.21), получим алгоритм последовательного уточнения оценки $\hat{\mathbf{Y}}$.Покажем, что правило (5.21) минимизирует целевую функцию (5.20). С этой целью рассмотрим значения функции (5.20) на последовательности (5.21):

$$J(\mathbf{Y}^{k+1}) = J\left(\mathbf{Y}^k - q\nabla^{-2}\left(\mathbf{Y}^k\right)\nabla\left(\mathbf{Y}^k\right)\right).$$

Функцию в правой части этого выражения разложим в ряд Тейлора в окрестности точки $\hat{Y}^k$, ограничив разложение линейным членом. В результате получим

$$J(\hat{\mathbf{Y}}^{k+1}) = J(\hat{\mathbf{Y}}^k) - q\nabla^T(\hat{\mathbf{Y}}^k)\nabla^{-2}(\hat{\mathbf{Y}}^{\mathbf{k}})\nabla(\hat{\mathbf{Y}}^k). \qquad (5.27)$$

Так как матрица вторых производных, а равно и обратная к ней, является положительно определенной, т.е. при $\forall \mathbf{b} \neq 0$ выполняется неравенство $\mathbf{b}^{\mathbf{T}}\nabla^{-2}(\hat{\mathbf{Y}}^{\mathbf{k}})\mathbf{b} > 0$, то справедливо

$$\Delta J(\hat{\mathbf{Y}}^k) = J(\hat{\mathbf{Y}}^{k+1}) - J(\hat{\mathbf{Y}}^k) < 0, \qquad (5.28)$$

и движение на множестве точек (5.21) осуществляется в направлении уменьшения целевой функции (5.20). Это движение производят до срабатывания некоторого *правила останова*, среди многочисленных вариантов которого отдадим предпочтение следующему: движение на m-й итерации прекращается, если обнаруживается

$$\| \nabla(\hat{\mathbf{Y}}^m) \| \leq \varepsilon, \qquad (5.29)$$

где $\varepsilon -$ выбранная малая величина и $\| \ldots \| -$ символ нормы.

Результаты решения задачи методами многомерного шкалирования часто сопровождают некоторыми графическими иллюстрациями, известными как *диаграммы Шепарда*. Они строятся так.

На плоскости вводят двумерную систему координат. По оси ординат в этой системе откладываются воспроизведенные расстояния, а по оси абсцисс – истинные расстояния. Разброс точек относительно друг друга является своеобразным показателем эффективности

применения многомерного шкалирования. Часто в целях большей наглядности и информативности диаграммы строят график ступенчатой функции, представляющей собой результат некоторого монотонного преобразования исходных данных. Заметим, что большинство пакетов прикладных программ, предназначенных для решения задач многомерного шкалирования, содержат и подпрограммы построения диаграмм Шепарда. И ими целесообразно руководствоваться в своей научно-исследовательской деятельности.

5.4. Метод многомерного шкалирования в задаче анализа структуры общества

Для иллюстрации техники применения идей многомерного шкалирования рассмотрим результаты применения метода при анализе структуры общества, проведенного в [23]. Под анализом структуры общества понимается результат его расслоения на отдельные части в зависимости от степени удовлетворенности условиями жизни. Изначально были выделены восемь слоев общества в зависимости от их социально-экономического положения и интересов. Эти группы таковы: бедные-бесправные; бедные-приближенные к власти; благополучные-бесправные; благополучные-приближенные к власти; благополучные-властные; богатые-бесправные; богатые-приближенные к власти; богатые-властные. В качестве исходных признаков рассматривались:

X_1 – наличие качеств, необходимых современной России;

X_2 – самооценка профессионализма;

X_3 – опасения потерять работу;

X_4 – самооценка общественного признания;

X_5 – удовлетворенность материальным положением;

X_6 – удовлетворенность жизнью;

X_7 – беспокойство о завтрашнем дне;

X_8 – способность самостоятельно изменить жизнь к лучшему;

X_9 – здоровье.

В следующей таблице (табл.1) приводятся результаты ранжирования этих признаков в диапазоне от единицы (минимальный ранг) до восьми.

Таблица 1

	X_9	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Бедные-бесправные	***1***	***1***	***1***	***1***	***1***	***1***	***1***	***1***	***1***
Бедные-приближенные к власти	***4***	***3***	***3***	***3***	***2***	***2***	***2***	***4***	***3***
Благополучные-бесправные	***2***	***3***	***4***	***4***	***2***	***3***	***3***	***3***	***3***
Благополучные-приближенные к власти	***5***	***5***	***5***	***5***	***5***	***4***	***4***	***4***	***5***
Благополучные-властные	***8***	***7***	***6***	***8***	***7***	***5***	***6***	***6***	***8***
Богатые-бесправные	***4***	***2***	***2***	***2***	***4***	***6***	***5***	***5***	***2***
Богатые-приближенные к власти	***7***	***8***	***7***	***7***	***6***	***7***	***8***	***8***	***6***
Богатые-властные	***6***	***6***	***8***	***6***	***8***	***8***	***7***	***7***	***7***

Далее была построена матрица различий Δ^* с двойным центрированием, причем в качестве меры расстояний использована евклидова метрика. При использовании идеологии метода главных компонентов были найдены собственные числа и векторы матрицы Δ^*. В табл. 2 систематизированы собственные числа $\lambda_i, i = \overline{1,7}.$

Таблица 2

λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
326.746	32.941	8.317	7.637	1.938	0.305.	0.112

Данные этой таблицы используются для определения количества r латентных переменных в составе шкалированой модели. Распространенный подход к этому процессу, известный как критерий "*каменистой осыпи*", заключается и построении графика зависимости стресса от размерности r с последующим выбором рационального r, гарантирующего приемлемый уровень стресса. Практически этот

способ сводится к построению зависимости $\lambda_i, i = 1,2,3,...,7,$ и выбора такого $i = r,$ начиная с которого, график зависимости λ_i как функции i начинает сглаживаться в направлении оси абсцисс. В данном случае, даже не прибегая к практическому построению графика $\lambda(i),$ легко обнаружить $r = 2.$ Как отмечается в [23], эти две латентные переменные можно интерпретировать следующим образом. Первая переменная (ось абсцисс) влияет на социальное самочувствие. Ею можно объяснить такие показатели, как образование, вид трудовой деятельности, материальный достаток и т.д. Вторая шкала (ось ординат) соответствует субъективному фактору, к которому можно отнести психологический склад личности, его самооценку, удовлетворенность жизнью, характеристики ценностно-мотивационной сферы, личностные амбиции и т.п.

Библиографический список

1. Дюран В., Оделл П. Кластерный анализ. М.: Статистика, 1977.
2. Классификация и кластер /под ред. Дж. Вэн Райзина. М.: Мир, 1980.
3. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988.
4. Калинина В.Н., Соловьев В.И. Введение в многомерный статистический анализ: учеб. пособие/ГУУ. М., 2003.
5. Дронов С.В. Многомерный статистический анализ: учеб. пособие. Барнаул: Изд-во Алт. гос. ун-та, 2003.
6. Чураков Е.П. Математические методы в экономических задачах: учеб. пособие/РГРТА. Рязань, 1996.
7. Певзнер Л.Д., Чураков Е.П. Математические основы теории систем. М.: Высшая школа, 2009.
8. Боровиков В.П., Боровиков И.П. Statistica. Статистический анализ и обработка данных в среде Windows. М.: Филинъ, 1998.
9. Боровиков В.П., Ивченко Г.И Прогнозирование в системе Statistica в среде Windows. М.: Финансы и статистика, 2006.
10. Айвазян С.А. и др. Прикладная статистика. Классификация и снижение размерностей. М.: Финансы и статистика, 1989.

11. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М.: Финансы и статистика, 2000.
12. Загоруйко Н.Г. Методы распознавания и их применение М.: Советское радио, 1972.
13. Чураков Е.П. Многомерные статистические методы. Ч.1: учеб.пособие. /РГРТУ. Рязань, 2013.
14. Torgenson W.S. Multidimensional scaling. Theory and method. – Psychometrika, 1952, v.17, № 4.
15. Torgenson W.S. Theory and methods of scaling.- N.Y. John Wiley, 1958.

Известен русскоязычный вариант: Торгенсон УС. Многомерное шкалирование. Теория и метод. Статистическое измерение качественных характеристик.- М.: Статистика, 1972.

16. Shepard R.N. The analysis of proximities multidimensional scaling with an unknown distance function. Psychometrika, 1962, v.27, № 2-3.
17 .Kruskal J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrica, 1964, v.29, № 1-2.
18. Терехина А.Ю. Методы многомерного шкалирования и визуализации данных (обзор)// Автоматика и телемеханика .1973,№7
19. Терехина А.Ю. Анализ данных методами многомерного шкалирования. – М.: Наука, 1986.
20. Клигер С.А., Косолапов М.С., Толстова Ю..А. Шкалирование при сборе и анализе социологической информации. М.: Наука, 1978.
21. Толстова Ю.А. Основы многомерного шкалирования. М.: 2006.
22. Дейвисон М. Многомерное шкалирование и методы наглядного представления данных. - М.: Финансы и статистика, 1988.
23. Трусова А.Ю. Многомерное шкалирование структуры общества. Вестник СамГУ. 2008. № 7 (66).

Предметный указатель

Агломеративный метод……..9
Алгоритм Ньютона- Рафсона……36
Взвешенное евклидово расстояние……5
Гельдерово расстояние……4
Гессиан……37
Градиент стресса……37

Дендрограмма……9
Диаграмма Шепарда…….42

Задача кластерного анализа……4
Евклидово расстояние……..4

Кластер……4.
Критерий “каменистая осыпь”…….44

Матрица экспериментальных данных……..…3
Минковского расстояние……...4
Метрические и неметрические методы анализа…… 40
Метод Ньютона - Рафсона……..41

Положительно определенная функция……..42
Порядок модели……36

Расстояние по принципу “ближайшего соседа”……..6
Расстояние по принципу “дальнего соседа”……6
Расстояние по принципу “средней связи”……7
Расстояние “city-block”…….……..5
Расстояние Махалонобиса…….…….5
Расстояние множества общее……..……...6
Расстояние по “центрам тяжести”……7

Супремум расстояния……..…...4
Стандартизованная форма…….....3
Среднестатистическое расстояние……6
Стресс……..…….……...40

Теорема Таргерсона……..…….38
Таксон……...4
Цель многомерного шкалирования……..37

Оглавление

4. Основы кластерного анализа……3
4.1. Назначение кластерного анализа…….……3
4.2. Расстояние между объектами…….……...4
4.3. Расстояние между кластерами…….……6
4.4. Индивидуальные свойства кластеров…….……..7
4.5 Целевые функции классификации…….……8
4.6. Основные алгоритмы кластерного анализа……...10
4.7. Примеры компонентного и кластерного анализа……..…15

5. Многомерное шкалировние…… 34
5.1. Основное назначение, цели……..34
5.2. Немного истории…..…….35
5.3. Вычислительные алгоритмы многомерного шкалироваия…. .40
5.4. Метод многомерного шкалироваиия анализа
структуры общества……..43