1. Методы систематизации наблюдений

1.1. Метод главных компонент в случае $k \geq 2$ переменных

Рассмотрим ситуацию с числом признаков, большим двух. В большинстве случаев эти признаки коррелированы, поскольку являются результатами частных измерений некоторых интегральных характеристик. В медицине эти скрытые признаки чаще всего отражают степень тяжести того или иного заболевания и мощность защитных сил организма и складываются из влияния целого комплекса характеристик. Часто оказывается, что эти интегральные характристики описываются главными компонентами. Идея метода главных компонент состоит в том, чтобы заменить k-мерную случайную величину при наименьшей потере информативности на m- мерную $(m < k)$.

Что означает наименьшая потеря информации? Эффективность любого выбора зависит от того, в какой степени эти m линейных функций дают возможность реконструировать или восстановить k первоначальных величин. Один из методов реконструкции этой первоначальной случайной величины состоит в построении ее наилучшего предиктора на основе m линейных функций. Оказывается, что наилучшим выбором линейных функций является выбор первых m главных компонент.

Определим понятие главной компоненты. Пусть имеется k признаков $X_1, \ldots, X_k$. Первой главной компонентой Y_1 называется сохраняющая расстояние между точками линейная комбинация исходных признаков

$$Y_1 = \alpha_{11} X_1 + \ldots + \alpha_{k1} X_k,$$

где коэффициенты $\alpha_{11}, \ldots, \alpha_{k1}$ выбираются таким образом, чтобы дисперсия $D(Y_1) = \lambda_1$ была максимальной. Это соответствует тому, что по первой главной компоненте индивиды должны отличаться наибольшим образом.

Вторая главная компонента также является линейной комбинацией исходных признаков:

$$Y_2 = \alpha_{12}X_1 + \ldots + \alpha_{k2}X_k,$$

где коэффициенты $\alpha_{12}, \ldots, \alpha_{k2}$ выбираются таким образом, что компоненты Y_1 и Y_2 некоррелированы, а дисперсия $D(Y_2) = \lambda_2$ является максимальной из всех линейных комбинаций, некоррелированных с Y_1, то есть вторая компонента должна нести наибольшую новую информацию, не имеющую отношения к первой главной компоненте. Аналогично строятся остальные главные компоненты

$$Y_j = \sum_{i=1}^{k} \alpha_{ij}X_i, \ j = 1, \ldots, k.$$

Суммарная дисперсия остается неизменной:

$$V = D(X_1) + \ldots D(X_k) = \lambda_1 + \ldots + \lambda_k.$$

Значимость главных компонент определяется долей объясняемой ими дисперсии, равной $\frac{\lambda_i}{V} \cdot 100\%$.

Факторами называются нормированные главные компоненты: $Y_j/\sqrt{\lambda_i}$.

Для того чтобы сделать факторную обработку в *StatSoft*, нужно запустить программу *Statistics - Multivariate Exploratory Techniques - Factor Analysis*. Далее нужно выбрать переменные, по которым строятся факторы (*Variables*). Нажимаем *OK*.

В окошке *Max.no.of factors* нужно указать необходимое для анализа число факторов, оно должно быть больше нуля и не больше количества участвующих в анализе признаков.

В окошке *Mini. eigenvalue* лучше сразу поставить 0. Читаем, сколько переменных и сколько наблюдений участвует в анализе. Далее *OK*. Для правильного чтения результатов необходимыми являются сведения, излагаемые далее.

Вычисление коэффициентов главных компонент

Пусть признаки $X = (X_1, \ldots, X_k)^T$ центрированы $EX_i = 0$ и имеют ковариационную матрицу $\Sigma = EXX^T$. Обозначим через A_i собственные векторы матрицы Σ, соответствующие собственным числам λ_i

$$\Sigma A_i = \lambda_i A_i, \quad A_i^T A_j = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad i = 1, \ldots, k. \tag{1}$$

Упорядочим собственные числа по убыванию: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$. Соберем все собственные вектора в одну ортогональную матрицу

$$\mathcal{A} = [A_1, \ldots, A_k] = \begin{bmatrix} a_{11}, & \ldots, & a_{1k} \\ . & . & . \\ a_{k1}, & \ldots, & a_{kk} \end{bmatrix}, \quad \mathcal{A}^T\mathcal{A} = I = \begin{bmatrix} 1, & \ldots, & 0 \\ . & . & . \\ 0, & \ldots, & 1 \end{bmatrix}$$

и определим главную компоненту как

$$Y_j = A_j^T X = \sum_{i=1}^{k} a_{ij} X_i. \tag{2}$$

Для всех главных компонент справедливо выражение:

$$Y = \begin{bmatrix} Y_1 \\ . \\ Y_k \end{bmatrix} = \begin{bmatrix} A_1^T X \\ . \\ A_k^T X \end{bmatrix} = \mathcal{A}^T X, \quad \text{откуда} \quad X = \mathcal{A}Y. \tag{3}$$

Теорема (о некоррелированности главных компонент и инвариантности суммарной дисперсии).

$$1. \quad EY_iY_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \qquad 2. \quad \sum_{i=1}^{k} DX_i = \sum_{j=1}^{k} DY_j.$$

Доказательство.

$$1. EY_iY_j = EY_iY_j^T \overset{(2)}{=} EA_i^T X (A_j^T X)^T = A_i^T \Sigma A_j \overset{(1)}{=}$$

$$= A_i^T \lambda_j A_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j. \end{cases}$$

Следовательно, дисперсия главной компоненты равна соответственному собственному числу ковариационной матрицы: $DY_j = \lambda_j$. Из упорядоченности собственных чисел получаем упорядоченные по убыванию дисперсии главные компоненты: $DY_1 \geq DY_2 \geq \ldots \geq DY_k$.

$$2.\ DX_1 + \ldots + DX_k = \sum_{i=1}^{k} EX_i^2 = EX^T X = \overset{(3)}{=}$$

$$= EY^T \mathcal{A}^T \mathcal{A} Y = EY^T Y = \sum_{j=1}^{k} EY_j^2 = DY_1 + \ldots + DY_k.$$

Факторные нагрузки и формула восстановления

Корреляция $\beta_{ij} = cor(X_i, Y_j)$ между признаком X_i и главной компонентой Y_j называется *факторной нагрузкой*. Пусть σ_i – стандартное отклонение X_i, λ_j – дисперсия главной компоненты Y_j. Тогда факторная нагрузка имеет вид:

$$\beta_{ij} = cor(X_i, Y_j) = \frac{EX_i Y_j}{\sigma_i \sqrt{\lambda_j}} \overset{(3)}{=} \frac{E \sum\limits_{t=1}^{k} \alpha_{it} Y_t Y_j}{\sigma_i \sqrt{\lambda_j}} = \frac{\alpha_{ij} \sqrt{\lambda_j}}{\sigma_i} \tag{4}$$

Формулу восстановления X_i по главным компонентам можно выразить через факторы $G_j = Y_j / \sqrt{\lambda_j}$:

$$X_i = \sum_{j=1}^{k} \alpha_{ij} Y_j = \sigma_i \sum_{j=1}^{k} \frac{\beta_{ij}}{\sqrt{\lambda_j}} G_j \sqrt{\lambda_j} = \sigma_i \sum_{j=1}^{k} \beta_{ij} G_j \,.$$

При восстановлении переменных по m главным компонентам, меньшему количества исходных признаков k, значения признаков могут восстанавливаться с ошибками. Чем больше вклад используемых в восстановлении главных компонент, тем меньше ошибки восстановления.

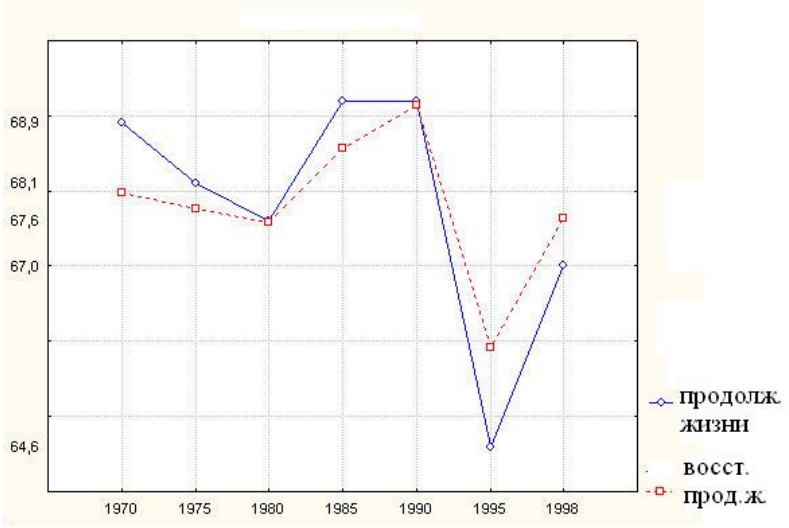


Рис. 1. Восстановление средней продолжительности жизни по первому и второму факторам. $\widehat{L} = (0.896 f_1 + (-0.398 f_2)) \cdot 1.64 + 67.8$,

факторные нагрузки

	Factor 1	Factor 2
L	0.896	−0.398
M	−0.815	−0.564
P	0.905	−0.045
A	−0.847	−0.486
V	−0.772	0.613
Дисп.гл.комп. λ	3.60	1.06
вклад в дисп.	72%	21.8%

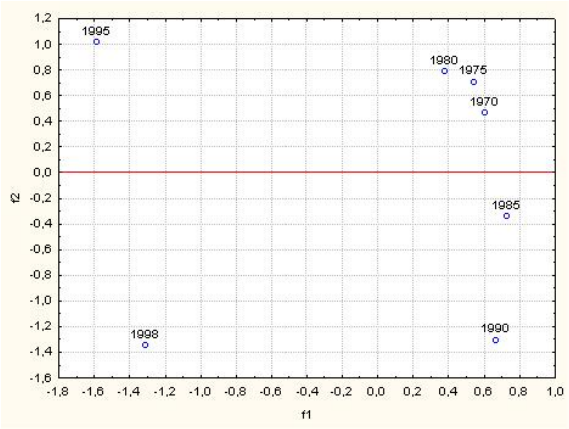
Значения факторов

годы	f_1	f_2
1970	0.600	0.465
1975	0.540	0.710
1980	0.375	0.790
1985	0.724	−0.339
1990	0.662	−1.301
1995	−1.587	1.019
1998	−1.314	−1.345

ПРИМЕР 1. Признаки: L – средняя продолжительность жизни; M – количество чиновников; A – количество автомобилей; P – доходы бедных; V – объемы продажи водки.

Для того чтобы увидеть матрицу факторных нагрузок, нужно открыть вкладку *Loading* и нажать кнопку *Summary* : *Factor loadings*. Для получения значений факторов G_j нужно во вкладке *Scores* выбрать *Factor Scores* или *Save factorscores*.

Вклад первого фактора равен 72%. Значения в i-й строке и j-м столбце соответствуют коэффициенту корреляции между i-м признаком и j-й главной компонентой. Чем больше первый фактор, тем больше продолжи-



тельность жизни и доходы бедных, меньше чиновников и автомобилей и не много водки – фактор благополучия. С первым фактором имеет значимую корреляцию количество заключаемых браков.

Второй фактор менее значим, его вклад равен 22%. Большие значения второго фактора соответствуют тому, что много пьют, а автомобилей мало, соответственно малые наоборот: автомобилей много, пьют мало. Этот фактор имеет значимо положительную корреляцию с числом разводов.

1.2. Кластеризация индивидов в случае метрических признаков

Пусть имеется n индивидов и p признаков, x_{ij} – наблюдение i-го признака и j-го индивида. В качестве мер сходства и различия используются следующие метрики:

1. евклидова метрика, равная геометрическому расстоянию между точками (например, индивидами 1 и 2) в p-мерном пространстве:

$$\sqrt{\sum_{k=1}^{p}(x_{i1}-x_{i2})^2}$$

2. квадрат евклидовой метрики:

$$\sum_{k=1}^{p}(x_{i1}-x_{i2})^2$$

3. манхеттенская метрика:

$$\sum_{k=1}^{p}|x_{i1}-x_{i2}|$$

4. Метрика Брея-Картиса:

$$\frac{\sum\limits_{k=1}^{p}|x_{i1}-x_{i2}|}{\sum\limits_{k=1}^{p}x_{i1}+\sum\limits_{k=1}^{p}x_{i2}}$$

5. Канберровская метрика:

$$\sum_{k=1}^{p}\left(\frac{|x_{i1}-x_{i2}|}{x_{i1}+x_{i2}}\right).$$

6. Метрика Адкинса:

$$\sum_{k=1}^{p}\left(\frac{|x_{i1}-x_{i2}|}{|x_{i1}|+|x_{i2}|}\right).$$

Последние три метрики используются часто при большом различии масштабов измерения. Например, если единицы измерения одного признака миллиметры, а другого километры, то ненормированные евклидовы и манхеттенские метрики будут измерять меру сходства и различия только по признаку с большими единицами измерения.

Кластеризация индивидов осуществляется на основе выбора специальной стратегии объединения. Сначала рассматриваются всевозможные пары индивидов и пары с минимальным различием объединяются в группы. Далее встает вопрос, как измерить расстояние между группами. Существует

несколько стратегий измерения расстояния между группами. Все они по формальному признаку сводятся к выражению:

$$d_{kh} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|,$$

где через h обозначена группа, полученная в результате объединения индивидов i и j, $h = i \cup j$, через k – группа, расстояние до которой от группы h определяется в результате выбранной стратегии. Параметры $\alpha_i, \alpha_j, \beta$ и γ называются параметрами стратегии.

Например, параметры вида $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = -0.5$ определяют стратегию ближнего соседа, когда за расстояние между группами принимается расстояние между ближайшими соседями. Эта стратегия уменьшает расстояния между группами или, как говорят, сжимает пространство.

Параметры вида $\alpha_i = \alpha_j = 0.5$, $\beta = 0$, $\gamma = 0.5$ определяют стратегию дальнего соседа, когда за расстояние между группами принимается самое дальнее расстояние между элементами групп. Эта стратегия расширяет пространство.

Параметры $\alpha_i = n_i/n_h, \alpha_j = n_j/n_h$, $\beta = 0$, $\gamma = 0$ определяют стратегию группового среднего, когда за расстояние между группами принимается расстояние между ближайшими соседями. Через n_i обозначено количество элементов в i-й группе. Недостаток этой стратегии – немонотонность, но она не изменяет пространство.

Существует еще центроидная стратегия, когда расстояние между группами определяется как расстояние между центрами тяжести.

Стратегия, удовлетворяющая условиям монотонности и сохранения метрики пространства, называется гибкой стратегией (метод $ward$). Соответствующие ей параметры имеют вид: $\alpha_i = \alpha_j = 0.625$, $\beta = -0.25$, $\gamma = 0$.

Пример кластерной обработки в R с параметрами $metric =' manhattan'$ и $method =' average'$.

$library(\text{"}cluster\text{"})$

$data.0 < -scale(data)$

$dist < -daisy(data.0, metric = \text{"}manhattan\text{"})$

$h < -hclust(dist, method = \text{"}average\text{"})$

$plot(h)$

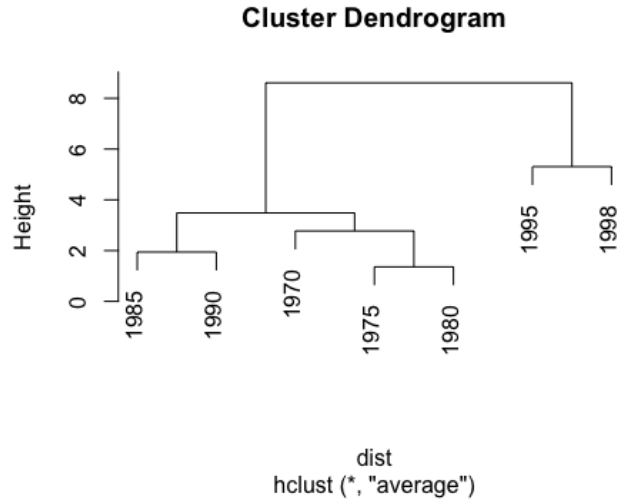


Рис. 2. Дерево кластеризации по метрическим признакам.

Для того чтобы все признаки участвовали в кластеризации индивидов, полезно исходные данные центрировать (вычитание математического ожидания) и нормировать (деление на стандартное отклонение). В этом случае кластеризация осуществляется с учетом вклада всех признаков, а не только признака с наибольшим масштабом измерения.

1.3. Кластеризация признаков

Разбиение признаков на кластеры можно осуществлять также как и в случае индивидов. Для этого нужно только транспонировать матрицу наблюдений. Наряду с перечисленными метриками используются также метрики, связанные с коэффициентом корреляции. Напомним, что коэффициент корреляции является характеристикой меры линейной зависимос-

ти между признаками ξ и η:

$$\varrho = \frac{E(\xi - E\xi)(\eta - E\eta)}{\sqrt{E(\xi - E\xi)^2}\sqrt{E(\eta - E\eta)^2}}.$$

Если признаки не зависимы, то коэффициент корреляции равен нулю. Если признаки связаны линейной зависимостью вида $\eta = k\xi + b$, то коэффициент корреляции равен ± 1 в зависимости от знака коэффициента k.

При наличии матрицы наблюдений выборочный коэффициент кореляции между первым и вторым признаками вычисляется по формуле:

$$r = \frac{\sum\limits_{j=1}^{n} x_{1j}x_{2j} - n\bar{x_1}\bar{x_2}}{\sqrt{\sum\limits_{j=1}^{n} x_{1j}^2 - n\bar{x_1}^2}\sqrt{\sum\limits_{j=1}^{n} x_{2j}^2 - n\bar{x_2}^2}},$$

где также считаем, что имеется n индивидов и p признаков, x_{ij} – наблюдение i-го признака и j-го индивида, $\bar{x}_i$ – выборочное среднее i-го признака.

В качестве меры сходства между признаками выберем метрику, равную $1 - r^2$. Чем ближе коэффициент корреляции к нулю, тем больше метрика. Чем меньше метрика, тем больше связаны линейной зависимостью, т.е. можно сказать, тем меньше расстояние между признаками.

ПРИМЕР 1. Рассмотрим данные о жизни России за последние 30 лет.

Года	Продолж-ть жизни L	Кол-во (тыс.чел.) чиновников M	Водка (бут.в год) V	Доходы бедных P	Автомобили (шт.на 1000 чел.) A
1970	68.9	1060	25.3	7.8	5.5
1975	68.1	1101	28	9.5	15.3
1980	67.6	1147	30	10.1	30.2
1985	69.2	1204	23.5	10	44.5
1990	69.2	1602	18	9.8	58.6
1995	64.6	1893	38.4	5.5	93.3
1998	67	2777	29.6	6.2	122

ПРИМЕР. Если в качестве метрики рассматривать характеристику $1 - r$, показывающую степень удаленности от положительной корреляции, то получится матрица расстояний следующего вида:

$1-r$	L	M	P	A	V
L	0				
M	1.50	0			
P	0.23	1.70	0		
A	1.60	**0.05**	1.67	0	
V	1.93	0.70	1.68	0.63	0

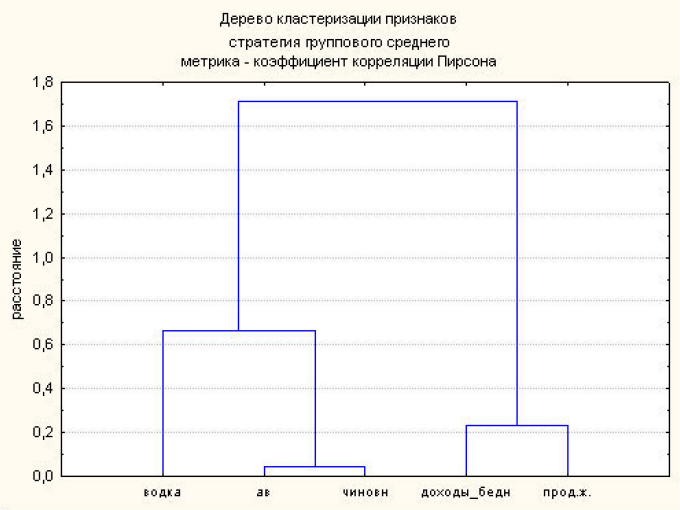


Рис. 3. Дерево кластеризации метрических признаков.

Также близкими являются признаки A - автомобили и M - чиновники.

d	AM	L	P	VM
AM	0			
L	1.92	0		
P	2.09	**0.23**	0	
V	0.82	1.93	1.68	0

$$d_{AM,L} = 0.625(d_{A,L} + d_{M,L}) - 0.25d_{A,M} = 0.625(1.6 + 1.5) - 0.25 \cdot 0.05 = 1.92,$$

$$d_{AM,V} = 0.625(d_{A,V} + d_{M,V}) - 0.25d_{A,M} = 0.625(0.63 + 0.7) - 0.25 \cdot 0.05 = 0.82,$$

$$d_{AM,P} = 0.625(d_{A,P} + d_{M,P}) - 0.25d_{A,M} = 0.625(1.67 + 1.7) - 0.25 \cdot 0.05 = 2.09.$$

Находим, что минимальное расстояние оказывается между продолжительностью жизни L и доходами бедных P.

Итак, имеется пара признаков AM и LP. К какой паре можно отнести

количество продаваемой водки?

$$d_{V,LP} = 0.625(d_{V,L} + d_{V,P}) - 0.25d_{L,P} = 0.625(1.93 + 0.70) - 0.25 \cdot 0.23 = 2.2$$
$$d_{LP,AM} = 0.625(d_{L,AM} + d_{P,AM}) - 0.25d_{L,P} = 0.625(1.92 + 2.09) - 0.25 \cdot 0.23 = 2.45,$$
$$d_{V,AM} = 0.82.$$

Минимальное расстояние оказывается между признаками автомобили-чиновники и количеством продаваемой водки. Автоматически это можно получить при помощи следующих функций.

$dist < -daisy(t(data.0), metric = "euclidean")$

$h < -hclust(dist, method = "ward.D")$

$plot(h)$

1.4. Кластерный анализ по категориальным признакам

Качественными или категориальными называются признаки, градации которых можно менять местами. Например, цветы могут быть красными, синими, белыми, желтыми или лилиями, розами, гиацинтами и т.д. Пусть имеется группа n индивидов (букет цветов), характеризуемая признаком ξ (тип цветка) со значениями $x_1, \ldots, x_m$ (ромашка, колокольчик, гвоздика и т.д.), $n = a_1 + a_2 + \ldots + a_m$, где a_i – количество индивидов со значением x_i признака ξ (a_1 число ромашек, a_2 колокольчиков и так далее).

Информационное разнообразие I этой группы измеряется формулой:

$$I = n \ln n - \sum_{i=1}^{m} a_i \ln a_i.$$

Например, информационное разнообразие букета, состоящего из двух гвоздик, одного тюльпана и двух хризантем равно

$$I_1 = 5 \ln 5 - 2 \ln 2 - 1 \ln 1 - 2 \ln 2 = 5.275.$$

В группе с одинаковыми значениями признака информационное разнообразие равно нулю, так как $I = n \ln n - 0 \cdot \ln 0 - n \ln n = 0$. Очевидно, что

разнообразие группы, состоящего из одного индивида, также равно нулю.

Если группа индивидов характеризуется несколькими признаками (например, в букете гвоздики и тюльпан красные, а хризантемы белые), то разнообразие складывается из информационных разнообразий по каждому признаку. Разнообразие по цвету равно:

$$I_2 = 5\ln 5 - 3\ln 3 - 2\ln 2 = 3.365,$$

отсюда общее разнообразие равно $I = I_1 + I_2 = 5.275 + 3.365 = 8.64$.

Информационный выигрыш от слияния двух групп A и B с разнообразиями I_A и I_B равен разности между разнообразием объединенной группы и разнообразиями исходных групп до объединения:

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B.$$

Информационный выигрыш от слияния групп может быть нулевым или положительным. Например, если объединяются букет роз и букет ромашек, разнообразие которых по-отдельности равно нулю, то выигрыш положительный, так как разнообразие смешанного букета положительно. Если к ромашке с колокольчиком присоединить еще одну пару ромашки с колокольчиком, то информационный выигрыш окажется нулевым, так как

$$I_A = 2\ln 2 - 1\ln 1 - 1\ln 1 = 2\ln 2,$$
$$I_B = 2\ln 2,$$
$$I_{A+B} = 4\ln 4 - 2\ln 2 - 2\ln 2 = 4\ln 4 - 4\ln 2 = 4(\ln 4 - \ln 2) = 4\ln 2,$$
$$\Delta I = 4\ln 2 - 2\ln 2 - 2\ln 2 = 0.$$

Рассмотрим процедуру построения кластеров для шести лиц пяти типов, условно изображенных на рис.4. Каждый индивид характеризуется четырьмя признаками глаза: 1 - большие, круглые, 0 - узкие, мелкие; нос: 1 -

большой, толстый, 0 - тонкий, маленький; рот: 1 - большой, 0 - узкий; уши: 1 - большие, музыкальные, 0 - маленькие. Первый индивид характеризуется вектором признаков (1110), второй (1010), два третьих (1101), четвертый (0101) и пятый (0000). Сначала для каждой пары индивидов вычисляем количество признаков, по которым индивиды отличаются,

	1	2	3	4	5
1	–	1	2	3	3
2		–	3	4	2
3			–	1	3
4				–	2
5					–

и получаем, что больше всего друг на друга походят индивиды первый и второй (отличаются только носами) , а также два третьих и четвертый (только глазками). Таким образом, мы агрегируем этих индивидов в три группы: $A(1,2)$, $B(3,3,4)$ и $C(5)$. Соответственно их разнообразия равны

$$I_{A(1,2)} = 2\ln 2 - 1\ln 1 - 1\ln 1 = 1.37,$$
$$I_{B(3,3,4)} = 3\ln 3 - 2\ln 2 - 1\ln 1 = 1.91,$$
$$I_C = 0.$$

В табл.1 показано, как вычисляются эти разнообразия и разнообразия от объединения групп. Отсюда информационные выигрыши от объединения групп соответственно равны:

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B = 11.73 - 1.37 - 1.91 = 8.44,$$
$$\Delta I_{A+C} = I_{A+C} - I_A - I_C = 5.73 - 1.37 - 0 = 4.34,$$
$$\Delta I_{B+C} = I_{B+C} - I_B - I_C = 7.27 - 1.91 - 0 = 5.36.$$

Наименьший выигрыш достигается от объединения групп $A(1,2)$ и

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{A+C}	1.91	1.91	1.91	0	5.73

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
C	5		0	0	0	0	
		I_C	0	0	0	0	0
		I_{B+C}	2.77	2.25	0	2.25	7.27

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
A	2		1	0	1	0	
		I_A	0	1.37	0	0	1.37
	3		1	1	0	1	
	3		1	1	0	1	
B	4		0	1	0	1	
		I_B	1.91	0	0	0	1.91
		I_{A+B}	2.5	2.5	3.37	3.37	11.73

Таблица 1. Вычисление разнообразия объединений групп $A(1,2)$-$C(5)$, $B(3,3,4)$-$C(5)$ и $A(1,2)$-$B(3,3,4)$.

$C(5)$. Поэтому получаем два кластера: (1, 2, 5) и (3,3,4), которые отличаются ушками (рис.4).

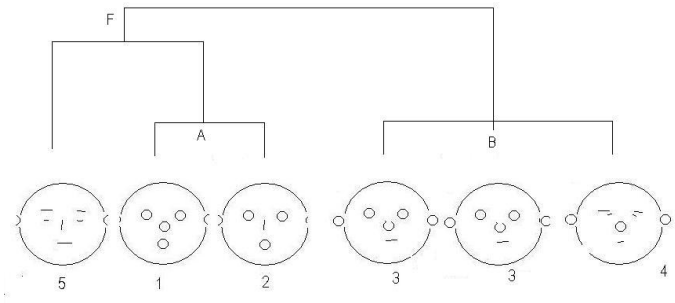


Рис. 4. Дерево кластеризации.

Задание 1. Рассмотрим кластеры $F(1, 2, 5)$ и $B(3, 3, 4)$. Нужно заполнить таблицу 2, для этого не обязательно пересчитывать, можно воспользоваться результатами табл.1, затем вычислить информационные выигрыши от объединения

$$\Delta I_{F+D} = I_{F+D} - I_F - I_D,$$
$$\Delta I_{B+D} = I_{B+D} - I_B - I_D,$$

и по наименьшему из этих выигрышей выяснить, к кому из них примкнет $D(6)$ индивид (1001) шестого типа (рис.5).

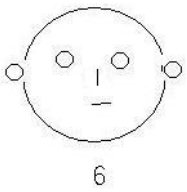


Рис. 5. Индивид (1001) шестого типа.

группа	индивиды	разнообразие	глаза	нос	рот	уши	Сумма
	1		1	1	1	0	
	2		1	0	1	0	
F	5		0	0	0	0	
		I_F					
D	6		1	0	0	1	
		I_D					
		I_{F+D}					

Таблица 2. Форма для вычисления разнообразия объединений групп $F(1,2,5)$-$D(6)$, $B(3,3,4)$-$D(6)$.