FÜSTÖS LÁSZLÓ—MESZÉNA GYÖRGY—SIMONNÉ MOSOLYGÓ NÓRA

Cluster analízis: fogalmak és módszerek.

1. Bevezetés

A számítógépes adatfeldolgozás elterjedésével gyakorlati lehetőség nyílt a különböző sokváltozós matematikai statisztikai módszerek széles körű alkalmazására az empirikus vizsgálatok eredményeinek értékelésénél. A klasszifikációs technikák – igen változatos körülmények között – a megfigyelt objektumok osztályokba sorolását teszik lehetővé. Az objektumokat a lehető legáltalánosabban értelmezhetjük, objektumok összességének tekintünk minden kvantitatív vagy kvalitatív jellemzőkkel definiált egyedekből álló rendszert. Az osztályok meghatározása tanulási folyamat eredménye, melynek két fő típusát különböztetjük meg:

1. Tanulás tanítóval,
2. Tanulás tanító nélkül (cluster analízis).

Az első esetben a gép kiértékelt tananyagot kap és – a megfelelő algoritmus segítségével – ennek az információnak az alapján végzi az osztályozást. A tanító nélküli tanulásnál az osztályokat kizárólag a tananyag (a minta) felhasználásával alakítják ki, valamilyen előre megadott osztályozási kritérium alapján. A döntési szabálynak a probléma szempontjából legfontosabb információkat kell tartalmaznia, az alakfelismerésnek ezt a lépését *lényegkiemelésnek* nevezzük, s ez magában foglalja:

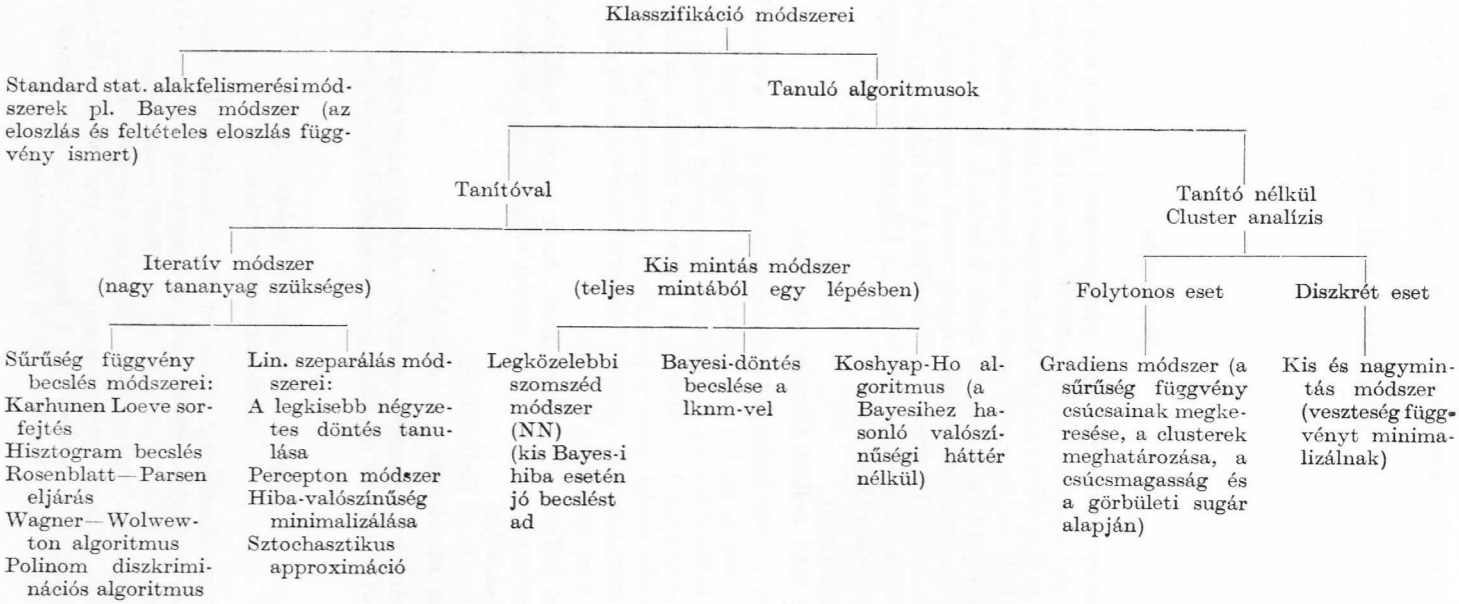
1. A lényeges jellemzők kiválasztását, azok mérési skálájának definiálását.
2. Az adatrendszer egységesítését, azaz a változók mérési skáláinak szükséges transzformációját.
3. A mértékrendszer definiálását.
4. A súlyozás problémájának megoldását.

A klasszifikációs módszerek elmélete eléggé szerteágazó (l. 1. táblázat), a továbbiakban a *cluster analízis* szempontjából lényeges aspektusokat tárgyaljuk.

Jellemzők kiválasztása, mérési skálák, skálatranszformációk

A jellemzők kiválasztása a vizsgálat szempontjából lényegesnek ítélt tulajdonságok számbavételét, ezen tulajdonságokhoz mérési skálával ellátott változók hozzárendelését jelenti. Bár tulajdonság lehet bármely ismérv, mennyiségi érték, minőségi állapot, földrajzi megjelölés stb. mégis a tulajdonságokhoz rendelt változó értékét tekintjük matematikai változónak. A változók típusától függenek elsősorban a kapcsolatok mérésének módszerei. Célszerű ezért a különböző típusok rövid áttekintése.

1. táblázat


Klasszifikáció módszerei
Standard stat. alakfelismerési módszerek pl. Bayes módszer (az eloszlás és feltételes eloszlás függvény ismert)
Tanuló algoritmusok
Tanítóval
Tanító nélkül
Cluster analízis
Iteratív módszer (nagy tananyag szükséges)
Kis mintás módszer (teljes mintából egy lépésben)
Folytonos eset
Diszkrét eset
Sűrűség függvény becslés módszerei:
Karhunen Loeve sorfejtés
Hisztogram becslés
Rosenblatt—Parsen eljárás
Wagner—Wolwewton algoritmus
Polinom diszkriminációs algoritmus
Lin. szeparálás módszerei:
A legkisebb négyzetes döntés tanulása
Percepton módszer
Hiba-valószínűség minimalizálása
Sztochasztikus approximáció
Legközelebbi szomszéd módszer (NN) (kis Bayes-i hiba esetén jó becslést ad
Bayesi-döntés becslése a lknm-vel
Koshyap-Ho algoritmus (a Bayesihez hasonló valószínűségi háttér nélkül)
Gradiens módszer (a sűrűség függvény csúcsainak megkeresése, a clusterek meghatározása, a csúcsmagasság és a görbületi sugár alapján)
Kis és nagymintás módszer (veszteség függvényt minimalizálnak)

Alapvetően két szempont alapján teszünk különbséget: az értékkészlet nagysága és a mérési mód szerint.

2. táblázat

A változók típusai

Mérési mód	Értékkészlet nagysága		
	folytonos	diszkrét	bináris
Nominális	—	születési hely	nő — férfi igaz — hamis
Ordinális	hangintenzítás, fényerősség	tanulmányi eredmény munkahelyi beosztás	kicsi — nagy jó — rossz
Intervallum	hőmérséklet C°-ban	jövedelem	feleség száma (0 vagy 1)
Arány	életkor	családonkénti gyermekek száma	két különböző egységpár

[1] L. [1]. 27. o.

A változók típusainak differenciált megkülönböztetésével és a különböző típusú változók közötti skálatranszformációk segítségével lehetővé vált a kevert változótípusok együttes kezelése. Az egységesítéshez alkalmazandó skálatranszformációt a változórendszer struktúrája határozza meg, általános alapelv az információveszteség minimalizálása.

2. A kapcsolatok mérésének módszerei

Legyen adva egy n elemű statisztikai sokaság, amelyet S-el jelölünk, S az osztályozandó objektumok véges, nem üres halmaza:

$$S = \{s_1 \dots s_n\}.$$

Adottnak tekintjük még a vizsgálat céljára kiválasztott tulajdonságok

$$T = \{X_1 \dots X_m\}$$

m-elemű halmazát.

Az osztályozás kiinduló adatbázisa az objektumokat és azok tulajdonságait tartalmazó T, $n \times m$-es adat-mátrix.

	T				
S	X_1	...	X_j	...	X_m
s_1	x_{11}	...	x_{1j}	...	x_{1m}
.					
.					
.					
s_i	x_{i1}	...	x_{ij}	...	x_{im}
.					
.					
.					
s_n	x_{n1}	...	x_{nj}	...	x_{nm},

ahol x_{ij} az i-edik objektum j-edik tulajdonságának megfigyelt vagy mért értéke. Az osztályozás tényleges input adata a T mátrixból a sorok vagy oszlopok páronkénti összehasonlításával keletkező szimmetrikus DC mátrix (dissimilarity coefficient). Az előbbi esetben a DC mátrix az objektumok közötti, az utóbbi esetben pedig a tulajdonságok közötti ún. taxonomikus távolsági vagy hasonlósági mérőszámokat tartalmazza. A továbbiakban a T → DC transzformáció kérdését vizsgáljuk. A kapcsolatok mérési módszerét egyrészt a T mátrix változóinak típusa határozza meg, másrészt pedig az, hogy az objektumokat vagy a tulajdonságokat kívánjuk összehasonlítani. Ez utóbbi esetre a matematikai statisztika számos jól használható mérőszámot dolgozott ki. Bizonyos esetekben úgy tűnhet, hogy az ilyen fajta feladatok nem is képezik az automatikus osztályozás feladatát. Azonban a botanikában, zoológiában, pszichológiában egyre elterjedtebben használják a cluster analízis módszereit az ismérvek osztályozására. Az ilyen típusú mérőszámok rövid áttekintése azért is indokolt, mert bizonyos esetekben – más értelmezéssel – ezeket is felhasználhatjuk az objektumok közötti kapcsolatok mérésére.

Az ismérvek közötti hasonlósági mutatók

A taxonomikus hasonlósági mérőszámok általános (de nem minden esetben érvényes) tulajdonságai az alábbi formában írhatók fel, ha s_i, s_j két tetszőleges összehasonlítandó objektum és $A(s_i, s_j)$ a hasonlósági mérőszám, akkor

1. $A(s_i, s_j) = A(s_j, s_i)$ (szimmetria),
2. A értéke általában a $0 \leq A \leq 1$ vagy a $-1 \leq A \leq 1$ intervallumba esik,
3. $A(s_i, s_i) = 1$.

A mérési módszereket a változók egyes típusaira külön-külön ismertetjük.

Nominális és ordinális változók

A mérés alapja a statisztikából ismert kontingencia tábla

A \ B	1	2	...	q	
1	f_{11}	f_{12}	...	f_{1q}	$f_{1.}$
2	f_{21}	f_{22}	...	f_{2q}	$f_{2.}$
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
r	f_{r1}	f_{r2}	...	f_{rq}	$f_{r.}$
	$f_{.1}$	$f_{.2}$	...	$f_{.q}$	n

,

ahol f_{ij} az i és j tulajdonság együttes előfordulásának – az n elemű mintából számított – gyakorisága.

A mérési módszerek jelentős része az ismert χ^2 statisztikára épül. A kontingencia táblázatból számítható és a változók közötti függetlenséget feltételező χ^2 formulából:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{q}\frac{f_{ij}^2}{f_{i.}f_{.j}} - 1\right)$$

látható, hogy χ^2 közvetlenül függ a tábla méretétől és n növekedésével minden határon túl nő. Ezért χ^2-nek különféle normált értékei jöhetnek számításba. Ilyen normalizáló faktor nyilvánvalóan az n, a kapott érték 0 és 1 közé esik. Ezt figyelembe véve javasolta *Pearson* a P, *Csuprov* a T, *Cramer* a C kontingencia együtthatót.

$$P = \left(\frac{\Phi^2}{1+\Phi^2}\right)^{1/2}, \quad \text{ahol} \quad \Phi^2 = \frac{\chi^2}{n};$$

$$T = \left(\frac{\chi^2}{n(r-1)(q-1)}\right)^{1/2};$$

$$C = \left(\frac{\chi^2}{n \cdot \min[(r-1),(q-1)]}\right)^{1/2},$$

Kendall és *Stuart* mutatott rá a χ^2 statisztikán alapuló mértékek torzító hatásainak okaira. Ezek a mértékek arra a hipotézisre épülnek, hogy a kontingencia tábla olyan kétváltozós normális eloszlást reprezentál, amelyre teljesül az alábbi összefüggés:

$$\lim_{n\to\infty} P^2 = r^2, \qquad \text{(ahol } r \text{ a korrelációs együttható).}$$

A gyakorlatban ez a feltevés általában nem jogos, így ezek a mértékek csak korlátozottan alkalmasak az asszociáció mérésére.

Másik hiányosságukra *Goodman* és *Kruskal* mutatott rá: (hiv. *Anderberg*, 740 o.) a változó párok egymás között nem összehasonlíthatók e mértékek alapján.

Ebből kiindulva javasolták a γ statisztika bevezetését; ez az asszociációs mérték az optimális osztály becslésén alapul.

Nominális változók esete

Ha a kontingencia tábla minden elemét n-el elosztjuk a kapott értékek relatív gyakoriságok, amelyek n növelésével jól közelítik a megfelelő valószínűségeket, indokolt tehát a következő jelölések bevezetése

$$p_{ij} = \frac{f_{ij}}{n}; \qquad p_{.j} = \frac{f_{.j}}{n}.$$

A következő valószínűségi modellt használjuk: válasszunk ki az n elemű sokaságból véletlenszerűen egy elemet. Becsüljük meg a lehető legkisebb hibával, hogy melyik A_i, ill. B_j ismérv-osztályba tartozik. A becslést két esetben végezzük el:

1. csak azt tudjuk a kiválasztott elemről, hogy besorolható valamelyik két osztályba;
2. ismerjük a kiválasztott elem A_i osztályát.

Nyilvánvaló, hogy az utóbbi esetben több információnk van; az elkövetett hiba legfeljebb akkora lehet, mint az első esetben.

Legyen P_1 a besorolás hibájának valószínűsége az 1. esetben
P_2 a 2. esetben
Ekkor a Γ_B asszociációs mérőszámot így definiáljuk:

$$\Gamma_B = \frac{P_1 - P_2}{P_1},$$

Γ_B a besorolási hiba valószínűségének azt a relatlv csökkenését mutatja, amely az A_i osztály ismeretéből származó információtöbbletből ered. Ha bevezetjük a $p_{.m} = \max_j p_{.j}$ és a $p_{im} = \max_j p_{ij}$ jelöléseket, akkor

$$P_1 = 1 - p_{.m}, \quad P_2 = 1 - \sum_i p_{im},$$

$$\Gamma_B = \frac{\sum_{i=1}^{r} p_{im} - p_{.m}}{1 - p_{.m}}.$$

Ha nem az A_i, hanem egy B_j osztály azonosítható a véletlenszerűen kiválasztott elemmel, akkor a Γ_A hasonlósági mérőszám a fentivel teljesen analóg módon definiálható:

$$\Gamma_A = \frac{\sum_{j=1}^{q} p_{mj} - p_{m.}}{1 - p_{m.}}.$$

A fenti gondolatmenetet megismételhetjük akkor is, ha az A és B ismérvváltozatok között a kapcsolatoknak nincs kitüntetett iránya. Tehát egy tetszőleges elem kiválasztásakor $\frac{1}{2}$ valószínűséggel az A_i vagy a B_j osztályba tudjuk sorolni. Ekkor ismeretlen prediktor osztály esetén a hiba valószínűsége

$$P_1 = 1 - \frac{1}{2}(p_{.m} + p_{m.}),$$

ismert prediktor osztály esetén pedig

$$P_2 = 1 - \frac{1}{2}\left(\sum_{i=1}^{r} p_{im} + \sum_{j=1}^{q} p_{mj}\right).$$

Az asszociációs mutató értéke:

$$\Gamma = \frac{\frac{1}{2}\left(\sum_i p_{im} + \sum_j p_{mj} - p_{m.} - p_{.m}\right)}{1 - \frac{1}{2}(p_{.m} - p_{m.})}.$$

Γ értékei a $\Gamma_A \leq \Gamma \leq \Gamma_B$ intervallumba esnek.

Az asszociációs mutató tulajdonságai

a) Γ akkor és csak akkor nem határozható meg, ha az egész sokaság egy osztályba tartozik; egyébként $0 \leq \Gamma \leq 1$,

b) $\Gamma = 1$ akkor és csak akkor, ha A_i ismerete egyértelműen definiálja a megfelelő B_j osztályt, azaz függvényszerű kapcsolat van a két változó között.

c) $\Gamma = 0$, ha a vizsgált osztályok statisztikailag függetlenek (nem megfordítható állítás).

d) Γ invariáns a kontingencia tábla sorainak (vagy oszlopainak) permutációjára.

A cluster analízisben a nominális változók közötti kapcsolatok jellemzésére jól használhatók még a

– kanonikus korreláció és az
– entrópia elméleten alapuló mértékek.

A γ statisztika ordinális változók esetén

Most az A és B ismérvváltozatok közül legalább az egyik természetes módon rendezhető. Így a kontingencia táblázat sorainak vagy oszlopainak permutációjára γ nem lehet invariáns. A valószínűségi modell: válasszunk ki a sokaságból véletlenszerűen (visszatevéssel) két elemet. Tegyük fel, hogy az első valamilyen $(A_{i1}; B_{j1})$, a második pedig valamilyen $(A_{i2}; B_{j2})$ kategóriába tartozik, ahol $1 \leq i_k \leq r$ és $1 \leq j_k \leq q$ $(k = 1, 2)$. Függetlenség esetén joggal várhatjuk, hogy az i_k indexek rendezettsége nincs összefüggésben a j_k indexek rendezettségével, míg kapcsolat esetén ez a rendezettség általában megegyezik.

Jelöljük a hasonló rendezettség valószínűségét P_h-val

$$P_h = P\{i_1 < i_2 \quad \text{és} \quad j_1 < j_2, \quad \text{vagy} \quad i_1 > i_2 \quad \text{és} \quad j_1 > j_2\},$$

az eltérő rendezettség valószínűségét P_e-vel

$$P_e = P\{i_1 < i_2 \quad \text{és} \quad j_1 > j_2, \quad \text{vagy} \quad i_1 > i_2 \quad \text{és} \quad j_1 < j_2\},$$

valamint az azonosság valószínűségét P_a-val

$$P_a = P\{i_1 = i_2, \quad \text{vagy} \quad j_1 = j_2\}.$$

Az egyértelműség kedvéért ez utóbbi esetet a statisztika definiálásakor nem engedjük meg, vagyis P_h és P_e helyett az $\{i_1 = i_2$ vagy $j_1 = j_2\}$ esemény negáltjára vonatkozó feltételes valószínűségeket tekintjük. Pl. P_h helyett a $P_h/(1 - P_a)$ valószínűséget.

Az asszociációs mutató:

$$\gamma = \frac{P_h - P_e}{1 - P_a}.$$

A γ mutató tulajdonságai

a) γ nem határozható meg, ha a kontingencia tábla nem nulla elemei egy sorban vagy egy oszlopban vannak.
b) $-1 \leq \gamma \leq 1$.
c) $\gamma = 1$, ha a nem nulla elemek a $p_{11} \rightarrow p_{rq}$ irányú átlóban vannak; ekkor $P_e = 0$;
d) $\gamma = -1$, ha a nem nulla elemek a $p_{r1} \rightarrow p_{rq}$ irányú átlóban vannak;
e) $\gamma = 0$, ha teljesül a függetlenség. Ez az állítás nem megfordítható (kivétel a 2×2-es tábla).

Arány és intervallum változók

A T mátrix elemei mérhető értékek, feladatunk két tetszőleges oszlop hasonlóságának a mérése. Jelöljük a T mátrix két oszlopát X és Y vektorokkal, $X,\ Y \in R^n$.

A hasonlóság mértékéül a vektorok hajlásszöge és a szorzatmomentum korrelációs együttható tekinthető.

$$A(X, Y) = \cos \alpha = \frac{X * Y}{|X| \, |Y|} .$$

Képezzük az $\hat{X} = X - \bar{X}$ és $\hat{Y} = Y - \bar{Y}$ nulla átlagú vektorokat. A szorzatmomentum korrelációs együtthatója

$$r = r(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \cdot \operatorname{var}(Y)}} ,$$

ahol $\operatorname{cov}(XY) = \dfrac{\hat{X}\,\hat{Y}}{n}$ és $\operatorname{var}(X) = \dfrac{X * X}{n}$.

Könnyen belátható, hogy $r(X, Y) = A(\hat{X}, \hat{Y})$.

$A(X, Y)$ invariáns a nyújtásra, $r(X, Y)$ pedig a nyújtásra és az eltolásra. Ebből adódik, hogy $A(X, Y)$ az arány, az $r(X, Y)$ pedig az intervallum változók esetén alkalmazható eredményesen.

Bináris változók

A bináris változók sajátos tulajdonsága miatt célszerű a külön kiemelés mert:

– az előző formuláknak bináris esetre általában létezik egyszerűbb alakja,
– a tulajdonságok asszociációs mérőszámai bizonyos esetekben, mint említettük alkalmazhatók objektumok összehasonlítására is, ez különösen bináris változókra áll fenn.

A T mátrix most csak a 0 és az 1 számokat tartalmazza. Két tetszőleges oszlop összehasonlítása nyilvánvalóan minden esetben redukálható egy 2×2-es táblára.

A \ B	1	0	
1	a	b	$a+b$
0	c	d	$c+d$
	$a+c$	$b+d$	$a+b+c+d=n$

Példaként a már bevezetett *Csuprov*-együttható bináris alakját mutatjuk be:

$$A_{es} = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} .$$

A bináris mérőszámok konstruálásánál a problémát egyrészt a d érték figyelembevétele jelenti, ez ugyanis a közös tulajdonságok hiányát méri, másrészt az, hogyan súlyozzuk az illeszkedéseket és nem illeszkedéseket.

A mutatókat a felvetett problémák szerint osztályozva a 3. táblázat foglalja össze.

3. táblázat

Súlyozás	0—0 Illeszkedés a nevezőben	0—0 Illeszkedés a számlálóban	
		nem szerepel	szerepel
Egyenlő súlyok	szerepel	1. *Russel* és *Rao* $\frac{a}{a+b+c+d} = \frac{a}{n}$	2. *Sokal* és *Michner* $\frac{a+d}{a+b+c+d} = \frac{a+d}{n}$
	nem szerepel	3. *Jaccard* $\frac{a}{a+b+c}$	4. —
Dupla súlyozás a kapcsolódó pároknál	szerepel	5. *nem ajánlott* $\frac{2a}{2(a+d)+b+c}$	6. $\frac{2a+d}{2(a+d)+b+c}$
	nem szerepel	7. *Dice* $\frac{2a}{2a+b+c}$	8. —
Dupla súlyozás nem kapcsolódó pároknál	szerepel	9. *nem ajánlott*	10. *Rogers—Tanimoto* $\frac{a+d}{a+d+2(b+c)}$
	nem szerepel	11. $\frac{a}{a+2(b+c)}$	12. —
A kapcsolódó párok kizárva a nevezőből	—	13. *Kulczynski* $\frac{a}{b+c}$	14. $\frac{c+d}{b+c}$

[2] L. [1]. 89. o.

3. Az objektumok közötti távolság, hasonlóság mértékei

Ebben a fejezetben a T mátrix sorainak páronkénti összehasonlításával foglalkozunk. Számos alkalmazási területen kizárólag így vetődik fel a kérdés. A tárgyalást a mérhető változókkal kezdjük. Az eddigiekhez képest alapvetően új eljárásokkal ezek esetében találkozunk, mert az objektumok között értelmezhető a taxonomikus távolság fogalma.

A taxonomikus távolság metrikus mértékei

A cluster analízis kvantitatív módszereinek gyakorlati alkalmazásánál az egyik központi probléma a pontok, ill. ponthalmazok közötti távolság definiálása. A távolság megfelelő megválasztása legalább olyan körültekintést igényel, mint az adekvát osztályozási algoritmus kiválasztása. Tételezzük fel, hogy a T mátrix elemei mérhető változók. Minden objektum egy pontnak tekinthető a p dimenziós absztrakt térben [4]. E pontok között értelmezhetők metrikus tulajdonsággal rendelkező távolságmérő függvények.

Jelöljük az (x, y) pontpár távolságát $d(x, y)$-al, amely minden $x, y, z \in M$ esetén, az alábbi tulajdonságokkal rendelkezik:

1. $d(x, y) = d(y, x)$,
2. $d(x, x) = 0$,
3. $d(x, y) > 0$, ha $x \neq y$,
4. $d(x, y) \leq d(x, z) + d(y, z)$.

Ezek a metrikus tér általános tulajdonságai és az ezeket kielégítő $d(x, y)$ függvényt metrikus függvénynek vagy röviden *metrikának* nevezzük. Ha a 3. feltétel nem teljesül, akkor d-t *pszeudo metrikának* nevezzük.

Érdemes megvizsgálni, hogy milyen zavarhoz vezet ha a 4. tulajdonság az ún. háromszögegyenlőtlenség nem teljesül. Ezt az esetet *szemimetrikának* nevezzük. Azon metrikákat, amelyek kielégítik a $d(x, y) \leq \max[d(x, z) + d(y, z)]$ egyenlőtlenséget *ultrametrikának* nevezzük.

Tegyük fel, hogy 5 pontunk van, az i-edik és a j-edik távolságát jelöljük d_{ij}-vel. Távolságaink legyenek a következők:

$$\begin{array}{lll} d_{12} = 2 & d_{23} = 10 & d_{15} = 1 \\ d_{13} = 10 & d_{24} = 10 & d_{25} = 100 \\ d_{14} = 10 & d_{34} = 2 & d_{35} = 1{,}5 \\ & & d_{45} = 100 \end{array}$$

Az első két oszlop alapján két jól elkülöníthető osztályt kapunk, az $S_1 = \{x_1, x_2\}$ és az $S_2 = \{x_3, x_4\}$ osztályokat, de hova soroljuk az x_5 pontot? Ha S_1-hez vesszük hozzá, akkor x_2-től távolabb lesz, mint x_3-tól, pedig az előbbivel egy osztályba tartozik. De ugyanilyen észszerűtlen S_2-be sorolni is, mert akkor x_1-hez lesz aránytalanul közelebb, mint x_4-hez. Marad még egy lehetőség, x_5 külön osztályt alkot. De ez sem kielégítő megoldás, mert x_1-től is és x_3-tól is kisebb távolságra van, mint a velük egy osztályba tartozó x_2, ill. x_4 pontok.

Ha ismerjük adatrendszerünk valószínűség eloszlását, jól alkalmazható a következő mérték.

Tegyük fel, hogy az $x_1 \ldots x_n$ pontokat egy p-dimenziós valószínűségi vektorváltozó értékeinek tekintjük és rögzített k mellett az $x_1^k \ldots x_n^k$ számok a megfelelő egydimenziós valószínűségi változó értékei.

Legyen D a p-dimenziós változó szórásmátrixa (kovariancia mátrixa),

$$D = \begin{pmatrix} D_{11} & D_{12} \ldots D_{1p} \\ D_{p1} \ldots & D_{pp} \end{pmatrix},$$

ahol $D_{ij} = M[(X_i - M(X_i))(X_j - M(X_j))]$;

D^{-1} legyen D inverze. Ekkor a következőképpen definiálhatunk pontjaink között egy metrikát:

$$d_{ij} = \sqrt{(x_i - x_j)^* S' D^{-1} S(x_i - x_j)},$$

ahol S diagonális elemeket tartalmazó súlymátrix.

Ha változóinkat nem akarjuk súlyozni, akkor S elhagyható a formulából. Ezt a metrikát akkor célszerű alkalmazni, ha ismerjük az eloszlást, ill. ha a minta elengedően nagy ahhoz, hogy D értékét kielégítő pontossággal becsüljük. Ebben az esetben a d_{ij} távolság nemcsak az x_i és x_j pontok koordinátáitól függ, hanem az összes többi ponttól is, ellentétben a következőkben ismertetendő metrikákkal. Az sem elhanyagolható, hogy figyelembe vettük a különböző változók kapcsolatát is. Ha a változók páronként korrelálatlanok, akkor $D_{ij} = 0$, ha $i \neq j$, és ennek megfelelően a D diagonális mátrix, D^{-1} pedig olyan diagonális mátrix, amelynek főátlójában az egyes változók szórásának reciproka áll, d_{ij} ekkor a következő egyszerűbb alakba írható

$$d_{ij} = [w_1(x_i^1 - x_j^1)^2 + \ldots + w_p(x_i^p - x_j^p)^2]^{1/2},$$

ahol a w_i-k tetszőleges nem negatív súlyok.

Az egyes változók súlyozása valamennyi metrikánál elvégzendő, de annak megítélése, hogy milyen súlyrendszert alkalmazzunk, elsősorban a kutató feladata. Meg kell jegyeznünk, hogy ha minden változót azonos súllyal akarunk figyelembe venni, akkor is el kell végezni a súlyozást; egyenlő súlyozáshoz akkor jutunk, ha minden változót a szórásával normálunk. A továbbiakban feltesszük, hogy változóink már normáltak.

A Minkowski-metrika és speciális esetei

Az egyik legáltalánosabb metrikaosztály, amely tetszőleges $1 \leq r < \infty$ érték mellett egy-egy metrikát ad, a következőképpen definiálható:

$$d_r(x, y) = \left(\sum_{i=1}^{p} | x_i - y_i |^r \right)^{1/r}.$$

A Minkowski metrika $r = 2$ esetben az ismert euklidesi metrikával azonos.

$$d_2(x, y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}.$$

$r = 1$ esetén a távolságmérő függvény a koordinátánkénti eltérések összegével egyenlő

$$d_1(x, y) = \sum_{i=1}^{p} | x_i - y_i | .$$

A cluster analízis során ezt a két esetet szokták alkalmazni. Sok esetben a metrika helyett a pontok közötti hasonlóságot definiáljuk. A hasonlóság is egy nemnegatív szám, de a metrikával ellentétben célszerű úgy megválasztani, hogy értékei nulla és egy közé essenek. h-val jelölve a hasonlóságot, megköveteljük, hogy tetszőleges x pontra $h(x, x) = 1$ legyen, azaz minden pont (objektum) saját magához hasonlítson a legjobban. Az ismertetendő módszerek szempontjából közömbös, hogy az adott objektumok között távolságot vagy hasonlóságot értelmezünk, csak arra kell ügyelni, hogy a minimális távolság maximális hasonlóságnak felel meg és fordítva.

Ha egy, a pontpárok távolságán értelmezett, monoton csökkenő függvényt adunk meg, amelynek értékei 0 és 1 közé esnek, akkor a metrikához egy hasonlóságot rendelünk hozzá.

A legegyszerűbb hasonlósági mérték az

$$R_{ij} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{[\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2]^{1/2}}$$

korrelációs együttható.

Mivel $-1 \leq R_{ij} \leq 1$, a megfelelő hasonlósági mértéket pl. az $R'_{ij} = (1 + R_{ij})/2$ egyenlőséggel definiálhatjuk.

Egy további lehetőség, ha a pontok távolságát a hozzájuk tartozó vektorok hajlásszögével mérjük. A hajlásszög koszinusza,

$$\cos(x, y) = \frac{\Sigma x_i y_i}{(\Sigma x_i^2 \cdot \Sigma y_i^2)^{1/2}} \qquad x, y \neq 0$$

tekinthető a két pont hasonlóságának, távolságuk pedig a

$$d_{ij} = \sqrt{1 - \cos^2(x, y)} \qquad x, y \neq 0$$

képlettel definiálható, d_{ij} ebben az esetben pszeudo-metrika lesz, vagyis a 3. feltétel itt nem teljesül; két pont távolsága akkor lesz nulla, ha a vektorok egy egyenesbe esnek, így $x \neq y$ pontok távolsága is lehet nulla.

Bináris változók esetében – az eddigieken kívül – más távolságot is használhatunk.

A kétdimenziós tábláknál szokásos jelölésekkel

$$d_{ij} = \frac{a + d}{a + b + c + d}$$

egy lehetséges metrika.

De mérhető a pontok távolsága a

$$d_{ij} = \frac{2a}{2a + b + c}$$

értékkel is.

Ez utóbbi metrikánál alaposan mérlegelni kell a clusterezés alkalmazási szempontjait, mert a dichotom változók nulla és egy értékeinek megválasztása gyakran esetleges (pl. a nemeknél férfi = 0, nő = 1 vagy fordítva), ez pedig azt jelenti, hogy a jelöléstől függően más lesz a pontok távolsága. Ha pl.

$$x = (1,0,0,0,0,0), \qquad y = (1,1,1,0,0,0),$$

akkor $a = 1$; $b = 0$, $c = 2$, $d = 3$ alapján: $d_{ij} = \frac{1}{2}$; ha most a változók értékeiben felcseréljük a nulla és egy jelölést, akkor:

$$x = (0,1,1,1,1,1), \qquad y = (0,0,0,1,1,1),$$

vagyis $a = 3$, $b = 2$, $c = 0$, $d = 1$ és ennek megfelelően $d_{ij} = \frac{3}{4}$.

A korábban tárgyalt metrikáknál ilyen probléma nincs, a d_{ij} távolság ott független a változók értékeinek számozásától.

A Minkowski metrika alkalmazásakor figyelemmel kell lenni arra, hogy a metrika a változók különböző tartalmát, dimenzióját nem változtatja meg, ezt a kapott távolsági érték interpretációjánál kell számba venni. A metrika a változók közötti függetlenséget tételezi fel, így előfordulhat, hogy a változók közötti kapcsolatok esetén egyfajta hatást többször veszünk figyelembe.

Nem metrikus mértékek

A nem metrikus mértékek egyik típusa az objektumok között relációkat definiál és ezek relációelméleti alapon történő feldolgozásával csoportosít. A másik típus tulajdonképpen feltételez valamilyen – az előzőektől eltérő – metrikát, amelyre támaszkodva az objektumok egy rendezését végzi, de a rendezésnek már nincsenek meg a metrikától megkövetelt tulajdonságai.

Intervallum változók. Calhoun-távolság

Ez a távolságfogalom a kérdéses két pont és a vonatkoztatási koordinátatengelyek iránya által meghatározott hiperfelületek közé eső többi pontra épül. Pl. az x_1 és x_2 pont közötti távolságot a bevonalazott területbe eső pontok segítségével határozhatjuk meg (1. ábra).

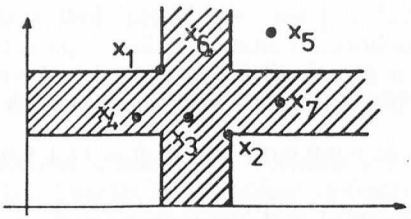
Két pont *Calhoun*-távolsága definíció szerint:

$$D_c = 6N_i + 3N_b + 2N_z,$$

ahol N_i – azon pontok száma, amelyek a két pont által meghatározott hipersíkba vagy meghosszabbításába esnek, legalább egy változójuk szerint.

N_b – azon pontok száma, amelyek egyetlen dimenzióban sem esnek a két pont közé, de egy vagy több változó szerint határra esnek.

[3] L. [1]. 111. o.



1. ábra

N_z – azon pontok száma, amelyek egy vagy több változó szerint mindkét ponttal azonos értékűek, de nem esnek a hipersík belsejébe vagy a határra.

Ha N az alappontok száma, akkor D_c maximális értéke: $6(N-2)$.

A Calhoun-távolság, mint mérték, nem felel meg a metrika követelményeinek, mert két pont távolsága akkor is lehet nulla, ha a két pont nem esik egybe. Ez a mérték hasznos lehet olyan esetekben, ha a clusterek egy vagy több változó szerint átfedik egymást.

Lance és Williams-féle mérték

Két objektum közötti távolság definíciója:

$$d_{LW} = \frac{\Sigma \mid x_i - y_i \mid}{\Sigma \, (x_i - y_i)} .$$

A számláló Minkowski metrika, $r = 1$ esetre, a nevező pedig a maximális kiterjedés mértéke. Bináris változó esetén a következő alakú:

$$d_{LW} = \frac{b + c}{(a + b)\,(a + c)} = 1 - \frac{2a}{2a + b + c} .$$

Ezek a mértékek ritkán használatosak, speciális eseteket elégítenek ki.

Nominális változók (l. [1] 123. o.)

Az ismérvek közötti hasonlóság mérésére alkalmazott mutatók az objektumok közötti hasonlóság mérésére is alkalmasak.

Ekkor az összehasonlítás azon alapul, hogy összeszámoljuk a közös és eltérő ismérveket. Ha az ismérvek jelenléte vagy hiánya egyértelműen megállapítható, akkor a bináris változóknál bevezetett 2×2-es táblához jutunk. Előfordulhat, hogy egyes ismérvek nem jellemzőek a kérdéses objektumra, ezt figyelembe véve a két lehetséges alternatíva (0 és 1) helyett hármat bevezetve, a bináris változók kiterjesztéséről beszélhetünk.

Legyen:

n_{a+d} azon ismérvek száma, amelyekben a két objektum megegyezik;

n_d azon ismérvek száma, amelyekkel a két objektum nem jellemezhető;
n_{b+c} azon ismérvek száma, amelyekben a két objektum eltér egymástól.

A 3. sz. táblázat képleteit felhasználva megkapjuk a 4. táblázatban összefoglalt kapcsolódási együtthatókat.

4. tábla (L. [1]. Table S. S.)

Kapcsolódási együtthatók nominális változók esetén

1.) $\dfrac{n_{a+d} - n_d}{n_{a+a} + n_{b+c}}$ 2.) $\dfrac{n_{a+d}}{n_{a+d} + n_{b+c}}$

3.) $\dfrac{n_{a+d} - n_d}{n_{a+d} - n_d + n_{b+c}}$

6.) $\dfrac{2n_{a+d}}{2n_{a+d} + n_{b+d}}$

7.) $\dfrac{2(n_{a+d} - n_d)}{2(n_{a+d} - n_d) + n_{b+c}}$

10.) $\dfrac{n_{a+d}}{n_{a+d} + 2n_{b+c}}$

11.) $\dfrac{n_{a+d} - n_d}{n_{a+d} - n_d + 2n_{b+c}}$

13.) $\dfrac{n_{a+d} - n_d}{n_{b+c}}$ 14.) $\dfrac{n_{a+d}}{n_{b+c}}$

(A formulák sorszáma a 3. sz. táblázat megfelelő számozására utal.)

Az ordinális változóknak kitüntetett szerepe van a cluster elemzésében, mert a legáltalánosabb osztályozási algoritmusok monoton invariánsak. Ebből következik, hogy ezek alkalmazása esetén elegendő a távolsági értékek helyett azok egymáshoz viszonyított rangsorát tekinteni.

A változók súlyozása

A T mátrix két oszlopának (két ismérv) összehasonlításakor két egyenként homogén koordinátájú vektorunk van, míg két sor (két objektum) egybevetésekor az egyes koordináták gyakran különböző tartalmú és típusú változókat reprezentálnak. Ebből adódik, hogy a mértékeket befolyásolják a nagyságrendek, és felmerül a különböző mértékek additivitásának problémája.

E problémák megoldására szolgál a súlyozás. Például, ha a vizsgálat körébe bevont változók nem egyformán fontosak az adott kérdés szempontjából: szubjektív súlyozást alkalmazunk.

Más jellegű a probléma, amikor a változók különböző dimenziójából adódó esetlegességet kívánjuk kiszűrni. A statisztikában leggyakrabban használt standardizálást itt is alkalmazhatjuk, ekkor minden változó azonos súlyú. Euklideszi távolság esetén ez a normalizálás $w = 1/s^2$ súlyozást jelent. Ez azzal a veszéllyel jár, hogy éppen azon ismérvek szerepét csökkentjük, amelyek a

legalkalmasabbak a csoportok megkülönböztetésére. Lényegesen elfogadhatóbb eredményre jutnánk, ha a teljes szórás helyett a csoporton belüli szórással normálnánk – a vizsgálat előtt ez persze nem ismert.

Az apriori ismeretek hiánya miatt bírálhatók azok a módszerek is, amelyek a súlyozással a korreláció hatását kívánják kikapcsolni, ilyen a *Mahalanobis* távolság-fogalom is.

Megfelelően szűri ki a korrelációnak a kapcsolatok mérését torzító hatását a faktoranalízis főfaktor módszere, ami azzal a további előnnyel is jár, hogy az eredeti adatmátrix mérete jelentősen csökken.

Meg kell jegyezni, hogy a súlyozás a különböző típusú változók együttes megjelenéséből adódó problémákat nem oldja meg, ehhez a megfelelő skálatranszformációt kell elvégezni.

Clusterek távolsága

A ponthalmazok között is többféle távolságot értelmezhetünk, itt azonban a metrika 1–4. követelményei közül általában csak a szimmetria teljesül. Legkézenfekvőbb megoldás, ha a ponthalmazok távolságán a legközelebbi, ill. a legtávolabbi pontjaik távolságát értjük.

1. $D(S_i, S_j) = \min_{m,k} d(x_{im}, x_{jk})$,

2. $D(S_i, S_j) = \max_{m,k} d(x_{im}, x_{jk})$.

Itt d a pontok között értelmezett tetszőleges távolság lehet.

A továbbiakban az i-edik és j-edik osztály távolságát D_{ij}-vel jelöljük, a most definiált távolságokat pedig $D_{\min}$, ill. $D_{\max}$ távolságoknak nevezzük. Szokásos az osztályok távolságának a centroidok távolságát tekinteni. Az n_i pontból álló S_i osztály centroidja a

$$C_i = \frac{1}{n_i} \sum_{k=1}^{n_i} x_{ik}$$

vektor, ahol x_{ik} az S_i osztály k-adik pontja. Ekkor az S_i és S_j osztályok távolsága

$$D_{ij}^c = d(C_i, C_j).$$

Az osztályok távolsága az egyes osztályokból vett pontpárok átlagos távolságaként is értelmezhető.

$$D_{ij}^r = \frac{1}{n_i n_j} \left[\sum_{m,k} \left(d(x_{im}, x_{jk}) \right)^r \right]^{\frac{1}{r}}.$$

r tetszőleges értékére az osztályok közötti távolság egy-egy lehetséges definícióját kapjuk. $r = 1$ esetben D^1 az előbbi távolsággal azonos. Ha $r \to \infty$, akkor $D^\infty \to D_{\max}$, míg ha $r \to -\infty$, akkor $D^{-\infty} \to D_{\min}$.

Az osztályozás során nemcsak két osztály távolságát kell meghatározni, hanem – különösképpen az összevonásos módszereknél – ismerni kell két egymással összevont osztálynak egy harmadiktól vett távolságát is. Az egyes összevonások után célszerű a távolságokat ezek felhasználásával számítani. A D^r távolságnál pl. ha az S_i és S_k osztályokat vonjuk össze S_{jk}-ba, akkor ennek egy S_i-től vett távolsága:

$$D^r(S_i, S_{jk}) = \left[\frac{n_j(D_{ij})^r + n_k(D_{ik})^r}{n_j + n_k}\right]^{\frac{1}{r}}.$$

Ha $r = \pm \infty$, a $D_{\min}$ és $D_{\max}$ távolságoknál ez a formula nem használható, helyette az alábbi összefüggést adjuk meg:

$$D(S_i, S_{jk}) = a\ D_{ij} + b\ D_{ik} + c\ D_{jk} + d|\ D_{ij} - D_{ik}|\ ,$$

ahol $a = b = \frac{1}{2}$, $c = 0$, $d = -\frac{1}{2}$, ha $D = D_{\min}$;

$a = b = \frac{1}{2}$, $c = 0$, $d = \frac{1}{2}$, ha $D = D_{\max}$;

$a = \frac{n_j}{n_j + n_k}$, $b = \frac{n_k}{n_j + n_k}$, $c = d = 0$, ha $D = D^1$.

4. A cluster analízis módszerei

A tudományos klasszifikációtól megköveteljük, hogy az identifikációt objektív kritériumok alapján lehessen elvégezni; ez a követelmény általában csak igen nagy megszorításokkal teljesül. Ennek egyik oka, hogy egy osztályozási sémáról önmagában nem lehet eldönteni, hogy jó-e vagy rossz, ez azon múlik, hogy a felosztás a vizsgálat szempontjából mennyire megfelelő. Ugyanakkor meghatározhatunk általános követelményeket.

1. Objektivitás
2. Stabilitás
3. Prediktivitás

Az első feltétel alatt azt értjük, hogy az adott kutatási terület szakemberei a vizsgált objektumokat jellegében azonos módon csoportosítsák. A második feltétel azt jelenti, hogy a klasszifikációtól megköveteljük, hogy új adatok kevéssé befolyásolják; ez akkor kerül előtérbe, amikor új adatok és eredmények a régi fogalmi struktúrát megkérdőjelezik és ezáltal a klasszifikációs rendszer instabillá válik. A harmadik feltétel az osztályozás igen magas kvalitását jelzi, s ennek megfelelően csak ritkán teljesül (pl. *Mendelejev* periódusos rendszere).

A klasszikus logikára épülő osztályozási modellek két alaplépése különböztethető meg:

– típus és koncepcióalkotás, kategóriák definiálása,
– az események kijelölése a már definiált kategóriákhoz.

A klasszikus logika azonban csak olyan kategóriák definiálását adja meg, amelyeknek minden egyede minden szempontból ekvivalens. Az ilyen elven alapuló osztályozást monotetikus osztályozásnak nevezzük. Ez a módszer sokváltozós, nagy méretű adatrendszer esetén még számológéppel sem kezelhető, de ha ez mégis sikerülne, az eredmények áttekinthetetlen felaprózottsága miatt a gyakorlatban használhatatlan lenne. Ezért nagy jelentőségű a cluster analízis, amely utat nyit a sokváltozós nagy minták áttekinthető numerikus értékeléséhez.

A cluster analízis modellje három szempontból is eltér a klasszikus osztályozási modellektől:

a) Nem definiál típusokat mielőtt kijelölné a mintaegyedeket; az eljárások során a típusokat definiáló fogalmak hozzárendelődnek az osztályozással kialakított csoportokhoz. Ez a meggondolás az alábbi feltételezéseken alapul:
   - léteznek típusok
   - a típusfogalom ismerete nélkül is létezik olyan kritérium, amelynek felhasználásával a clusterek felismerhetők;
   - a felismert clusterhez az egyedek ismérvei alapján megadhatók a típus-jellemzők.

Mindez szemléletesen azt jelenti, hogy az n-dimenziós térben az egyes típusokat elkülönítő hipersíkok akkor válnak láthatókká, ha az azonos clusterbe tartozó elemeket meghatároztuk, és ez azt jelenti, hogy a csak empirikusan előforduló típusok is felismerhetők.

b) A cluster analízis megengedi a politetikus osztályokat. Politetikusnak tekintünk egy osztályt, ha elemei több, de nem minden jellemző szerint ekvivalensek vagy hasonlóak. Az osztályhatárokat nem előre határozzuk meg. Ez a cluster analízisnek a gyakorlat szempontjából további előnyös tulajdonsága, hogy ugyanis minden jellemzőt figyelembe véve is jelentősen csökkenthető az osztályok száma.

c) A klasszikus modellek csak diszkrét változókkal dolgoznak, a cluster analízis megenged folytonos, sőt vegyes változó típusokat is.

Az osztályozás kritériumai

Az alábbiakban megadjuk az osztályozás gyakorlati követelményeit. A módszerek nem mindegyike teljesíti az összes feltételt egyszerre, de a konkrét értékelés alapján megítélhető a használhatóságuk.

1. *Egyértelműség*

Adott adathalmazból mindig ugyanazt az eredményt kapjuk egy adott M rendszer esetén.

2. *Monoton invariáns*

Ha az osztályozás végeredménye csak a DC-k (DC = az osztályozás input mátrixa) sorrendjétől (rangjaitól) függ, akkor a módszer monoton invariáns:

$$[Mf(d)][f(h)] = (Md)(h), \quad \forall\, h \geq 0\text{-ra},$$

ahol f a rangot előállító leképezés.

3. *Skála függetlenség*

Ha $k > 0$ skalár konstans, akkor ez a feltétel az

$$M(kd) = k\,M(d)$$

egyenlőség teljesülését jelenti.

4. *Stabilitás*

Az adatok kis változtatása az eredményben is kis változást jelentsen, vagyis az

$$M: C(S) \to U(S)$$

leképezés folytonos.

5. *A csoportok megőrzése (monotonitás)*

Legyen $h > 0$ rögzített és $S_h \subset (Md)(h)$ egy kialakult osztály a h szinten, akkor a monotonitás követelménye:

$$S_h \subset S_l \quad \forall\, l > h \text{ esetén.}$$

Ez más alakban is felírható. Jelölje:

$$d' \leq d, \text{ hogy } d'(A, B) \leq d(A, B) \qquad \forall\, A; B \in S.$$

Ekkor a monotonitás így írható:

$$M(d) \leq d.$$

6. *Optimalitás*

Az osztályozandó objektumok kapcsolatairól a legtöbb információt az input DC mátrix hordozza, ezt egy többlépéses transzformációnak vetjük alá, amíg kialakulnak az osztályok. Az eljárás folyamán információt vesztünk, az input és az output DC mátrix között legyen a lehető legkisebb az eltérés; azaz ha $d \in U(S)$ és $M(d) \leq d' \leq d$, akkor teljesüljön a $d' = M(d)$ egyenlőség.

Nehezen osztályozható elemek; reprezentatív elemek

A kvantitatív osztályozási elveknél általában számszerűen mérhető, hogy egy elem milyen mértékben (statisztikai elvek szerinti osztályozásnál milyen valószínűséggel) sorolható egyik vagy másik osztályba. Nehezen osztályozható egy elem, ha több osztályba is közel azonos mértékben sorolható. Ilyenkor két lehetőségünk van:

1. eljárásunkat instabilnak minősítjük és evvel összhangban megváltoztatjuk a klasszifikáció elvét, vagy
2. a nehezen besorolható elemet „zaj" elemnek tekintjük (mérési vagy kódolási pontatlanság).

Reprezentatív elemnek az olyan elemeket nevezzük, amelyeket viszonylag egyértelműen, illetve minimális kockázattal tudunk osztályozni. Ez a meghatározás a reprezentáció tetszőleges értelmezését magába foglalja, ha meg-

felelően definiáljuk az osztályozás kritériumait. Például a centroid módszereknél, ahol az osztályokat a hozzájuk tartozó objektumok súlypontjával jellemezzük, az egyes súlypontokhoz – centroidokhoz – közel eső objektumok lesznek a reprezentatív elemek, másszóval az adott osztály közelítőleg átlagos tulajdonságú objektumai. Természetesen az osztályozás szempontjain változtatva a reprezentativitás fogalma is más lesz.

A döntésfüggvények típusai

Döntésfüggvény, ill. döntési kritérium alatt azt az elvet értjük, amely szerint a vizsgálandó objektumokat rendezve az n-dimenziós térben az osztályok kialakulnak.

A döntésfüggvény mérheti:

– a clusteren belüli elemek hasonlóságát,
– a clusterek közötti különbséget.

Az eddigiekből megállapítható, hogy a hasonlóság és különbözőség fogalma többféleképpen is értelmezhető; a tárgyalt különféle mértékek két objektum közötti hasonlóság, ill. távolság mérésére alkalmasak. A döntésfüggvény feladata ennél összetettebb, egyszerre több egyed több jellemző szerinti hasonlóságát, ill. különbözőségét kell mérnie vagy becsülnie.

Az alkalmazott módszerek szerint megkülönböztethetünk:

– sűrűségfüggvény-becslésen alapuló eljárásokat,
– valószínűségeloszlások keverékének szétválasztásán alapuló eljárásokat
– „kevert modell",
– csoporton belüli variancia becslésén alapuló módszereket,
– csoportok közötti diszkriminancia becslésén alapuló és
– gráfelméleten alapuló eljárásokat.

A különböző döntési kritériumokhoz különböző cluster fogalmak kapcsolódnak.

5. Sűrűségfüggvény-becslésen alapuló eljárások

Tekintsük megfigyeléseinket egy r-dimenziós valószínűségi vektorváltozó realizációinak. A felvetődő kérdésekre a sűrűségfüggvény ismeretében válaszolhatunk. Az elméleti sűrűségfüggvény ismeretlen, de a minta alapján becsülhető (ha az elengendően nagy). Jelentős engedmény, hogy a cluster analízis által felvetett problémák megoldásához nincs szükség a sűrűségfüggvény alakjának teljes ismeretére – a típusalkotás célja olyan clusterek körülhatárolása, ahol a pontok koncentrációja viszonylag sűrű – mert a clusterek a sűrűségfüggvény csúcsait magukba foglaló tartományok.

Tehát elegendő a sűrűségfüggvény csúcsait, lokális maximum-helyeit megkeresni. A csúcsok adják a clusterek magját, a körülhatároláshoz a csúcsmagasság és csúcsponti fő görbület értéke használható. A sűrűségfüggvény ezen jellemzőit a sztochasztikus approximáció alapján a gradiens módszer sztochasztikus változata segítségével becsüljük.

Legyen $\xi_i \in F^r$ a megfigyelt valószínűségi vektorváltozó,
$f(x)$ az együttes sűrűségfüggvény.

A becsléshez, mivel a gradiens vektor is ismeretlen, annak egy becslését használjuk.

$$z_{n+1} = z_{n+1}(Y_n, \xi_1 \ldots \xi_n),$$

ahol $Y_n \in F^r$ valószínűségi változó.

Az algoritmus:

$$Y_{n+1} = Y_n + \gamma_n z_{n+1}.$$

Az Y_n kezdeti érték tetszőleges lehet, de a lépésnagyságnak ki kell elégítenie az alábbi feltételeket.

$$\sum_{n=1}^{\infty} \gamma_n = +\infty; \qquad \sum_{n=1}^{\infty} \gamma_n^2 < \infty.$$

A sztochasztikus approximáción alapuló eljárások jó becslést adnak, a konvergencia sebessége általában lassúbb, mint gradiens algoritmusoknál. Ennek oka, hogy a gradiens vektor ismeretlen és egy becslését használjuk. (Az állítás bizonyítását lásd [15]-ben.)

A sűrűségfüggvény becslésén alapul két fontos döntési kritérium:

a) a súlypontok módszere,
b) a sűrűségfüggvény csúcsainak lokalizálása.

a) *Súlypontok módszere*

Tegyük fel, hogy ismert a kívánt clusterek száma: s; a cluster-rendszert pedig jelölje: $C_1, C_2, \ldots C_s$. A sűrűségfüggvény olyan speciális függvény, hogy a C_i cluster egy M_i pontjával – a súlypontjával – azonosítható.

Ez teljesül, ha $f(x)$ a C_i összefüggő halmazokon konstans, rajtuk kívül eltűnik, és minden cluster átmérője kisebb a súlypontjához legközelebb eső más cluster súlypontjától vett távolságnál. Azaz, ha feltételezzük, hogy a clusterek diszjunkt rendszerén belül a sokaság egyenletes eloszlású. A clustereket elkülönítő kritérium az, hogy az egyenletes eloszlás jellemző paraméterei clusterenként eltérőek. E feltételeket kielégítő döntésfüggvény a következő ún. veszteségfüggvény:

$$J(M) = \sum_{i=1}^{s} \int_{C_i} || x - M^{(i)} ||^2 f(x)\, dx = \int_{E^k} \min_{1 \le i \le k} || x - M^{(i)} ||^2 f(x)\, dx.$$

Optimális a cluster-rendszer, ha $J(M)$ minimális. A veszteségfüggvény az egyes clustereken belül a súlypont és a többi pont közötti eltérés várható értékét becsüli.

A J függvény differenciálható, deriváltja a gradiens vektor, jelölje $U^{(i)}(M)$; $i = 1,2, \ldots s$.

$$U^{(i)}(M) = 2 \int_{E^r} \varepsilon^{(i)}(x, M)\,(M^{(i)} - x)\, f(x)\, dx,$$

ahol $\varepsilon^{(i)}(x, M) = \begin{cases} 1, \text{ ha } || x - M^{(i)} || < || x - M^{(j)} ||, \text{ ha } i \neq j, \\ 0, \text{ egyébként.} \end{cases}$

Ha adott az $\xi_1, \ldots, \xi_n \ldots$ minta, akkor

$$Z_{n+1}^{(i)} = 2\varepsilon^{(i)}(\xi_{n+1}, M)\,(M^{(i)} - \xi_{n+1})$$

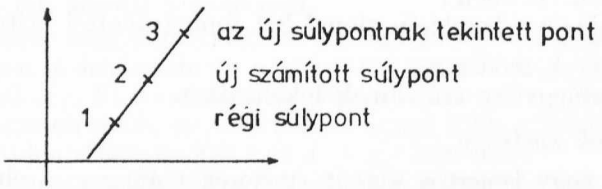
torzítatlan becslése $U^{(i)}(M)$-nek.

Tehát J minimumát $M_0 \in E_k$ kezdőérték mellett az

$$M_{n+1}^{(i)} = M_n^{(i)} - 2\gamma_n \varepsilon^{(i)}(\xi_{n+1}, M_n)(M_n^{(i)} - \xi_{n+1})$$

algoritmus adja, ahol $M_n = (M_n^{(1)}, M_n^{(2)} \ldots M_n^{(s)})$ az n-edik lépésben kialakított súlypontrendszer és $\gamma_n > 0$ eleget tesz a korábbi követelményeknek. Ezen a klasszikus súlypont módszeren alapul *Forgy* [6] konvergens nem hierarchikus módszere. Minden egyedet besorol a legközelebbi súlypontú clusterbe, majd az egész adatrendszer osztályozása után kiszámítja az új súlypontot és újra indítja az osztályozó algoritmust. Ha két egymásutáni ciklus cluster struktúrájában nincs változás, akkor megkaptuk a döntésfüggvény minimumát, az osztályozás optimális.

A konvergencia sebességének gyorsítása céljából a módszer több módosítása ismert. *Jancey* az előző ciklus súlypontjának az új súlypontra vonatkozó tükörképét tekinti az új lépés súlypontjának (2. ábra).



2. ábra

A módosítás alapgondolata, hogy az 1. pontból a 2.-ba húzott egyenes a veszteségfüggvényben az adott súlypont melletti gradiens becslése. Ha ebbe az irányba mozdul el a súlypont, akkor csökkenthető a legjobban a veszteségfüggvény értéke.

Egy másik módszert *MacQueen* [24] javasolt a konvergencia gyorsítására.

$$M_{n+1}^{(i)} = M_n^{(i)} + \gamma_n^{(i)} \varepsilon^{(i)}(\xi_{n+1}, M_n)(M_n^{(i)} - \xi_{n+1}),$$

ahol $\gamma_n^{(i)} = (1 + \omega_n^{(i)})^{-1}$; $M_0 \in E_k$

$$\omega_0^{(i)} > 0; \quad \omega_{n+1}^{(i)} = \omega_n^{(i)} + \varepsilon^{(i)}(\xi_{n+1}, M_n).$$

Teljes indukcióval belátható, hogy

$$M_{N+1}^{(i)} = \frac{\omega_0^{(i)} M_0^{(i)} + \sum_{n=0}^{N} \varepsilon^{(i)}(\xi_{n+1}, M_n)\, \xi_{n+1}}{\omega_0^{(i)} + \sum_{n=0}^{N} \varepsilon^{(i)}(\xi_{n+1}, M_n)}.$$

Ha $\omega_0^{(i)} = 1$, akkor $M_{N+1}^{(i)}$ éppen az $M_0^{(i)}$ kezdőpont és a korábban a C_i osztályba sorolt tanulópontok számtani átlaga súlypontja. Az eljárás minden elem clusteresítése után módosítja a súlypontot. Az eljárás *MacQueen*-féle közép módszerként ismert.

A súlypont-módszerek alkalmazásánál problémát jelent, hogy az induló súlypont-rendszert (magpontokat) előre kell megadni. Ennek feloldására születtek meg a súlypont-módszer olyan változatai, ahol a clusterszám előzetes megadása helyett elegendő bizonyos paraméterek megadása (pl. a clusterek maximális átmérője vagy elemszáma); a clusterek számát maga az algoritmus határozza meg. A súlypont módszer ilyen változata a MacQueen által kidolgozott paraméter becsléssel működő algoritmus, létezik ennek *Wishart* [42] által kidolgozott optimalizáló módosítása, valamint ismertek *Ball* és *Hall* ISODATA néven ismert módszerei [4].

b) *Sűrűségfüggvény csúcsainak lokalizálása*

A súlypont módszernél általánosabb feltételek mellett keresi az optimális rendszert. Míg a súlypont módszer alaphipotézise szerint az $f(x)$ sűrűségfüggvény a cluster egy pontjával jellemezhető és így egy egyszerű szerkezetű veszteségfüggvény definiálható, addig ez az algoritmus a clustereket a sűrűségfüggvény csúcsainak, lokális maximum-helyeinek környezeteként értelmezi.

Legyen $\xi_1, \xi_2 \ldots \xi_n \ldots$ független, azonos eloszlású r-dimenziós vektorváltozók sorozata. Az $f(x)$ sűrűségfüggvény differenciálható és gradiens vektora $U(x)$ egyenletes Lipschitz feltételnek tesz eleget, azaz létezik olyan L konstans, hogy

$$|| U(x) - U(y) || \leq L || x - y || \text{ és}$$

$$|| U(x) || \leq M(1 - || x ||), \qquad \text{ahol } M < \infty.$$

$f(x)$ maximum helyeinek megkeresésére az

$$Y_{n+1} = Y_n + \gamma_n Z_{n+1}, \qquad Y_0 \in E^r$$

gondolatmenet használható, ahol

$$Z_{n+1} = U(Y_n) + \lambda_{n+1}.$$

A számítógépes algoritmust *Wishart* dolgozta ki e módszer alapján [43].

6. „Kevert modell"

Nagy minta esetén jól alkalmazható modelltípus. A hipotézis az egyes csoportokra azonos típusú valószínűségeloszlást tételez fel, az egyes csoportok a paraméterekben különbözhetnek. Az algoritmus az eloszlás típusát ismertnek tételezi fel, és az optimális cluster struktúrát az eloszlások paramétereinek becslése alapján határozza meg. A többváltozós normális eloszlás feltételezése mellett *Wolfe* dolgozott ki számológépes algoritmust (NORMIX; NORMAP).

7. Csoporton belüli variancia becslésén alapuló eljárások

Az eljárás célja olyan osztályok létrehozása, amely rendszer a csoportokon belüli variancia összsokaságbeli összegét minimalizálja.

A többváltozós variancia-elemzés lineáris modellje a következő azonosságból indul ki:

$$x_{ij} = m + (m_j - m) + (x_{ij} - m_j),$$

ahol x_{ij} az i-edik megfigyelés a j-edik csoportban,
m a teljes mintaátlag,
m_j a j-edik csoportátlag.
Átrendezve a fenti formulát

$$x_{ij} - m = (m_j - m) + (x_{ij} - m_j),$$

vagyis az i-edik egyednek a j-edik csoportban a teljes átlagtól való eltérése két részre bontható: a csoportátlagnak a mintaátlagtól való eltérésére és az x_{ij} eltérésére a csoportátlagtól.

Ennek alapján a clusterezés céljára a többváltozós variancia elemzés alapegyenlete a következő:

$$\underbrace{\sum_{j}^{s}\sum_{i}^{n_j}(x_{ij}-m)(x_{ij}-m)'}_{T} = \underbrace{\sum_{j}\sum_{i}(m_j-m)(m_j-m)'}_{K} + \underbrace{\sum_{j}\sum_{i}(x_{ij}-m_j)(x_{ij}-m_j)'}_{B}$$

$$T = K + B$$

$$i = 1 \ldots n_j;\ j = 1, \ldots s,$$

ahol n_j a j-edik csoport elemeinek száma,
s a csoportok száma.

A cluster analízis célja a csoporton belüli homogenitás növelése, azaz a csoporton belüli variancia minimalizálása, ill. a csoportok közötti variancia maximalizálása. Ez a cél többféle típusú döntésfüggvény segítségével valósítható meg.

a) A B mátrix nyomának minimalizálása. Ez a kritérium a csoporton belüli átlagtól mért eltérések négyzetösszegének minimalizálását jelenti. A hierarchikus módszerek között mind agglomeratív, mind divizív algoritmus készült a $tr[B] \rightarrow$ min döntésfüggvény alapján (medián módszer, *Ward*-módszer).

b) A hiba minimalizálása a teljes sokaság varianciájához viszonyítva a *Wilks*-féle Λ döntésfüggvény alapján,

$$\Lambda = \frac{|B|}{|T|}.$$

Az algoritmus olyan cluster struktúrát alakít ki, amelynél Λ értéke minimális.

8. Diszkriminancia analízisen alapuló eljárások

A diszkriminancia analízis lényege, hogy az eredeti ponthalmazt a diszkriminancia függvény segítségével egy olyan térre vetítjük, ahol a pontokból kialakítható csoportok a legjobban elkülönülnek, azaz a pontok közötti diszhomogenitás a lehető legnagyobb. Ez a $\frac{|K|}{|B|}$ hányados maximalizálását

jelenti. A többváltozós variancia elemzés eredményeinek felhasználásával a diszkrimináns függvény:

$$\lambda = \frac{v' Kv}{v' Bv} \to \max!$$

Átalakítva a fenti egyenlőséget

$$(B^{-1} K - \lambda E) v = 0$$

alakúra a λ és v egyszerűen meghatározható. λ a $B^{-1}K$ mátrix sajátértéke, v pedig a sajátvektora. A különböző sajátértékek száma határozza meg a diszkriminációs függvények számát.

A diszkriminációs hatás próbájára is a *Wilks*-féle Λ-t használják, ami ekvivalens az alábbival:

$$\Lambda = \prod_{j=1}^{m} \frac{1}{1 + \lambda_j}.$$

A hierarchikus módszerek közül *Casetti*, *Hung* és *Dubes* módszere épül a diszkriminációs döntésfüggvényre.

9. Gráfelméleti alapokon álló eljárások

A hierarchikus módszerek két típusa gráfelméleti probléma megoldásán alapul.

a) *Egyszerű lánc-módszerek* (single linkage)

Legyen adott az $x_1 \ldots x_n$ pontrendszer az X absztrakt térben, és egy d metrika.

Állítsuk elő az adott pontok által kifeszített minimális fát. A fa konstruálásához a következő algoritmust alkalmazzuk:

- sorbarendezzük a $d(x_i, x_j)$ távolságokat és minden lépésben két pontot összekötünk, amely az alábbi két feltételnek tesz eleget:
- eddig még nem köti össze él a két pontot,
- az összekötött pontokon keresztül nem juthatunk el x_i-ből x_j-be.

Az előbbi két feltételt kielégítő pontpárok közül x_i és x_j távolsága a legkisebb. Az eljárás eredményeként kapott gráf összefüggő részei az egyes osztályok. Az algoritmus során a clusterek száma fokozatosan csökken, végeredményként megkapjuk az adatrendszer által kifeszített minimális gráfot.

A gráf sok értékes információt ad az adatrendszerről, de nagy elemszám esetén nehezen áttekinthető, ezért születtek meg a minimális fát újra felosztó eljárások. A clusterek számát ekkor előre meg kell adni s döntésfüggvény lehet a következő:

$$u = \frac{1}{s} \sum_{j=1}^{s} \bar{d}(j) \to \min,$$

ahol s a clusterek (összefüggő részgráfok) száma,
$\bar{d}(j)$ a j-edik részgráf éleinek átlagos hossza.

b) *Teljes láncmódszerek* (complete linkage)

Két cluster között a hasonlóságot a clusterek legtávolabbi elemei közötti távolsággal mérik, a clusterek közötti összes lehetséges párosításhoz kiszámítják ezt az értéket, és ahol ez minimális, azt a két osztályt vonja össze az algoritmus. Az eljárás lényegében a kialakuló clusterek „átmérőjét" minimalizálja.

10. Osztályozási módszerek. Cluster technikák

A döntési kritériumok sokfélesége, alaphipotéziseikben is lényegesen különböző típusai jelzik a módszertan változatosságát; tovább bővítik ezt a kört a különböző speciális szempontokat kielégítő cluster technikák. Az eljárások különbözősége azonos döntésfüggvény mellett is tovább differenciál (l. 5. táblázat).

5. táblázat

Cluster technikák	Egyszintű módszerek	Egyszintű felosztó módszerek	Objektív módszerek	6. tábla
			Szubjektív módszerek	7. tábla
		Egyszintű optimalizáló módszerek	Iteratív módszerek	8. tábla
			A hierarchia optimális felosztását végző módszerek	Minimálgráfot osztó algoritmus (PAGE)
	Hierarchikus módszerek	Hierarchikus divizív módszerek	Monotetikus módszerek	9. tábla
			Politetikus módszerek	
		Hierarchikus agglomeratív módszerek	Feltételes optimumot kereső eljárások	10. tábla
			A hierarchia adta keretek közt sem optimalizáló eljárások	Centroid módszer nagy adatrendszerekre (WOLFE)

6. táblázat

Objektív módszerek	*Outlierek = egyedülálló súlypontok*	Átlagos lánc Súlypont-módszer *k*-közép módszer (MCQUEEN)
	Egyszerű lánc	Elemi cl. analízis (MCQUITTY)
	Teljes lánc	Legtávolabbi szomszéd módszer "Rank order typal" elemzés (MCQUITTY)

7. táblázat

Szubjektív módszerek	*Egyszerű lánc*	Elemi cl. Paraméter: min. hasonlósági szint Adat átfedő clusterek Elemi cl. analízis (SOKAL és SNEATH)			
		Cluster magot kialakító eljárások	Paraméter: min. hasonlósági szint. Lehetnek átfedő cl.-ek A cl.-mag kialakítása variancia analízisen alapul	Objektív	Max: csoportok közötti diszhomogenitás (ZUBIN, FLEISS, BURDOOK)
					Max: csoporton belüli homogenitás (SNEATH)
				Szubjektív	Max: csoportok közötti diszhomogenitás (BAILEY)
					Max: csoporton belüli homogenitás (SAWREY, KELLER, CONGER)
			Paraméterek: min. hasonlóság és sűrűség Outlierek száma k-tól függ (WISHART)		
	Teljes lánc	A legtávolabbi szomszéd módszere Paraméter: min. hasonlósági szint (SRENSEN)			
	Átlagos lánc	Szakaszonként egy egyed clusteresítését engedélyezi Súlyozza a változókat (SOKAL és MICHENER)			
		Szakaszonként több egyed összevonását is engedélyezi Csoport-pár módszer (SOKAL és SNEATH) Egyenlőtlen súlyozással (LANCE és WILLIAMS)			

8. tálbázat

Iteratív módszerek — Egyszerű cl.-mag. — Politetikus cl.-ek — Természetes cl.-ek — Outlierek megengedettek — Tárolt hasonlósági mátrix — Súlypont-módszerek	*Adott a clusterek száma*	Fix magpontok (RORGY, JANCEY)
		Mozgó magpontok, konvergens k-közép módszer (MCQUEEN)
	A clusterek száma az eljárás során alakul ki	k-közép módszer (MCQUENN)
		A k-közép módszer egy változata (WISHART)
		„Isodata" (BALL és HALL)

A cluster technikák többféle szempont alapján osztályozhatók, leggyakoribb elv a következő:

a) átfedéses osztályozás,
b) diszjunkt osztályozás.

9. táblázat

Monotetikus módszerek (Outlierek nagyszámban lehetségesek)	Felosztás egyetlen, — a leginkább elkülönülő —, változó szerint (LAMBERT és WILLIAMS)
	A jellemzőket dichotomizálja, úgy hogy a max. információ-veszteség feltétele teljesüljön (LANCE és WILLIAMS)
	A csoporton belüli négyzetes hibaösszeget minimalizálja (SONQUIST és MORGAN)
Politetikus módszerek (A outliereket nem választja külön)	Mesterséges clusterek, diszkriminancia analízis (MAYER)
	Természetes clusterek, diszkriminancia analízis (CASETTI, HUNG, DUBES)
	Természetes clusterek, variancia analízis, nyomkritérium (EDWARD és CAVALLI—SFORZA)

10. táblázat

Feltételes optimumot kereső eljárások Tárolt hasonlósági mátrixszal	*Egyszerű lánc*		Legközelebbi szomszéd módszer Hierarchikus cl. analízis (MCQUITTY)
	Teljes lánc		Legtávolabbi szomszéd módszer (SAUNDERS, SCHUEMAN)
	Átlagos lánc	Egyedek páronkénti hasonlósága	Távolsági mérték min. a csoporton belüli (ORLÓCI) Korrelációs mérték max. a csop. belül (ORLÓCI) HOLZINGER-féle B-koeff. (TYRON)
		Centroid módszerek	WARD LANCE és WILLIAMS SOKAL és SNEATH Medián módszer (GOWER)
Tárolt adatrendszerrel (*átlagos lánc*)	WARD-módszer		
	Csoporton belüli variancia min.		
	Csoporton belüli eltérés négyzetösszegének min.		
	Centroid módszerek		
	Minden adat külső tárolón		Rendszerezett hasonlósági mátrix, egyszerű lánc, legközelebbi szomszéd módszer
			Módosított centroid módszer (PARK)

Az átfedéses osztályozás gyakorlati jelentősége kisebb, elmélete is kevésbé kidolgozott, relációelméleti alapon közelíthető meg.

Részletesen a diszjunkt átfedéseket nem tartalmazó osztályozási módszerekkel foglalkozunk, ezt gyakorlati jelentőségük és elméleti kidolgozottságuk indokolja. Az ilyen osztályozásnak két nagy csoportja van:

1. Nem hierarchikus osztályozás: az alapsokaságot k számú osztályra bontja.
2. Hierarchikus osztályozás: kezdetben minden elemet külön osztálynak tekint, majd az osztályok összevonásával lépésről-lépésre újabb osztályozási szinteket alakít ki, mindaddig, amíg az összes elem egyetlen osztályba nem kerül. (Lehetséges természetesen e gondolatmenet fordított irányú alkalmazása is.)

Általános tulajdonságok

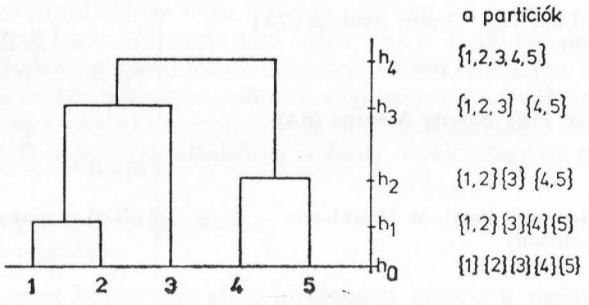
Az osztályozási algoritmus input adatának, a DC mátrixnak az előállítási módjaival a korábbi fejezetekben foglalkoztunk.

Az osztályozási algoritmus outputja az S halmaz diszjunkt partícióinak egy véges sorozata. Olyan fa struktúrával ábrázolható, ahol a fa csomópontjaihoz

$$h \in [0, \max d(A, B)]$$

értékek tartoznak.

Az elmondottakat az alábbi egyszerű példával illusztráljuk, ahol az egész számok objektumokat reprezentálnak (3. ábra).



3. ábra

S-nek egy adott partíciója a fa struktúra egy h szintjéhez tartozik, a sorozat első eleme az izolált pontok halmaza, az utolsó pedig az összes objektumokból álló halmaz. Az ilyen típusú fa-struktúrát *dendogramnak* nevezzük, és a következőképpen definiáljuk.

Jelölje $E(S)$ az S-halmazon értelmezett ekvivalencia relációkat, amelyek egyértelműen meghatározzák a halmaz diszjunkt partícióit, az ekvivalencia osztályokat.

A dendogram olyan

$$C: [0, \infty \to E(S)]$$

leképezés, amely az alábbi tulajdonságokkal rendelkezik:

1. Monotonitás: $C(h) \subseteq C(h')$, ha $0 \leq h \leq h'$.
2. Létezik a két triviális partíció.

3. A partíciók sorozata jól definiált, azaz adott $h > 0$-hoz létezik olyan $\delta > 0$, hogy

$$C(k + \delta) = x(h).$$

Minden objektumpárhoz azt a szintet rendeljük, ahol először egyesültek a dendogramban. Egy adott h szinten azok az objektumok vannak relációban, amelyek között a távolság kisebb vagy egyenlő mint h.

A kiinduló DC mátrix az iterációs lépések során mindig megváltozik, és az S elemei közötti távolságot minden iterációban újra kell számolni.

Ha az i-edik és j-edik csoport elemeinek páronkénti távolságát $d(i_l, j_m)$ jelöli, és ezen távolságok halmazát D_{ij}, akkor az i-edik és j-edik csoport közötti taxonomikus távolságot a módszerre jellemző

$$d(i, j) = f(D_{ij})$$

függvénnyel számítjuk ki.

Az eddigiekből következik, hogy az egyes módszerek a választott $f(D_{ij})$ függvényben különböznek, vagyis abban, hogyan értelmezzük a csoportok közötti távolságot. A leggyakrabban használatos definíciókat táblázatban foglaljuk össze.

Megnevezés	$d(i, j) = f(D_{ij})$
Single-Link (SL) vagy Nearest Neighbour (legközelebbi szomszéd)	$d(i, j) = \min d(i_l, j_m)$
Weighted-Average-Link vagy Group Average (GA) (súlyozott átlagos)	$d(i, j) = \dfrac{\sum_{l}^{r} \sum_{m}^{s} \mid i_l \mid \mid j_m \mid d(i_l, j_m)}{\mid i \mid \mid j \mid}$
Unweighted-Average vagy Simple Average (SA) (átlagos)	$d(i, j) = \dfrac{\sum_{l}^{r} \sum_{m}^{s} d(i_l, j_m)}{r\, s}$
Complete-Link (CL) vagy Farthest Neighbour (legtávolabbi szomszéd)	$d(i, j) = \max d(i_l, j_m)$

11. Nem hierarchikus osztályozás

A módszer lényegét a következő meggondolás alapján érthetjük meg. Induljunk ki abból, hogy objektumaink S halmazát két osztállyá kell bontani S_1 és S_2-re, a felbontást $2^{n-1} - 1$ féleképpen végezhetjük el. Az osztályozás hatékonyságát a

$$Q = \sum_{j=1}^{n_1} \sum_{k=j+1}^{n_1} d^2(x_{1j}, x_{1k}) + \sum_{j=1}^{n_2} \sum_{k=j+1}^{n_2} d^2(x_{2j}, x_{2k})$$

számmal mérjük, vagyis az azonos clusterbe tartozó objektumpárok eltérésének négyzetösszegével. Optimális a felbontás, ha Q értéke minimális. Minthogy az összes objektumpár eltérésének négyzetösszege

$$\sum_{i=1}^{n} \sum_{k=i+1}^{n} d^2(x_i, x_k) = c$$

konstans, ezért az előbbi feltétel ekvivalens a következővel, a:

$$Q' = \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} d^2(x_{1j}, x_{2k})$$

szám legyen maximális. A $Q + Q' = c$ összefüggés miatt a két követelmény ugyanannál az S_1, S_2 felbontásnál teljesül.

A problémát csak az jelenti, hogy milyen módszerrel találjuk meg a felbontást. Az összes lehetséges esetek száma $2^{n-1} - 1$ és csak kis n értékekre végezhető el az enumeráció még számítógép segítségével is. Éppen ezért közelítő módszereket alkalmazunk. Q értékét jól közelíti a

$$Q_1 = n_1 n_2 d^2(C_1, C_2)$$

szám, elegendő ezt maximalizálni. További egyszerűsítés, ha Q_1 helyett a centroidok távolságának négyzetét maximalizáljuk:

$$Q_2 = d^2(C_1, C_2) \to \max.$$

A megoldás során egy tetszőleges felbontásból kiindulva sorra áthelyezünk egy-egy pontot a másik clusterba; az algoritmus véget ér, ha az áthelyezés nem változtat a centroidok távolságán.

A Q' és Q_2 maximalizálása nem minden esetben ad azonos eredményt, általában csak akkor, ha a két osztályba ugyanannyi pont tartozik.

A módszer általánosítható, ha az eljárás k számú osztályra bontja a sokaságot. Két osztály esetén n lépésre volt szükség, hogy eldöntsük az osztályozásról, hogy optimális-e; k osztály esetén $(k-1) \cdot n$-féle áthelyezést kell megvizsgálni.

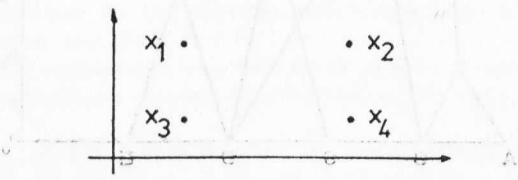
Az alkalmazott eljárások általában a fenti elvek alapján működnek, ilyenek az:

- összevonáson alapuló eljárások,
- reallokációs eljárások.

E két módszerrel kapcsolatban a következő kérdések merülnek fel:

1. Független-e az eredmény az induló osztályozástól?
2. Van-e olyan szempont, amely szerint az osztályozás optimálisnak tekinthető?

Az említett módszerek nem teljesítik a feltételeket, ha pl. a 4. ábrán levő pontokat tekintjük.



4. ábra

Az induló osztálybasorolás legyen $S_1 = \{x_1, x_2\}$, $S_2 = \{x_3, x_4\}$. Az algoritmus az első lépésben végetér, mert minden pont közelebb van a saját osztálya centroidjához, mint a másik osztályéhoz. Ugyanakkor nyilvánvaló, hogy az osztályozás nem minősíthető jónak. A centroid módszerek nem elégítik ki az egyértelműség és monotonitás követelményét.

12. Hierarchikus osztályozás

A módszer kialakítása *Jardin* és *Sibson* nevéhez fűződik. A hierarchikus módszereknek két fő típusa van

a) *Összevonó (agglomeratív) eljárások:*
Indulásnál minden pontot külön clusternek tekintünk és az egyes lépések során mindig két osztályt egyesítünk.

b) *Felosztó (divizív) eljárások*
Az előbbi módszerrel ellentétben itt az indulásnál egyetlen osztálynak tekintjük az objektumok halmazát, és az egyes iterációk során valamelyik osztályt mindig két osztállyá bontjuk.

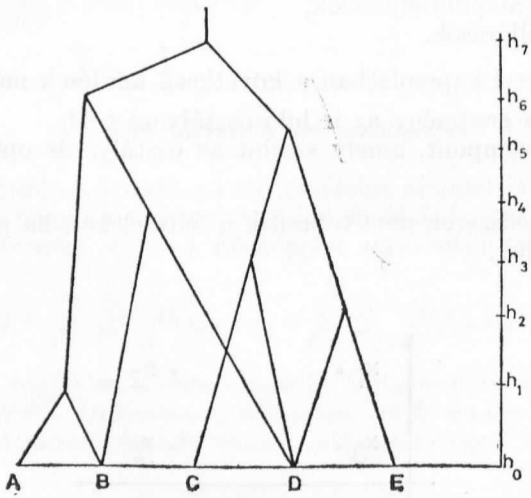
Az említett módszerek minden egyes h szinten ekvivalencia osztályokat határoznak meg az S halmazon. Általánosabb az ún. B_k-módszer, amely az egyes csoportok között átfedéseket is megenged, az átfedés mértékét a k paraméter határozza meg.

Legyenek $S_i, S_j \subset S$ a h szinten a B_k által kialakított csoportok, akkor

$$| S_i \cap S_j | \leq k - 1, \quad \forall h > 0\text{-ra.}$$

A B_k módszer $k = 1$ esetén az SL módszert adja.

A polihierarchikus elnevezés az osztályok kialakulását reprezentáló k-dendogram alapján indokolt. Ez olyan fa-struktúra, ahol minden csomópontból k számú út vezethet a magasabb szinteken levő pontokhoz. Az 5. ábrán egy 3-dendogramot mutatunk be.



5. ábra

A k-dendogram $k > 1$-re mint

$$c_n: [0, \infty) \to \Sigma(S)$$

leképezés adható meg, ahol (S) az S-en értelmezett szimmetrikus reflexív relációk halmaza. A k-dendogramnak olyan $M(d)$ feleltethető meg, amely az ún. gyenge k-ultrametrikus egyenlőtlenséget teljesíti az ultrametrikus helyett. Azaz, ha

$$P \subseteq S \text{ és } |P| = k, \text{ akkor } \forall (A, B) \in S\text{-re},$$

$$d(A, B) \leq \max \{d(x, y) \mid x \in PU \{A, B\}; y \in P\}.$$

Ezt az összefüggést a gyakorlatban előforduló DC mátrixok közül lényegesen több elégíti ki, mint az ultrametrikus egyenlőtlenséget.

Rohlf olyan minimális élhosszúságú fák előállításán alapuló számológépes algoritmust dolgozott ki, ahol fennáll, hogy az input DC mátrix azon elemei, amelyek kielégítik a gyenge k-ultrametrikus egyenlőtlenséget, nem változnak az osztályozás folyamán. Számos alkalmazásnál azonban nem engedhető meg az osztályok átfedése, ugyanakkor igény a valóban homogén osztályok megtalálása. Ebben az irányban jelent továbbfejlesztést a k-ad fokú hierarchikus osztályozás.

13. Értékelési szempontok

A különböző klasszifikáló módszerek minden esetben valamilyen osztályozást létesítenek az objektumok összességében. A kapott osztályozást többféle szempont szerint minősíthetjük.

1. Az eljárás eredményéhez valamilyen mérőszámot rendelünk. Pl. az egyes clusterek belső szórásának négyzetösszegét viszonyítjuk a teljes szórásnégyzethez. Ennek közelítő értéke

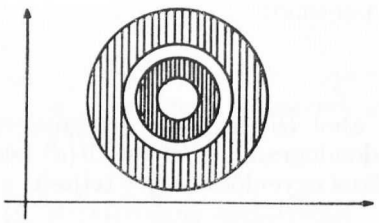
$$h = \frac{\sum_{i}^{k} \sum_{j}^{n_i} d^2(C_i, x_{ij})}{\sum_{i=1}^{n} d^2(C, x_i)},$$

$0 \leq h \leq 1$; h akkor lesz nulla, ha minden objektum egy-egy osztályt alkot, és akkor egy, ha minden osztály centroidja azonos.

Nyilvánvalóan nem állíthatjuk, hogy az osztályozás annál jobb, minél kisebb h értéke. Érdemi összehasonlításra csak akkor van lehetőség, ha az osztályok számát (k) rögzítjük, de még ezzel a megszorítással sem fogadható el, hogy h minden esetben az osztályozás hatékonyságát méri. Pl. a 6. ábra adekvát osztályozása esetében $h = 1$.

Ha két különböző osztályozásunk van és az osztályok száma azonos, akkor a h értékek összehasonlítása alapján dönthetünk egyik vagy másik osztályozás mellett.

Az értékelés egy másik szempontjánál az osztályozástól azt várjuk, hogy maximális információt adjon az objektumokról, vagyis a pontok eloszlása az



6. ábra

osztályok között legyen egyenletes. Ebben az esetben alkalmazható mérték az osztályozásnak, mint valószínűségi változónak az entrópiája:

$$h = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}, \qquad 0 \leq h \leq \log k.$$

$h = 0$ esetén egyetlen osztályunk van és semmi információt nem kapunk. Ha $h = \log k$, akkor minden osztályban azonos számú $\frac{n}{k}$ objektum található; ekkor maximális az átlagos információ.

A legtöbb esetben több szempontot kell egyidejűleg figyelembe vennünk, amikor egy osztályozást minősítünk. Így előfordulhat, hogy egy osztályozás valamely szempontból jobb, egy más szempontból rosszabb a másiknál. Kereshetünk tehát olyan kritériumot is, amely szerint nem lehet tetszőleges osztályozásokat összehasonlítani, de bizonyos osztályozások között mégis egyértelműen dönthetünk. Ez a meggondolás az alábbiak szerint általánosítható:

Legyen $S_1 \ldots S_n$ és $Z_1 \ldots Z_k$ az X halmaz két felbontása, és tegyük fel, hogy az $x_1 \ldots x_n$ elemek átrendezhetők egy $x_{i1}, x_{i2} \ldots x_{in}$ sorozattá oly módon, hogy ha x_j és x_k azonos osztályba tartoznak az S felbontásban, akkor a nekik megfelelő x_{ij} és x_{ik} pontok is azonos osztályba tartoznak a Z felbontásban és fordítva. Ilyen feltételek mellett azt mondhatjuk, hogy az S felbontás legalább olyan jó, mint a Z, ha

a) azonos osztályba tartozó tetszőleges pontpárra

$$d(x_j, x_k) \leq d(x_{ij}, x_{ik}) \text{ és}$$

b) különböző osztályokba tartozó tetszőleges pontpárra

$$d(x_j, x_k) \geq d(x_{ij}, x_{ik}).$$

Ha legalább egy helyen határozottan egyenlőtlenség áll fenn, akkor az S felbontás jobb, mint Z.

Az osztályozások között tehát egy részben rendezési relációt értelmezhetünk. Tehát ahhoz hogy két osztályozás összehasonlítható legyen szükséges (de nem elegendő) feltétel, hogy az osztályok megfeleltethetők legyenek egymásnak, abban az értelemben, hogy a megfelelő osztályokba ugyanannyi objektum tartozzék.

2. A klasszifikációs módszerektől megköveteljük, hogy az eredmény független legyen a kiinduló osztályozástól. Egy további gyakori követelmény, hogy a módszer a lineáris transzformációkkal szemben invariáns legyen: ha az $x_1 \ldots x_n$ pontok helyett az $ax_1 + b, \ldots ax_n + b$ pontokra alkalmazzuk az eljárást, az eredmény ne változzék. E követelmény teljesülése azon is múlik, hogyan definiáljuk a távolságot pontjaink között. Ha a vektorok hajlásszögének koszinuszát tekintjük távolságnak, akkor az $ax + b$ transzformáció hatására a pontok közötti távolság nem arányosan fog változni. A módszerek stabilitásának értékelésére egy lehetőség: Tegyük fel, hogy egy eljárással az $S_1 \ldots S_i \ldots S_k$ felbontást kaptuk, természetes követelménynek látszik, hogy ha az S_i osztály objektumait elhagyva megismételjük az eljárást, akkor az $S_1 \ldots S_{i-1}, S_{i+1} \ldots S_k$ osztályozást kell kapnunk.
3. Magának az adathalmaznak az értékelésére általánosan használható módszer nem adható. Adhatók ugyan absztrakt definíciók, hogy mikor „jó" egy ponthalmaz struktúrája, de ezek a feltételek a gyakorlatban szinte sohasem teljesülnek. Amit tehetünk: többféle eljárással megismételjük az elemzést, s ha eredményeink összhangja megfelelő, elfogadjuk azokat. Ellenkező esetben munkánkat tovább folytatjuk. Feladatunk ilyenkor a módszerek s az adathalmaz kölcsönhatásának felderítése, a reálisan működő algoritmus kiválasztása. Mindehhez célszerűen segítségül hívhatjuk a sokváltozós adatelemzés további módszereit is.
4. Minősíthetjük változóinkat is, olyan szempontból, hogy az egyes változóknak mekkora szerepük van az osztályok kialakításában. Ha az $x_1, x_2 \ldots x_r$ változók mellett a pontok osztályozását is egy további x változónak tekintjük, akkor vizsgálhatjuk x és a többi változó kapcsolatát úgy, hogy az $R(x, x_i)$ korrelációs együtthatók értékei szerint rendezzük a változókat. Ennél hatékonyabb ha az $I(x, x_i)$ kölcsönös információval mérjük, hogy az x_i változó milyen mértékben határozza meg az osztályozást. Ha $I(x, x_i) = 0$, az x_i változónak semmilyen szerepe nem volt az osztályok kialakításában; az ilyen változókat elhagyhatjuk.

Ha a változóinkat úgy számozzuk meg, hogy $i < j$ esetén $I(x, x_i) \leq I(x, x_j)$ legyen, akkor magukat a változókat is osztályozhatjuk. A metrika ebben az esetben:

$$d(x_i\, x_j) = |\, I(x, x_i) - I(x, x_j)\, |.$$

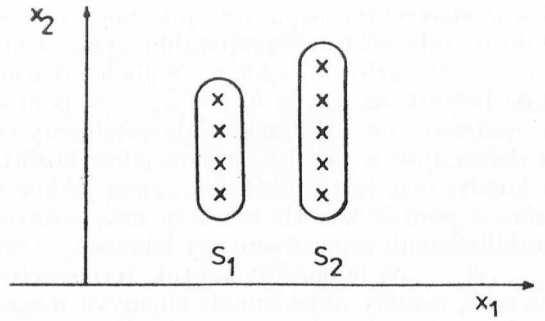
A legegyszerűbb eset, ha két csoportra osztjuk a változókat:

$$S_1 = x_1 \ldots x_j\text{: irreleváns változók,}$$

$$S_2 = x_{j+1} \ldots x_r\text{: releváns változók.}$$

A változók osztályozása után célszerű megismételni az objektumok osztályozását csak a releváns változók alapján. (Hasonló célt érhetünk el, ha az osztályozás előtt elvégezzük a változók faktoranalitikus vizsgálatát.) A 7. ábrán jól látható, hogy az x_1 változó teljes mértékben meghatározza az osztályozást, x_2 független az osztályozástól. Ha az objektumokat az x_1 tengelyre vetítjük, akkor a távolságok megváltoznak ugyan, de ugyanazok a pontok fognak egy osztályba tartozni.

Az ábrán jól látható, hogy x_2 elhagyása nemcsak azért indokolt, mert nincs szerepe az osztályozásban, hanem mert az osztályozás minősége is javítható: a külső és belső szórások hányadosa jelentősen csökken.



7. ábra

Az $I(x, x_i)$ kölcsönös információkat felhasználhatjuk a változók súlyozására is: ha az x_i változó w_i súlyát éppen $I(x, x_i)$-nek választjuk, akkor megismételhető az osztályozás ezekkel a súlyokkal. Az új felosztásnak, mint x' változónak, újra kiszámíthatjuk a kölcsönös információját az egyes változókkal; a következő lépésben a súlyokat $w_i' = I(x', x_i)$-re módosíthatjuk. Az iteráció végén kapott súlyok fejezik ki, hogy az egyes változóknak milyen szerepük van a végső osztályozásban.

(Beérkezett: 1977. aug. 15-én.)

IRODALOMJEGYZÉK

1. Anderberg, M. R.: Cluster Analysis for Applications. Academic Press, New York, 1973.
2. Everitt, B.: Cluster Analysis. London, 1974.
3. Ball, G. H.: Classification Analysis. California, 1970.
4. Ball – Hall – Isodata, A.: A Novel Method of Data Analysis and Pattern Classification. California, 1968.
5. Beale, E. M.: Euclidean cluster analysis. North Holland C.; Amsterdam, 1970.
6. Forgy, E. W.: Cluster Analysis of Multivariate Data. Biometrics. Vol. 21. No. 3. p. 768.
7. Birnbaum and Maxwell: Classification procedures based on Bayes formula. University of Illionois Press, 1965.
8. Bryan, J. K.: Classification and clustering using density estimation. Columbia, 1971.
9. Csibi-Gulyás: A számítógépek tanítása. Természet Világa, 1973. VIII.
10. Csibi, S.: Optimális döntésfüggvények iteratív tanulásáról. Preprint TKI, 1971.
11. Gulyás, O.: Tanuló algoritmusok reprodukáló magú Hilbert terekben. Szemináriumi Közlemények TKI, 1971.
12. Győrfi, L.: A potenciálfüggvényes algoritmusok konvergenciája. TKI, 1971.
13. Cole, A. J.: Numerical Taxonomy. Academic Press. New York, 1969.
14. Goodman and Kruskal: Measures of association for cross classifications. I. Amer. Statist. Assoc. Vol. 49.
15. Fritz, J.: Az alakfelismerés statisztikus módszerei. MTA Budapest, 1974.
16. Harman, H. H.: Modern Factor Analysis. University of Chicago Press, Chicago, 1960.
17. Harrison, I.: Cluster analysis. Metra, 1968.
18. Jardine and Sibson: The structure and construction of taxonomic hierarchies. Math. Biosciences. Vol. 1. No. 2. 1967.
19. Jardine, N.: Algorithm, methods and models in the simplification of complex data. Comput J. 13.
20. Jardine and Sibson: Mathematical Taxonomy. New York, 1971. p. 289.

21. KRUSKAL, J. B.: Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. Psychometrika 29, 1 – 27.
22. KRUSHAL, W. H.: Ordinal measures of association. J. Amer. Statist. Assoc. Vol. 33. 814 – 861.
23. MACNAUGHTON – SMITH – WILLEAMS: Numerical Classification of Sequences. The Australian Computer J. Vol. 2. No. 1.
24. MACQUEEN, J. B.: Some methods for Classification and Analysis of Multivariate Observations. Math. Statist. Vol. 1. 281 – 297.
25. ORLOCI: Összevonáson alapuló módszer növényi ökológiai rendszerek osztályozására. Journal of Ecology. 1967. 55. 193 – 206.
26. ROMNEY – SHEPARD: Multidimensional Scaling. Seminar Press. New York, 1972.
27. MAHALANOBIS: On the generalized distance in statistics. Proc. Natl. Inst. Sci. Vol. 12.
28. MARRIOT, F. H.: Practical problems in a method of cluster analysis. Biometrics Vol. 27. 501 – 514.
29. MORRISON, D. G.: Measurement problems in cluster analysis. Managment Sci. 13. 775 – 780.
30. PARKS, J. M.: Classification of mixed mode data by r-mode factor analysis and q-mode cluster analysis on distence functions. Academic Press. New York, 1969.
31. PARZEN, E.: On estimation of a probability density function and mode. Ann. Math. Statist. Vol. 33. 1065 – 1072.
32. RAO, M. R.: Cluster Analysis and Mathematical Programming. Amer. Statist. Assoc. Vol. 66. 1971.
33. ROHLF – SOKAL: Coefficient of correlation and distance in numerical taxonomy. Kausas University Sci. Vol. 45. 8 – 27.
34. SEBESTYÉN, G.: An algorithm for nonparametric pattern recognition. Electronic Computers Vol. 15. No. 6.
35. SOKAL – SNEATH: Principles of Numerical Taxonomy. San Francisco, 1963.
36. SPEARMAN, C.: Correlations of sums and differences. Brit. J. Psychol. 5. 417 – 426.
37. WALLACE – BOULTON: An information measure for classification. Comput J. Vol. 11. p. 185.
38. TOU – GONZALEZ: Pattern Recognition Principles. Addison – Wesley Publishing C. London – Amsterdam – Ontario – Sydney, 1974.
39. WARD, J. H.: Hierarchical grouping to optimize an objective function. J. Amer. Statist. Assoc. Vol. 58. No. 301.
40. LANCE – EILLIAMS: A general theory of classificatory sorting, strategies. Computer J. Vol. 3.
41. LANCE – WILLIAMS: Mixed-Data Classificatory Programs. The Australian Computer J. Vol. 1. No. 2.
42. WISHART, D.: Mode analysis: A generalization of nearest neighbor, which reduces chaining effects. Academic Press. New York, 1969.
43. WISHART, D.: An algorithm for hierarchical classifications. Biometrics. Vol. 22. No. 1.
44. ZADEH, H. A.: Fuzzy sets. Information and Control Vol. 8. 338 – 353.
45. JARDINE – SIBSON: Mathematical Taxonomy. London, 1971.
46. KEREKES ÁGNES – KISS PÉTER: A cluster analízis és egy lehetséges közgazdasági alkalmazása. (Szakdolgozat, Bp. 1977.)

CLUSTER ANALYSIS: CONCEPTS AND METHODS

In the introductory part a survey is given on classification methods, measuring scales of various type, their transformations as well as on the concepts of similarity both among criteria and objects and on measurement possibilities. All this is based on *Anderberg*'s book: Cluster Analysis for Applications.

In the subsequent parts the methodological aspects of cluster analysis are discussed. This includes also a summary of classification criteria and the types of decision functions. The authors briefly review various procedures based on durity functions estimation, on the concept of the "mixed model", on the estimation of variance within groups and on discriminancy analysis, as well as relying on graph theory, simple and complete chain methods, respectively.

In certain respects – e.g. dendogramme – technical details are also dealt with. Hierarchic, nonhierarchic, agglomerative, divisive procedures are discussed and the most relevant points of view of essessment reviewed.

The article is closed by a fairly comprehensive block diagramme reflecting systems approach and a rich bibliography.

КЛАСТЕРНЫЙ АНАЛИЗ

Во вводной части на основании книги Андерберга: „Cluster Analysis for Applications" дается обзор методов классификации, различных типов измерительных шкал, их трансформации, возможностях измерения а также понятия подобия критериев и объектов.

В дальнейшем рассматриваются методологические аспекты кластекрного анализа. Здесь происходит также и обобщение критериев классификации, типов функций принятия решений. Кратко излагаются методы, основанные на оценке функции частоты, опирающиеся на представления «смешанной модели», использующие вариационную оценку в рамках группы, в основе которых находится дискриминационный анализ, теория графов, простые методы цепей, а также полные методы цепей.

В некоторых случаях затрагиваются и, например, в отношении дендограммы, технические, неиерархические, агломеративные и дивизивные методы, излагаются наиболее важные точки зрения оценки.

Данная статья завершается довольно объемной блок — диаграммой, отражающей системный подход, а также и обширным списком литературы.