Лекция 10. Многомерен клъстерен анализ (КА)

Доц. д-р Снежана Гочева-Илиева
ПУ "П.Хилендарски", snow@uni-plovdiv.bg

Обща характеристика

Клъстерният анализ (КА) е общо название на множество изчислителни процедури, използувани при създаването на класификации на обекти. Той позволява да се открият нови зависимости и свойства в дадено множество от данни, които не биха могли да се установят по други известни теоретични и експериментални методи.

Класифицираните обекти могат да бъдат както наблюдения (случаи), така и променливи. В специализираната статистическа литература не съществува дефиниция на понятието клъстер, която може да претендира за универсалност. Най-общо може да се каже, че клъстерът е подсъвкупност

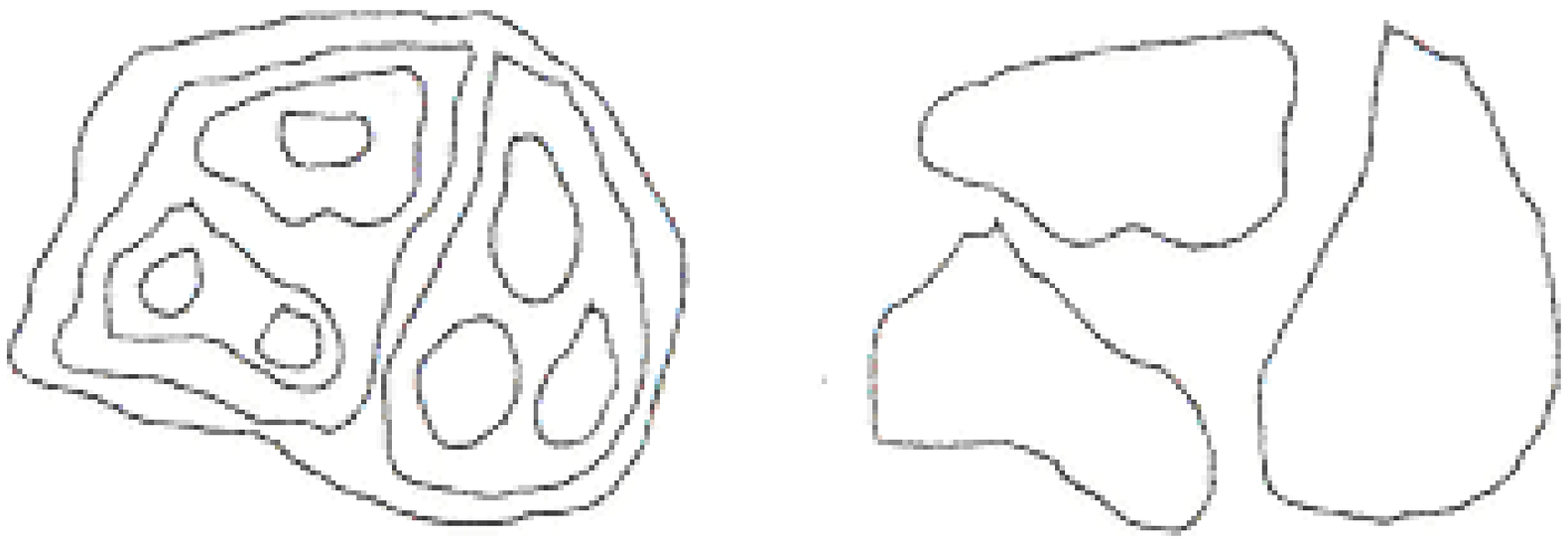
(група) от обекти, които са еднакви в определена степен и в съответствие с определен критерий.

Като редуцираща техника КА има общи елементи с дискриминантния и факторния анализ. Например при дискриминантния анализ също се класифицират единици, но в предварително определени групи. Различен е и познавателният подход при двата анализа. При дискриминантния анализ се изследват различията между известните групи и възможностите за разработване на ефективно правило за класификация на единиците, а при КА се изследват възможностите за формиране на логически обосновани и практически обясними еднородни групи (клъстери) от единици. При ФА се класифицират променливи величини в общи фактори, а не единици.

Целта на КА е да се намери оптимално групиране на наблюденията (или на променливите), при което елементите от даден клъстер са подобни, но клъстерите един с друг ясно се отличават.

Обикновено в КА броят на групите предварително е неизвестен и се определя от изследователя в процеса на изследването.

Съществуват различни видове КА, най-разпространените са: йерархични методи и КА по метода на К-средните (разделящи или нейерархични). В случая на интервални данни и сравнително малък размер на извадката (под 500), по-подходящи са йерархични методи, на които ще се спрем по-подробно.



Фиг. 1. а) Йерархично клъстеризиране б) Нейерархично клъстеризиране

Формираните групи (клъстери) трябва да бъдат еднородни (хомогенни) вътре и разнородни (хетерогенни) помежду си по зададени характеристики. Количественото оценяване на понятието „сходство” е свързано с понятието „метрика”. При този подход за сходство събитията се представят като точки в координатното пространство. Установените сходства и различия между точките се определят в съответствие с разстоянията между тях.

Два обекта са идентични, ако разстоянието между тях е нула. Колкото по-голямо е разстоянието между два обекта, толкова по-голямо е несходството между тях.

Нека таблицата от данни $\mathbf{X} = (x_{ij}),\ i = 1,...,n,\ j = 1,...p$ (1)

се записва във вида:

$$\mathbf{X} = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{pmatrix} = \left(x_{(1)}, x_{(2)}, \ldots, x_{(p)} \right),$$

където x_i' е векторът, съставен от i -тите съответни наблюдения за всяка променлива $x_{(j)}$, $x_{(j)}$ е вектор-стълб на j-тата променлива. Можем да групираме по редове (по наблюдения) или по стълбове (по променливи).

Основни изисквания за приложение на КА

За КА няма специални дискриминиращи изисквания към данните, техните разпределения и взаимозависимости. За сметка на това, не съществува механизъм за различаване на подходящите и неподходящите за изследването променливи. Ако някои съществени променливи бъдат игнорирани, резултатите могат да са неадекватни. Другият проблем е подходящият избор на оптималния брой клъстери, което е една неформализирана процедура.

При правилно определяне на решението, в случая на КА по променливи, много автори считат, че резултатите трябва формално да са много близки до тези от ФА за същите данни.

Мерки за сходство и различие в КА

Съществуват голям брой разстояния, за измерване на близостта (сходството), които се използват в КА:

1) Евклидово разстояние

Най-често използваната мярка за разстояние между интервални данни е обикновеното евклидово разстояние между два вектора $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ и $x_j = (x_{j1}, x_{j2}, \ldots, x_{jp})'$:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{p} (x_{im} - x_{jm})^2}, \quad i, j = 1, \ldots, n. \qquad (2)$$

2) Квадрат на евклидовото разстояние

$$d(x_i, x_j) = \sum_{m=1}^{p} (x_{im} - x_{jm})^2, \quad i, j = 1, \ldots, n. \qquad (3)$$

3) Чебишево разстояние

$$d(x_i, x_j) = \max \left| x_{im} - x_{jm} \right|, \quad i, j = 1, ..., n \,. \tag{4}$$

4) Разстояние на Минковски от ред *r*

$$d(x_i, x_j) = \left(\sum_{m=1}^{p} \left| x_{im} - x_{jm} \right|^r \right)^{\frac{1}{r}}, \quad i, j = 1, ..., n \,. \tag{5}$$

Така, в случай на n наблюдения трябва да пресметнем матрицата на разстоянията $D = d(x_i, x_j)$, която е симетрична и с нулеви елементи по главния диагонал.

Например, нека имаме три обекта (точки в двумерното пространство), с измервания за две променливи (x_1, x_2): (2, 5), (4, 2), (7, 9). Тук $p = 2, n = 3$. Използвайки обикновеното евклидово разстояние пресмятаме разстоянията: $d_{12} = \sqrt{(2-4)^2 + (5-2)^2} = \sqrt{13} \approx 3,6$, $d_{13} = \sqrt{(2-7)^2 + (5-9)^2} = \sqrt{41} \approx 6,4$ и $d_{23} = \sqrt{(4-7)^2 + (2-9)^2} = \sqrt{58} \approx 7,6$. За матрицата на разстоянията получаваме

$$D_1 = d(x_1, x_2) = \begin{pmatrix} 0 & 3,6 & 6,4 \\ 3,6 & 0 & 7,6 \\ 6,4 & 7,6 & 0 \end{pmatrix}.$$

Очевидно е обаче, че разстоянието зависи от мерната единица. Например, ако умножим първата променлива x_1 по 100 (да кажем смяна на метър със сантиметри), матрицата ще стане

$$D_2 = d(x_1, x_2) = \begin{pmatrix} 0 & 200 & 500 \\ 200 & 0 & 300 \\ 500 & 300 & 0 \end{pmatrix}.$$

Сега най-голямото разстояние е d_{13}, вместо предишното d_{23}.

Този проблем се решава обикновено чрез предварително стандартизиране на данните чрез z - трансформация или по друг обезмеряващ начин.

Може да се определят и други видове разстояния. Например, широко известна мярка се явява манхатъновото разстояние или „разстояние на градските квартали” (city-block) и др.

Формиране на клъстерите

След като се избере измерител на подобие (или различие) между обектите и се пресметне матрицата на разстоянията, се преминава към втория етап на КА - образуването на клъстерите. Множеството методи, които са разработени в статистическата литература за тази цел, се разделят на две групи: методи за йерархична (вертикална) клъстеризация и методи за нейерархична (хоризонтална) клъстеризация.

Методите за йерархична клъстеризация са разделени в две подгрупи: агломеративни и разделящи методи. При агломеративните методи се извършват последователни сливания на единици и клъстери. Започва се с n клъстера, представляващи отделните единици, и след последователни

сливания се стига до един клъстер, обединяващ всички единици. При разделящите методи подходът е обратен - започва се с един клъстер, който обединява всички единици, и след последователни разделяния се завършва до р клъстера, като всяка единица формира отделен клъстер. От разработените клъстерни методи най-често се използуват йерархичните агломеративни методи.

Най-разпространеният способ за представяне на резултатите на тези методи се явява дендрограмата (дървовидна диаграма), която графически изобразява йерархичната структура, породена от матрицата на сходство и правилата за обединение на клъстерите. Съществуват различни стратегии на обединение на обектите в клъстери и след това на самите клъстери. При

образуване на клъстери във вид на “верига” най-често се използват методите на междугруповото свързване (Between-groups linkage) и на най-близкия съсед (Nearest neighbor).

Ще опишем най-разпространените методи за групиране на клъстерите в йерархичните агломеративни методи.

1) Междугрупово свързване

Разстоянието между два клъстера A и B се дефинира като средната стойност на nA.nB на брой разстояния между nA точки от A и nB точки от B:

$$D(A,B) = \frac{1}{n_A.n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(x_i, x_j), \qquad (6)$$

където сумата се взима по всички x_i от А и всички x_j от В. Тук $d(x_i, x_j)$ е избраното разстояние между векторите x_i и x_j. На всяка стъпка се обединяват двата клъстера с най-малкото разстояние.

2) Вътрешногрупово свързване

Този метод е подобен на междугруповото свързване, но се изчисляват всевъзможните разстояния между всички точки на двата клъстера А и В, т.е. разстоянията между точките на един и същ клъстер. Формулата е:

$$D(A,B) = \frac{1}{(n_A + n_B).(n_A + n_B - 1)} \sum_{i,j} d(x_i, x_j), \qquad (7)$$

където сумата е върху всички точки x_i и x_j от А и В, а $d(x_i, x_j)$ е избраното

разстояние. На всяка стъпка се обединяват двата клъстера с най-малкото разстояние.

3) Най-близкия съсед

При този метод разстоянието между два клъстера A и B е равно на минимума от всички разстояния между точките на A и точките на B:

$$D(A,B) = \min\left\{d(x_i, x_j),\ x_i \in A,\ \ x_j \in B\right\}. \qquad (8)$$

4) Най-далечния съсед

Тук се пресмятат всички разстояния по формулата:

$$D(A,B) = \max\left\{d(x_i, x_j),\ x_i \in A,\ \ x_j \in B\right\}. \qquad (9)$$

Обединяват се в нов клъстер тези два клъстера, чието разстояние е минимално.

5) Центроиден метод

За всеки клъстер A се изчислява средният вектор $\bar{y}_A = \left(\sum_{i=1}^{n_A} y_i \right) / n_A$, където n_A е броят вектори (точки) на A. След това се изчисляват всички евклидови разстояния между средните вектори на всеки два клъстера A и B

$$D(A,B) = (\bar{y}_A, \bar{y}_B) . \qquad (10)$$

Обединяват се двата клъстера с минимално разстояние. Центроидът на новия клъстер AB се намира по формулата

$$\bar{y}_{AB} = \frac{n_A \bar{y}_A + n_B \bar{y}_B}{n_A + n_B} . \qquad (11)$$

Определяне броя на клъстерите

Това не е добре формализирана процедура. Първият начин е при разчитане на дендрограмата да се построят пресечни линии там, където разстоянието между един формиран клъстер и друг е по-голямо, например по-голям от поне 20% от максималната скала.

Друг начин е полученото решение да се сравни с решението от друг клъстеризационен метод и/или разстояние.

Валидация на резултатите от КА

Няма установени количествени оценки за валидация на резултатите. Препоръчват се следните методи:

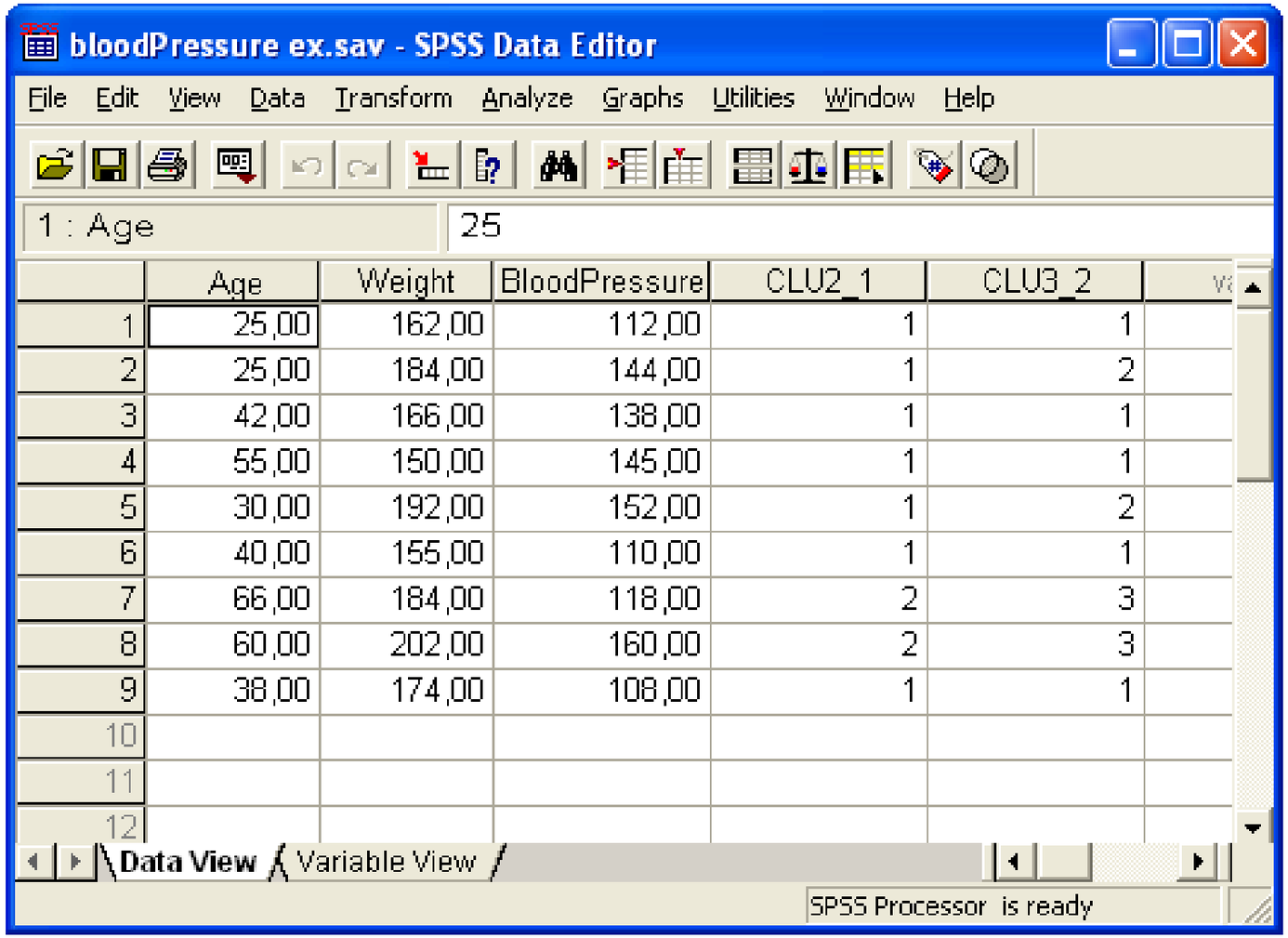
- проверка за стабилност на избрания клъстерен модел чрез сравняване с резултати от няколко метода
- “крос”-сравнения, което се проверява чрез случайно разделяне на извадката на две непресичащи се подмножества, провеждане на КА за всяко от тях и сравнение на моделите
- проверка за състоятелност на променлива. Когато някоя променлива (или наблюдение при КА на наблюденията) се класифицира в различни клъстери при различни методи, тя следва да се изключи от модела като несъстоятелна или да остане под наблюдение при следващи статистически анализи.

Как се прави КА със SPSS? ---> Повторете примерите в ЕВЛМ:

http://www.fmi-plovdiv.org/evlm/DBbg/database/studentbook/SPSS_CA_2.pdf
- Клъстерен анализ с SPSS: Клъстерен анализ на K-средните

http://www.fmi-plovdiv.org/evlm/DBbg/database/studentbook/SPSS_CA_3.pdf -
- **Клъстерен анализ с SPSS: йерархичен клъстерен анализ**

Пример 1. Данни за възраст, тегло и кръвно налягане на 9 пациента. Да се направи опит с КА за класифициране на пациентите.



	Age	Weight	BloodPressure	CLU2_1	CLU3_2
1	25,00	162,00	112,00	1	1
2	25,00	184,00	144,00	1	2
3	42,00	166,00	138,00	1	1
4	55,00	150,00	145,00	1	1
5	30,00	192,00	152,00	1	2
6	40,00	155,00	110,00	1	1
7	66,00	184,00	118,00	2	3
8	60,00	202,00	160,00	2	3
9	38,00	174,00	108,00	1	1
10					
11					
12					

Proximity Matrix

Case	Squared Euclidean Distance								
	1:Case 1	2:Case 2	3:Case 3	4:Case 4	5:Case 5	6:Case 6	7:Case 7	8:Case 8	9:Case 9
1:Case 1	,000	4,148	3,025	7,180	7,061	1,164	9,099	16,415	1,257
2:Case 2	4,148	,000	2,425	7,752	,480	6,638	9,126	7,114	4,325
3:Case 3	3,025	2,425	,000	1,705	3,334	2,379	4,606	6,875	2,537
4:Case 4	7,180	7,752	1,705	,000	8,641	4,149	6,135	9,497	6,591
5:Case 5	7,061	,480	3,334	8,641	,000	9,334	8,837	4,465	6,197
6:Case 6	1,164	6,638	2,379	4,149	9,334	,000	5,892	15,247	1,206
7:Case 7	9,099	9,126	4,606	6,135	8,837	5,892	,000	5,642	4,042
8:Case 8	16,415	7,114	6,875	9,497	4,465	15,247	5,642	,000	11,481
9:Case 9	1,257	4,325	2,537	6,591	6,197	1,206	4,042	11,481	,000

This is a dissimilarity matrix

Agglomeration Schedule

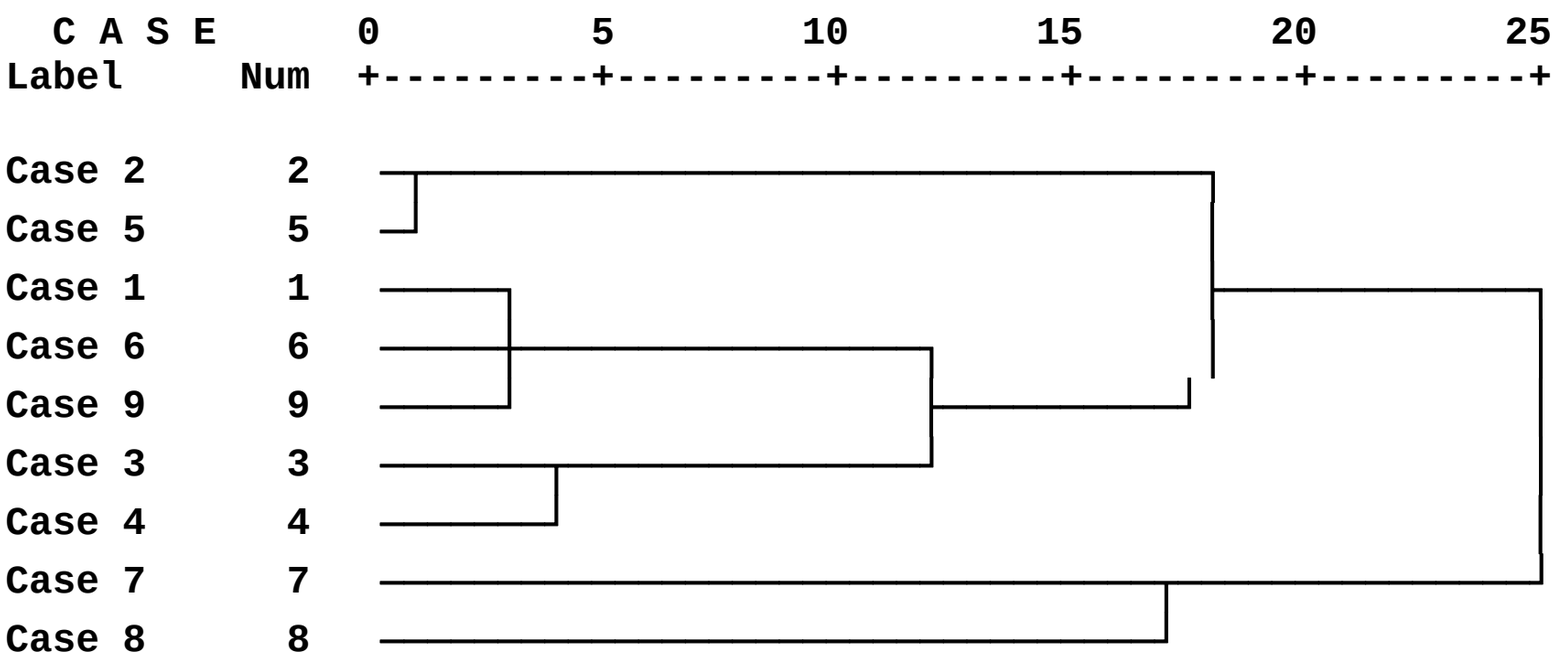
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	5	,480	0	0	7
2	1	6	1,164	0	0	3
3	1	9	1,231	2	0	5
4	3	4	1,705	0	0	5
5	1	3	4,310	3	4	7
6	7	8	5,642	0	0	8
7	1	2	5,985	5	1	8
8	1	7	8,488	7	6	0

Dendrogram

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine



Заключение:
Класифицирането в три клъстера дава 3 групи от хора:
1 клъстер – здрави (анкетирани 1, 3, 4, 6, 9)
2 клъстер – със средно здравословно състояние (аннкетирани 2, 5)
3 клъстер – голяма възраст (анкетирани 7, 8)

При 2 клъстера групите са 2:
1 клъстер – добро състояние (1, 2, 3, 4, 5, 6, 9)
2 клъстер – лошо състояние (7, 8)

Клъстеризирането в 3 групи е по-добро.

Пример 2. Приложение на клъстерен анализ за изследване основните характеристики на лазер с пари на меден бромид на базата на реални експериментални данни.

ОПИСАНИЕ НА ДАННИТЕ

Разглеждаме 7 основни характеристики на CuBr лазер:

Входни характеристики (независими променливи):

D– вътрешен диаметър на лазерната тръба,
dr– вътрешен диаметър на вътрешните пръстени, сложени в тръбата,
L – разстояние между електродите,
Pin – средна входна електрическа мощност,
PL – входна електрическа мощност на единица дължина, с 25% загуби,
PH2 – налягане на водорода от газовата смес.

Изходна характеристика (зависима променлива):

Eff– изходна лазерна ефективност

Използваме следната таблица с данни:

Табл. 1. Експериментални данни за CuBr лазер.

D, mm	dr, mm	L, cm	Pin, KW	PH2, Torr	PL, KW/cm	EFF, %
15	4.5	30	1	0.3	1.7	0.7
40	20	50	1.2	0.3	1.2	1.6
50	30	100	2.5	0.55	1.2	2.1
50	30	140	2.3	0.55	0.8	2
40	40	50	1.2	0.6	1.2	2
40	40	120	2.7	0.6	1.1	2.2
58	58	200	3.3	0.5	0.8	3

Ще проведем йерархичен КА по променливи

Клъстерният модел ще покаже групирането на променливите и структурата в групите. Така експериментаторът ще знае в какво съотношение на характеристиките е най-добре да планира бъдещ експеримент.

1. Набираме данните в SPSS.
2. Стартираме процедурата с: **Analize/ Classify/ Hierarchical cluster**
3. От основния прозорец внимаваме да маркираме **Variables**, вместо Cases.
4. В прозореца Statistics избираме **Proximity matrix** и **Cluster membership** от 2 до 5.
5. В Plots избираме **Dendrogram**.
6. В Method... избираме: **Between-groups Linkage**, distance: **Squared Euclidean distance** и Standardize: **Z-scores**.
7. В основния прозорец избираме OK.

Получаваме:

Табл.1. Матрица на разстоянията (Proximity matrix)
на независимите променливи.

Variable	D	dr	L	Pin	P_L	P_{H2}
D	**.000**	**2.330**	**2.402**	**2.845**	**23.004**	**5.123**
dr	**2.330**	**.000**	**2.584**	**2.921**	**21.417**	**3.638**
L	**2.402**	**2.584**	**.000**	**0.653**	**22.169**	**6.239**
Pin	**2.845**	**2.921**	**0.653**	**.000**	**20.800**	**5.310**
P_L	**23.004**	**21.417**	**22.169**	**20.800**	**.000**	**18.780**
P_{H2}	**5.123**	**3.638**	**6.239**	**5.310**	**18.780**	**.000**

Табл. 1 е началното състояние, когато всяка променлива се разглежда като отделен клъстер. Най-малко е разстоянието между L и Pin, което е 0.653. Затова тези променливи ще се групират най-напред в нов клъстер (виж по-надолу Фиг. 2).

Най-голямо е разстоянието между PL и D, равно е на 23.004. Изобщо PL има най-големи разстояния в сравнение с другите.

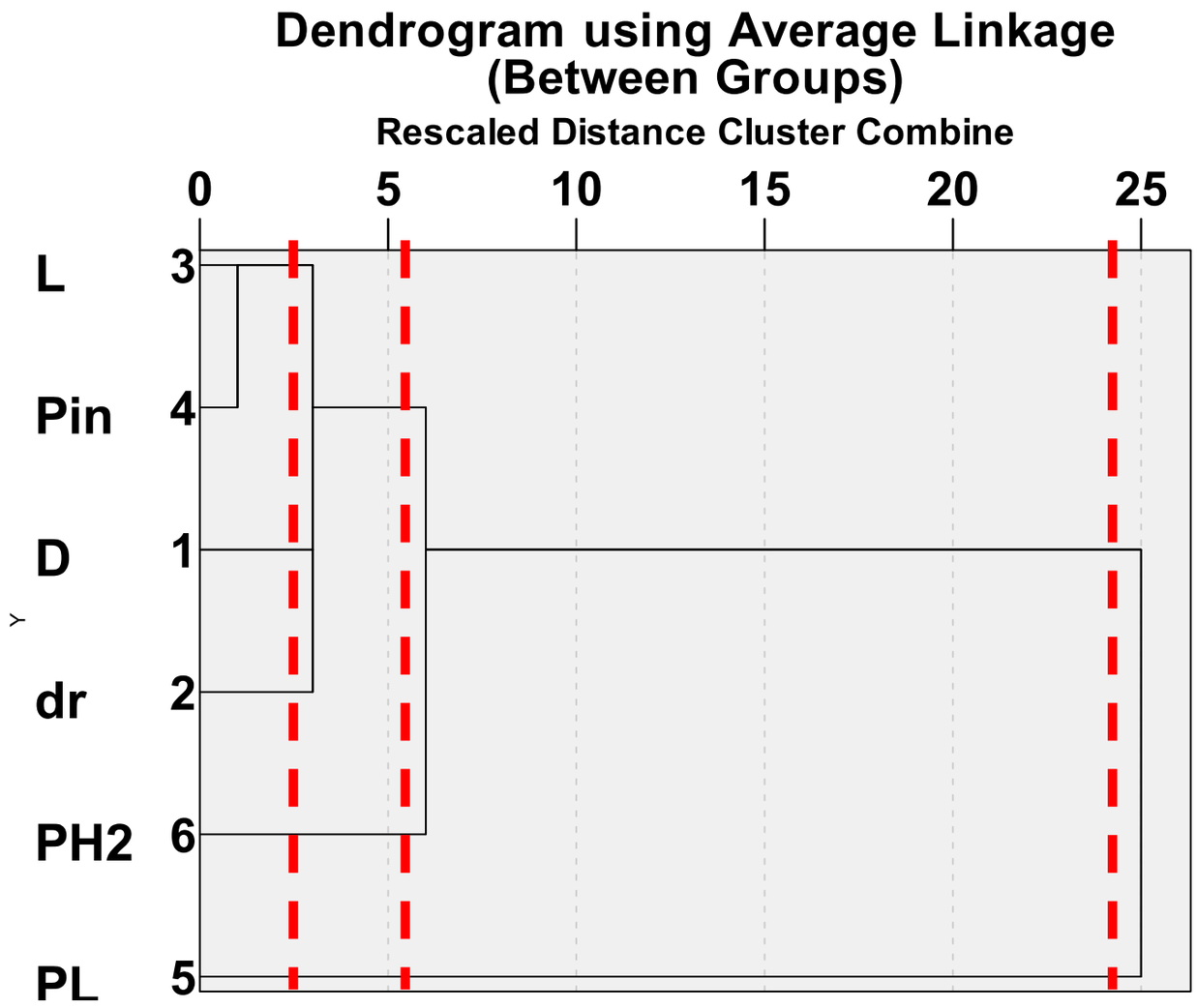
Крайното разпределение по клъстери с избрания метод на междугрупово свързане и квадрат на евклидовото разстояние е дадено на Табл. 2.

Табл. 2. Разпределение по клъстери.

Cluster Membership

Case	5 Clusters	4 Clusters	3 Clusters	2 Clusters
D	1	1	1	1
dr	2	1	1	1
L	3	2	1	1
Pin	3	2	1	1
PL	4	3	2	2
PH2	5	4	3	1

За да разберем кой е най-добрият клъстерен модел от дадените в Табл. 2, изследваме получената дендрограма:



Фиг. 2. Дендрограма на клъстерите с метода на междугруповото свързване и квадрат на евклидово разстояние.

За улеснение сме нарисували червени вертикални линии, които отразяват различните клъстерни решения. Броят на пресечените хоризонтални линии е броят на клъстерите за дадената вертикална права. Приема се, че голям отскок е при поне 5 относителни единици разстояние между 2 съседни линии. От дендрограмата се вижда, че най-голям отскок с почти 18 относителни единици имаме между най-дясната и средната червена линия. Оттук определяме, че най-доброто групиране за нашите данни е решението с 2 клъстера.

От табл. 2 това решение е:
1-ви клъстер= {L, Pin, D, dr, PH2}
2-ри клъстер={PL}

Променливата PH2 последна се групира към първия клъстер. При по-нататъшни изследвания с по-големи извадки се установява, че тя ще се отдалечи достатъчно, за да формира отделен клъстер.

Задача. Изследвайте близостта на 6-те променливи с изходната зависима променлива Eff, като определите в коя група попада с помощта на КА. Какви заключения можем да предложим?

Коментар. Разгледаните примери включват много малки по размер извадки и затова получените резултати не отразяват свойствата на генералната съвкупност, от която са взети. Трябва да се разглеждат само като учебни примери.