5

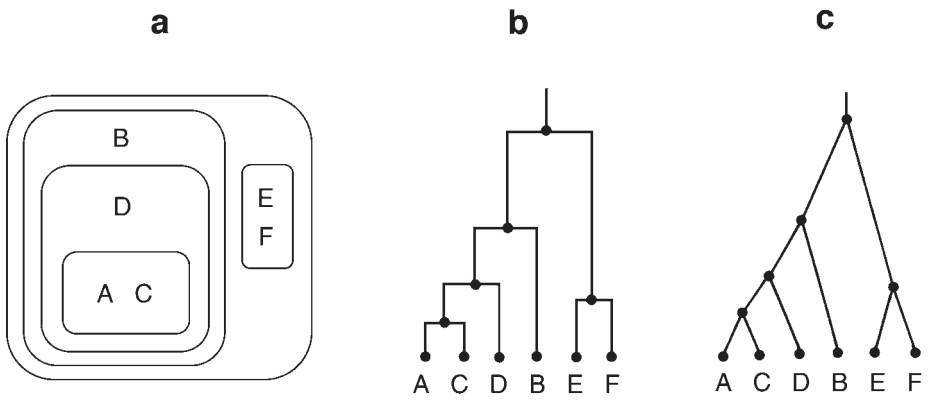
Hierarchikus osztályozás

(A "természet rendjének" keresése)

Az objektumok egyszerű felosztásán túlmenően a klasszifikációtól azt is elvárhatjuk, hogy megmutassa az egyes osztályok között fennálló kapcsolatokat is. Erre kétféle módon van lehetőségünk, az *exkluzív* és az *inkluzív* hierarchiák révén (Mayr 1982, Panchen 1992). Az első esetben a csoportokat egy lineáris rendezési reláció állítja sorba, és ez a sorbarendezés lesz az egyedüli többlet, amit az egyébként nem-hierarchikusnak is felfogható osztályozáshoz hozzáteszünk. Tipikus példa a rendfokozatok hierarchiája: egy katona csak egy csoportba tartozhat (váll-lapjának megfelelően) amely a magasabb rendfokozatúaknak alárendeltje, az alacsonyabb rendfokozatúaknak felettese. A biológiában sem ismeretlen az exkluzív hierarchia; gondoljunk a régen oly népszerű fejlettségi sorokra (*"scala naturae"*). Például, az állatvilág hierarchiájában legelöl "természetesen" maga az ember szerepel, majd a főemlősök, a többi emlős, a madarak, stb. következnek, az egysejtűekkel bezárólag (innen származik régies nevük: "véglények"). Könyvünkben ezzel a típusú hierarchiával nem foglalkozunk többet, és a hangsúlyt az inkluzív osztályozásokra helyezzük. Az inkluzív hierarchiában is van egy rendezettség: a kisebb osztályok nagyobb osztályokba vannak beágyazva. Egy objektum értelemszerűen több osztályba is beletartozik, a különböző hierarchikus szinteknek megfelelően. Ez a típus is régen ismert a biológiában, s példaként elegendő, ha a klasszikus rendszertani kategóriák (faj, genusz[1], család, rend, osztály, törzs) jól ismert kapcsolatrendszerére gondolunk. Az inkluzív hierarchia partíciók sorozatának is felfogható, és egy klasszikus logikai művelettel, a divízió szukcesszív alkalmazásával állítható elő. Mint majd rövidesen látni fogjuk, a divízió csak egy – és nem is a legfontosabb – módja a hierarchia előállításának.

Az inkluzív hierarchia-alkotás legalább olyan természetes képességünk, mint a particionálás. Az osztályok hierarchiába rendezése további könnyítést jelent a bennünket körülvevő világban való tájékozódáshoz, s korántsem korlátozódik a tudományos gondolkodásra. A hierarchiák könnyű intuitív értelmezhetősége az egyik oka annak, hogy a hierarchikus osz-

1 Magyar sajátosság: a genusz a növényeknél "*nemzetség*", az állatoknál viszont "*nem*", de ez az elkülönülés nehezen lesz tartható a legújabb makrotaxonómiai fejlemények tükrében.



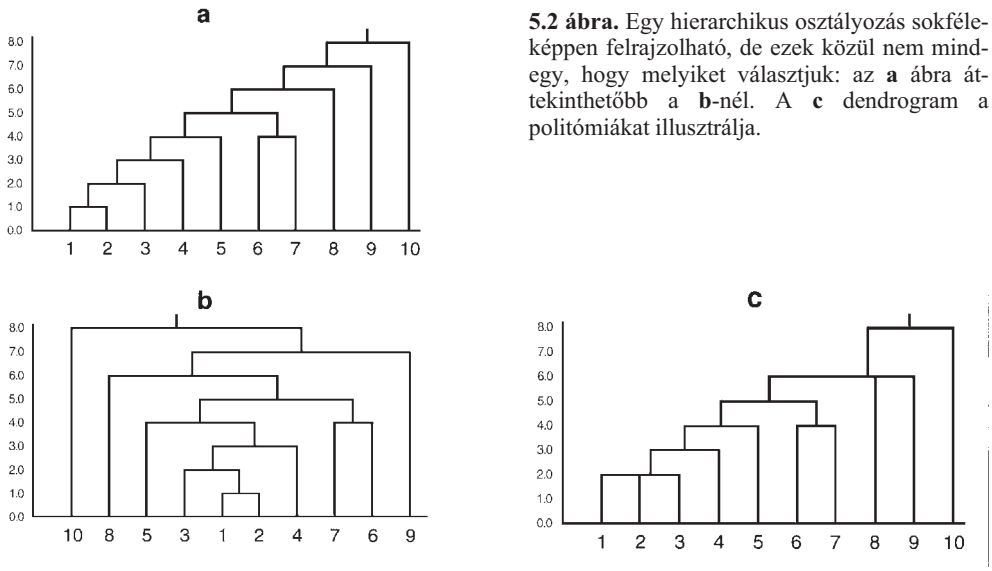
5.1 ábra. Hierarchikus osztályozások ábrázolásának alternatív lehetőségei.

tályozás központi szerepet tölt be a sokváltozós adatstruktúrák feltárásában is. Előnyük, hogy – ellentétben az előző fejezet módszereivel – az osztályok számát v. más paramétert nem kell előre megadnunk. Könnyű szívvel ajánlhatók tehát a célból, hogy segítségükkel gyors, kezdeti eredményre jussunk az adatelemzés hosszadalmas folyamatában. Mint a jelen fejezet példái is szemléltetik majd, nincs kitüntetett hierarchikus eljárás, ami bármely esetben alkalmazható lenne, tehát érdemes több módszert is alkalmazni egyidejűleg. De még ekkor is fennállhat az a veszély, hogy félrevezető eredményt (rossz szóval: "műterméket") kapunk (lásd Everitt 1980, ill. a példák), s ezért a hierarchikus módszerek csak az ordinációs eljárásokkal kiegészítve ajánlhatók még akkor is, ha vizsgálódásunk végső célja az osztályozás (pl. taxonómia).

A hierarchikus osztályozás többféleképpen ábrázolható, például egymásba ágyazott síkidomok segítségével ("kontúr-diagram", 5.1a ábra). Ennek megrajzolása azonban sok osztály esetén nehézkes és csak az osztályok topológiai viszonyai látszanak. A legáltalánosabb és jobban interpretálható ábrázolás[2] *dendrogramok* segítségével történik (5.1b-c ábra). A dendrogram egy fa-gráf, melynek végsõ (terminális) szögpontjai ("levelei") az osztályozott objektumoknak felelnek meg[3]. A kontúr-diagramokkal ellentétben a dendrogram az osztályok közötti kapcsolatot (pl. távolságot, hasonlóságot) numerikusan is ki tudja fejezni: ezt a dendrogram belsõ szögpontjainak magassága jelzi a vertikális tengelyen felmérve ("hierarchikus szint"). Ez a magasság jobban látszik, ha az éleket derékszögben megtörjük, amint az az 5.1b ábrán is látható. Ezzel teljesen egyenértékû az 5.1c ábra dendrogramja, bár ez az ábrázolásmód csak akkor célszerû, amikor nem tulajdonítunk különösebb jelentõséget a szinteknek, mert az elágazások rendszerén van a hangsúly (pl. kladogramok, 6. fejezet). A dendrogram voltaképp-

2 Vannak még más lehetőségek is, pl. a "jégcsap" diagramok (Ward 1963, Johnson 1967), de ezekre itt nem térünk ki. A kontúr diagramok egyébként nem vetendők el teljesen; az ordinációs térben alkalmazva hatékonyak lehetnek az eredmények interpretációjában (vö. 7.2 ábra).

3 A belső szögpontok nem azonosíthatók a vizsgálatban szereplő objektumokkal. Az ilyen gráfokat a szakirodalom *n*-fa néven ismeri (*n* objektumra, vö. Bobisud & Bobisud 1972), ellentétben a minimális feszítőfával (5.4.3 rész), amelyben csak annyi szögpont van, amennyi az objektumok száma. *n*-fák a később említendő additív fák is.



5.2 ábra. Egy hierarchikus osztályozás sokféleképpen felrajzolható, de ezek közül nem mindegy, hogy melyiket választjuk: az **a** ábra áttekinthetőbb a **b**-nél. A **c** dendrogram a politómiákat illusztrálja.

pen egy speciális fa-gráf, mert "gyökere" is van, a levelektõl legtávolabb esõ szögponthoz tartozó él (mint majd látni fogjuk az 5.4.3 részben, a gyökér nélküli fáknak is van szerepe a sokváltozós adatelemzésben). A fát rendszerint "lombozatával lefelé" szokták felrajzolni, azaz a levelek vannak legalul és a gyökér legfelül; a jelen kötet is többnyire ezt a konvenciót követi. Az ábrázolás persze fordítva is történhet sõt, a dendrogram fekvõ helyzetû is lehet; mindez teljesen a rajzoló ízlésére van bízva[4]. Bizonyos mértékben az objektumok sorrendje is önkényes: a belsõ szögpontokhoz tartozó rész-fák elfordíthatók a többihez képest (2^{m-1}-féleképpen). (A szögpontok felrajzolása egyébként felesleges is). Ugyanazon hierarchikus osztályozásnak tehát igen nagyszámú de azonos tartalmú ábrázolása lehetséges. Ezek közül a "legesztétikusabb", a legáttekinthetőbb elrendezést érdemes választani (5.2 ábra), ezt rendszerint a dendrogramot rajzoló számítógépes rutin automatikusan elintézi számunkra.

A dendrogram *dichotomikus*, ha minden belső szögponthoz három él tartozik (amint ez az 5.1b-c és az 5.2a-b ábrán látható). Ha van olyan szögpont, amelyhez ennél több él fut, akkor *politomikus* dendrogramról beszélünk (5.2c ábra). Az adatok szerkezete és maga a módszer is megszabhatja, hogy a dendrogram dicho- vagy politomikus lesz-e, pl. a kladisztika több eljárása (6. fejezet) szigorúan dichotomikus fák előállítását célozza. A jelen fejezetben tárgyalt módszereknél a politomikus rész-fák jelentkezése a dendrogramban határozott jelentésű, mert az adatstruktúra bizonyos tulajdonságaira utalhat.

A dendrogramok kapcsán egy speciális metrika-típusról is beszélnünk kell. Bármely dendrogram felírható egy szimmetrikus mátrix, **E**, formájában, amelyben e_{jk} az a legalacsonyabb hierarchikus szint, amelynél a j és k objektumok még egy osztályba tartoznak. Ha bármely három objektumra, függetlenül attól, hogy melyiket jelöljük h-val, j-vel, illetve k-val, az alábbi egyenlőtlenség teljesül:

4 Sneath & Sokal (1973) immár klasszikus numerikus taxonómia könyvében például a három ábrázolásmód egészségesen keveredik egymással.

$$e_{jk} \leq \max \{ e_{hj}, e_{hk} \} \quad (5.1)$$

akkor a dendrogrammal implikált *e* függvény *ultrametrika* (Johnson 1967). A háromszög-egyenlőtlenség axiómájánál szigorúbb megszorítást jelentő fenti összefüggés valójában azt fejezi ki, hogy bármely objektumhármast megvizsgálva két távolságértéket egyenlőnek találunk, a harmadik pedig szükségképpen nem lehet nagyobb a másik kettőnél. Mindez a dendrogramon a hierarchikus szintek monoton növekedésében nyilvánul meg. Vannak olyan hierarchikus osztályozó módszerek (pl. a centroid eljárás), amelyeknél a fenti egyenlőtlenség nem mindig áll fenn, ami a dendrogramon visszafordulások (*"reversal"*) formájában jelentkezik (5.9 ábra). Ebből nem következik az, hogy az illető módszer "rossz", hiszen a példaként említett módszer nagyon is értelmesen jellemezhető geometriailag. A dendrogramon esetlegesen jelentkező sok visszafordulás természetesen megnehezíti az eredmény értékelését.

5.1 A hierarchikus osztályozó algoritmusok főbb típusai

Hierarchikus osztályozások előállítására nagyon sok eljárás közül választhatunk. Ezeket a módszereket az alapalgoritmus jellege szerint sokféleképpen jellemezhetjük, s akár hierarchikusan osztályozhatjuk is.

Agglomeratív versus divizív algoritmusok

Az osztályozás folyamata alapvetően kétféle lehet. Az *agglomeratív* algoritmusok kiindulásképpen minden objektumot külön osztálynak tekintenek, s az egyes lépésekben ezeket az osztályokat páronként vonják össze növekvő tagszámú csoportokba a közöttük mért távolság (v. más mérték, pl. homogenitás) figyelembevételével. Az agglomeratív osztályozás utolsó lépésében minden objektum egy osztályba kerül. A *divizív* algoritmusok éppen fordítva járnak el: kezdetben az összes objektum egy osztályt alkot, amelyet alkalmas módon két osztályra bontunk, ezeket további divízióval még kisebb csoportokra osztjuk fel, s a felosztást addig folytatjuk, amíg az egyelemű osztályokhoz el nem jutunk (bár a felosztást előbb is abbahagyhatjuk valamilyen leállítási feltétel alapján). Egyik esetben sincs javítási lehetőség az elemzés közben: ha két objektum az elején egy csoportba került (agglomeratív módszerek), ill. elvált egymástól (divizív módszerek), akkor azon már nem lehet változtatni akkor sem, ha az egy másik szinten előnyös lenne. Az ember szubjektív osztályozó tevékenysége a divizív eljárásokhoz áll közelebb, a klasszifikáció számítógépes végrehajtása viszont az agglomeratív módszerekkel tűnik egyszerűbbnek.

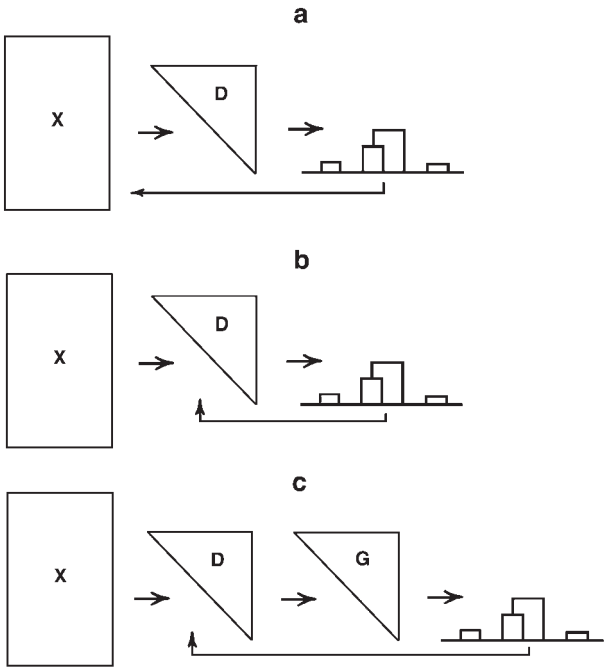
Monotetikus versus politetikus osztályozások

Ha az osztályozás egyes lépései *egy* kitüntetett tulajdonság szerint hoznak létre csoportokat, akkor *monotetikus* klasszifikációról beszélünk. Az osztályokon belül az objektumok szükségképpen *azonosak* erre a tulajdonságra nézve. A *politetikus* algoritmusok esetében viszont több változó "együttes véleménye" alakítja ki a csoportokat, az osztályon belül nem kell teljesen megegyezniük minden objektumnak egy változóra sem, mert az objektumok hasonlósága, ill. a sokdimenziós térbeli távolsága a döntő. Az agglomeratív eljárások mindegyike politetikus (bár elvileg monotetikus is elképzelhető, de ennek kevéssé lenne értelme), a divizív módszerek között viszont egyaránt találunk mono- ill. politetikusakat is. A régebbi osztályozások (akár pl. a Linné-féle törzsek) szigorú monotetikus felosztási elvéhez képest a politetikus klasszifikáció jelentős – de mondhatni: szükségszerű – engedménynek számít.

5.2 Agglomeratív módszerek

Az agglomeratív klasszifikáció során kétféle stratégia képzelhető el: a *távolság-optimalizáló* eljárások nemcsak az objektumok között, hanem a folyamat során képződő osztályok között is távolságokat (ritkábban: hasonlóságokat) mérnek ("route-optimizing methods", Williams 1971; *d*-SAHN módszerek, Podani 1989b, mely névben a betűszó a "sequential, agglomerative, hierarchical and nonoverlapping" jelzők kezdőbetűiből alakult ki, vö. Sneath & Sokal, 1973). Az osztályozás egyes lépéseiben a távolság minimalizálása (vagy a hasonlóság maximalizálása) a cél. E módszereknél döntő, hogy miképpen számítják ki a két v. többelemű csoportok közötti távolságokat (5.5 ábra, 5.1 táblázat), s geometriailag rendszerint jól értelmezhetők. A *homogenitás-optimalizáló* (=heterogenitás minimalizáló) módszerek, bár kiindulásképpen ugyanúgy távolság (hasonlóság) mátrixot alkalmaznak, az osztályok között már nem távolságokat mérnek. Két objektum vagy osztály összevonásának az ugyanis a feltétele, hogy a kapott új osztály valamilyen "homogenitási" mérőszáma optimális legyen a többi lehetséges összevonáshoz képest (*h*-SAHN módszerek, Podani 1989b). Ilyen mérőszám lehet a variancia, az entrópia vagy az osztályon belüli átlagos hasonlóság (gondoljunk vissza a 3.7 alfejezetre). E módszereknek nehézkes – vagy nem is létező – a geometriai interpretációja.

Mielőtt a konkrét algoritmusokat bemutatnánk, meg kell ismerkednünk néhány további alaptulajdonságukkal is, amelyek már inkább az osztályozás technikai kivitelezésével kapcsolatosak, s nem feltétlenül érintik az elveket. Először az adattárolási lehetőségeket említjük meg (5.3 ábra). Legkisebb memóriaigénye van azoknak a módszereknek, amelyek a távolságmátrix kiszámítása után már nem kérik többet a nyers adatokat; ekkor a dendrogram felépítéséhez a távolságmátrixba kezdetben beírt információ is elegendő, s e mátrix értékei íródnak felül az



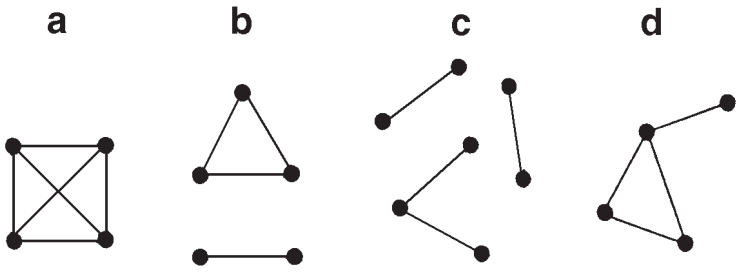
5.3 ábra. Az agglomeratív módszerek számításmenetéhez szükséges mátrixok.

algoritmus egyes lépéseiben (5.3a ábra). E módszereket *kombinatorikus* algoritmusok néven ismeri az osztályozás szakirodalma (Williams 1971, Lance & Williams 1966); e – kissé félrevezető – név arra utal, hogy az osztályok közötti távolságok vagy a homogenitás értékek a mátrix kezdeti értékeiből "kombinálhatók ki" alkalmas formulák segítségével. Az algoritmusok következő csoportja az adatmátrix és a távolságmátrix egyidejű tárolását igényli (5.3b ábra). Ekkor, az osztályozás egyes lépéseiben a távolságmátrix átszámításához továbbra is szükség van az eredeti adatokra ("*stored data approach*", Anderberg 1973). A centroid módszernek, például, jól ismert mindkét változata. A harmadik algoritmuscsoport két szimmetrikus mátrix egyidejű tárolását igényli (Podani 1989a, 1994, 5.3c ábra). A nyers adatokra a távolságmátrix kiszámítása után már nincs szükség, a távolságmátrixból azonban egy újabb mátrixot kell kiszámítanunk az osztályozás minden egyes lépésében. Erre példa az osztályok közötti és az osztályokon belüli távolságátlagok hányadosának minimalizálása (5.2.4 rész): ekkor a második szimmetrikus mátrix tartalmazza az összes lehetséges páronkénti összevonás után adódó ilyen átlagokat.

További szempont lehet az, hogy az analízis egyes lépéseiben hány összevonást hajtunk végre. Első közelítésben azt gondolhatnánk, hogy minden lépésben csak azt az objektum- (vagy később: osztály-) párt keressük ki, amelyre legkisebb távolságérték adódik, s csak ezeket vonjuk össze (*legközelebbi pár* algoritmus). Bizonyos módszerek azonban jelentékenyen felgyorsíthatók, ha a kölcsönösen legközelebbi párokat összevonjuk akkor is, ha a közöttük mért távolság messze nem a legoptimálisabb a mátrixban (azaz, ha az *A* osztályhoz *B* van a legközelebb, és viszont; *reciprok-pár* algoritmus). Bruynooghe (1978) és Gordon (1987) mutatta meg, hogy mely módszerekre alkalmazható ez a felgyorsítás az eredmény megváltozása nélkül (5.1 és 5.2 táblázatok utolsó oszlopai).

Az agglomeratív osztályozás egy kritikus, és gyakran figyelmen kívül maradó problémája az *egyezések* feloldása. Egyezésnek ("*tie*") nevezzük azt a – bináris adatok esetében nem ritka – esetet, amikor a legkisebb távolságérték több helyen is szerepel a mátrixban. Ekkor sok módszer önkényesen kiválasztja valamelyiket, s az ahhoz tartozó két osztályt vonja össze. Nem kell hangsúlyoznunk, hogy ez a döntés nagymértékben befolyásolhatja a kapott eredményt (Podani 1980, ad meg egy konkrét cönológiai példát). Ha valamelyest objektíven akarunk dönteni, akkor figyelembe kell vennünk az alábbiakat.

Az egyezéseket legjobban gráfokkal illusztrálhatjuk (Podani 1989a). Tekintsük a szóbanforgó objektumokat egy *G* "egyezés gráf" szögpontjainak. Két pont között akkor legyen él, ha a megfelelő távolság éppen minimális a távolságmátrixban. A négy lehetséges alapesetet az 5.4 ábra foglalja össze.



5.4 ábra. Az agglomeratív osztályozás során felmerülő egyezések különféle típusai (Podani 1989a).

a) *G* egy teljes gráf (minden szögpont össze van kötve a többivel);
b) *G*-ben izolált részgráfok vannak, azok mindegyike önmagában teljes;
c) *G*-ben az izolált részgráfok legalább egyike nem teljes; és
d) *G* nem teljes gráf, de nem esik szét izolált részgráfokra sem.

Az *a*-*b* esetekben az egyezések feloldása eléggé egyértelmű: egy *többszörös fúzióval* minden objektumot összevonunk (*a* eset) vagy pedig *szimultán* (egyidejű) fúziókkal több osztályt alakítunk ki egyszerre, amelyek mindegyike egy részgráfnak felel meg (*b* eset). A másik két szituációban kétféle megoldás is lehetséges:

– az *egyszerű lánc feloldás* annyi csoportot hoz létre, amennyi részgráf van (3 ill. 1 csoport az 5.4c-d ábrán).

– a *szuboptimális fúzió* révén figyelmen kívül hagyjuk az egyező távolságértékeket, s a következő legkisebb távolságot keressük meg a mátrixban, melyre nézve már nincsenek egyezések.

Ha tehát kétségeink vannak az analízis egyértelműségét illetően – s ez különösen prezencia/abszencia adatok esetén lehet így – akkor célszerű az elemzést az egyezések mellőzésével és feloldásával is végrehajtani s utána összehasonlítani az eredményeket. A **NT-SYS** programcsomag (Rohlf 1993a) pedig lehetőséget ad arra, hogy az egyezések önkényes feloldásából adódó összes lehetséges dendrogramot megvizsgáljuk (bár ez áttekinthetetlenül sok is lehet!). Backeljau et al. (1996) összefoglalója azt vizsgálja meg, hogy egyes programcsomagok miként kezelik az esetleges egyezéseket.

Most pedig már valóban itt az ideje, hogy a konkrét módszerekkel részletesen is megismerkedjünk.

5.2.1 Távolság-optimalizáló kombinatorikus módszerek

Kiindulópontjuk az objektumok **D** távolság- vagy különbözőség-mátrixa (amennyiben hasonlóságokkal van dolgunk, azokat előzetesen különbözőséggé kell átalakítani a 3.4 formula alapján, hogy az 5.1 táblázat érvényes legyen). Az eljárás egyes lépéseiben megkeressük az egymáshoz legközelebbi objektumpárokat s ezeket egy osztályba vonjuk össze. Az összevonás szintjét a dendrogram mellé rajzolt tengelyen olvashatjuk le. Ezután kiszámítjuk az újonnan kapott osztályok és a többi osztály vagy objektum távolságait, miközben a távolságmátrix felesleges sorai és oszlopai kiesnek (két objektum összevonásával egy sor, ill. oszlop válik feleslegessé **D**-ben). A kulcskérdés az új távolságok kiszámításának módja, ehhez a Lance - Williams (1966, 1967a) féle rekurziós formula alkalmazható:

$$d_{h,ij} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta\, d_{ij} + \gamma \mid d_{hi} - d_{hj} \mid \qquad (5.2)$$

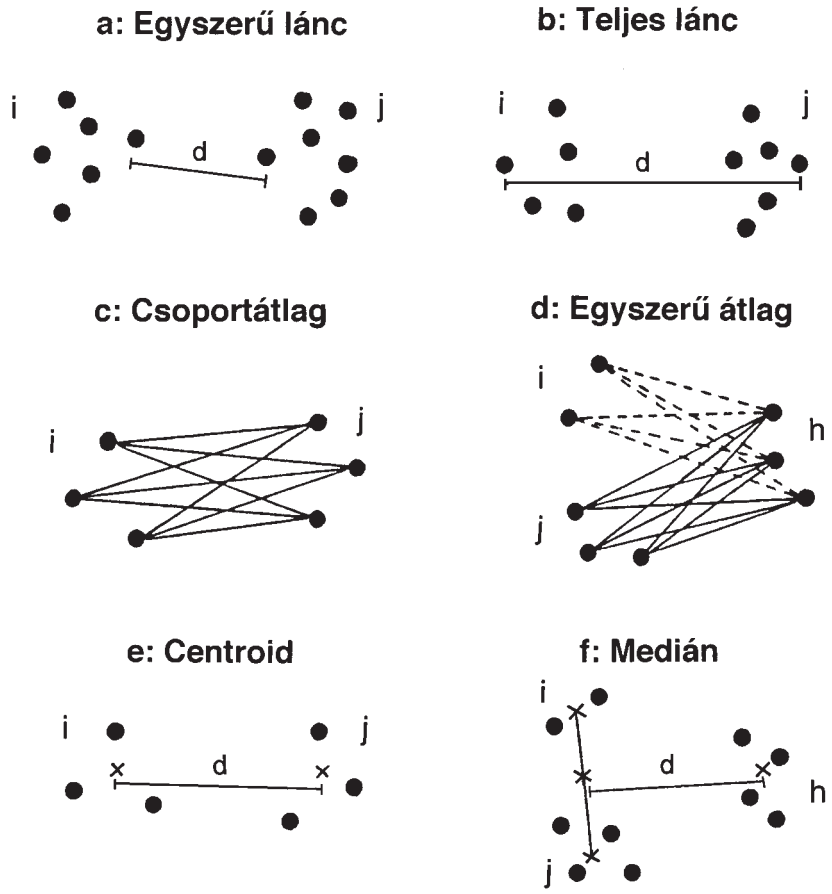
Amit keresünk, a $d_{h,ij}$, az *i* és *j* objektumokból éppen létrehozott új osztály, valamint egy másik *h* osztály (vagy objektum) távolsága (vagy távolságnégyzete, 5.1 táblázat). d_{hi}, d_{hj} és d_{ij} a megfelelő objektumpárok távolságai. A paraméterek az algoritmusra jellemzőek, sokszor az osztályokban előzőleg egyesített objektumok számától függenek (5.1 táblázat).

Egyszerű lánc (legközelebbi szomszéd) módszer (Florek et al. 1951, Sneath 1957). Két osztály távolságát az egymáshoz legközelebb eső, de nem egy osztályba tartozó objektumaik távolságaként definiáljuk (5.5a ábra). Ez a módszer az osztályok szeparáltságát emeli ki: megnyúlt pontfelhőket is felismer, viszont "zavarba jön", ha az osztályok között nincs éles elválás. Az osztályok belső kohéziója szinte teljesen mellékes, és könnyen előadódik az az eset, hogy egy már meglévő kis osztály egyenként magához vonzza a többi objektumot (ez a dendrogramon

"*lánchatásként*" jelentkezik). A módszer rendkívüli előnye viszont – a többivel szemben –, hogy az osztályozást nem befolyásolják az egyezések, és az eredmény változása arányos az adatok megváltoztatásának mértékével (Jardine & Sibson 1971).

Az elmondottakat megerősítik a 4.3a-f ábrák kétdimenziós ponteloszlásainak elemzései (5.6 ábra). Az egyszerű lánc módszer jól "felismerte" a **b** és **e** esetek elkülönülő osztályait, alakjuktól függetlenül, és csaknem sikeresen elkülönítette a **d** ábra három megnyúlt pontfelhőjét is (itt a zavart a 8. objektum okozta, amely túlságosan távol esik mindentől, s így a módszer kívülállóként ["outlier"] értékelte). Az egyszerű lánc módszer csoportosulásokat fedezett fel a random esetben is (**a**), ellenben nem lehetett "becsapni" a csaknem szabályos ponteloszlással (**f**). Leginkább zavarba ejtő az egyszerű lánc módszer kudarca a **c** esetben, hiszen a két fő osztály teljesen összekavarodik az erős lánchatást mutató dendrogramon, s csak kisebb "csoportocskák" ismerhetők fel az eredményben.

Teljes lánc (legtávolabbi szomszéd) módszer (Sorensen 1948, Lance & Williams 1967a). Minden szempontból az előző ellentéte; két osztály távolságát a legtávolabbi objektumaik tá-



5.5 ábra. Hat távolság-optimalizáló osztályozó algoritmus alapelvének geometriai ábrázolása (Podani 1994).

5.1 táblázat. A távolság-optimalizáló kombinatorikus algoritmusok paraméterei és főbb jellemzői. n_i és n_j az éppen összevont i és j osztályban előzőleg meglevő objektumok száma.

Név	α_i	α_j	β	γ	Kezdeti érték **D**-ben	Reciprok-pár algoritmus használható (+)
Egyszerű lánc	1/2	1/2	0	–1/2	d_{ij}	+
Teljes lánc	1/2	1/2	0	1/2	d_{ij}	+
Csoportátlag	$n_i / (n_i+n_j)$	$n_j / (n_i+n_j)$	0	0	d_{ij}	+
Egyszerű átlag	1/2	1/2	0	0	d_{ij}	+
Centroid	$n_i / (n_i+ n_j)$	$n_j / (n_i+n_j)$	$-n_i n_j / (n_i+n_j)^2$	0	d^2_{ij}	–
Medián	1/2	1/2	–1/4	0	d^2_{ij}	–
β-flexibilis	1/2 (1-x)	1/2 (1–x)	x (<1)	0	d_{ij}	–
(β,γ)-flexibilis	1/2 (1-x)	1/2 (1–x)	nincs korlát	nincs korlát	d_{ij}	–
Flexibilis csoportátlag	$(1–x) (n_i / (n_i+n_j))$	$(1–x) (n_j / (n_i+n_j))$	x (<1)	0	d_{ij}	–

volságával definiálja (5.5b ábra). Új csoportok kialakulásának jóval nagyobb esélye van a klasszifikáció folyamatában, mint az előző módszernél. A láncképzéssel ellentétes effektus könnyen előfordul: a dendrogramok szabályos "építkezésűek", lépcsős hierarchiát mutathatnak még akkor is, ha az adatok szerkezete egyáltalán nem indokolja (ezt "lépcsőhatásnak" is nevezhetnénk). A módszer viszont jól kimutatja a nem szeparálódó, de viszonylag erős kohéziójú csoportokat.

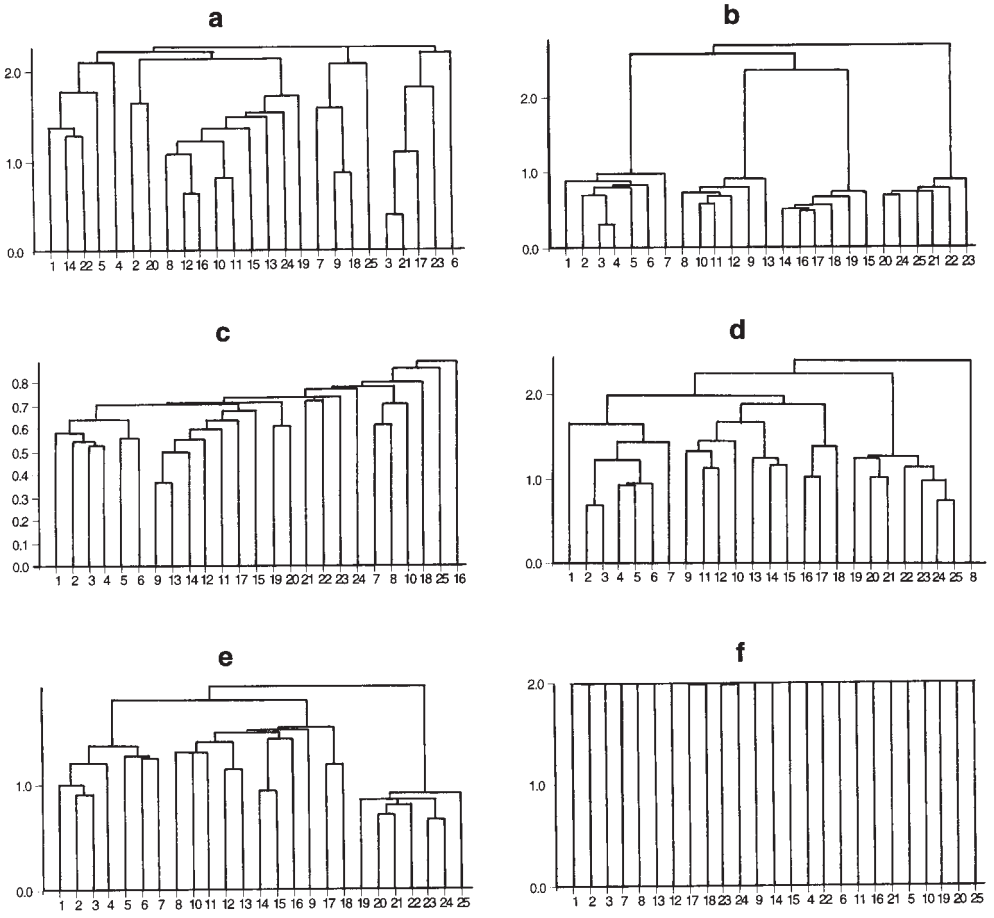
Az 5.7 ábra alapján meggyőződhetünk a fenti jellemzés helyességéről. A random (**a**) és a majdnem szabályos (**f**) ponteloszlásokra kapott dendrogramok jól elkülönülő osztályok létezését sejtetik. A random esetben valóban vannak pontsűrüsödések, amelyek a dendrogramon is felismerhetők, a szabályos ponteloszlásra kapott hierarchia viszont határozottan "műtermék jellegű". Való igaz, hogy ezekhez képest a **b** dendrogramon jóval nagyobbak az ugrások, amelyek egyértelműen jelzik a négy csoport elkülönülését. Ez azonban csak egy ilyen alapos összehasonlítás során válik nyilvánvalóvá, egyébként könnyen félrevezethet bennünket a dendrogram alakja. A **c** eset éppen összeérő két osztályát viszont jól érzékeli az elemzés, a kritikus 14. objektum a jobboldali csoportba került. A **d** és **e** esetek "nem szabványos" pontfelhőit a módszer nem tudja egyértelműen kimutatni. 3-osztály szinten az 5.7d dendrogram a 4.3d ábra – k-közép módszerrel kapott – felosztására emlékeztet, a 19-25 objektumok alkotta csoport itt is elkülönül, a másik kettő viszont elkeveredik. Az **e** ábra {19-25} csoportja pedig magával "rántja" az 1-7 objektumokat. Összefoglalva tehát: a teljes lánc módszer voltaképpen csak két esetben "vizsgázott" megfelelően.

Csoportátlag eljárás (UPGMA, Sokal & Michener 1958, Rohlf 1963). Az egyszerű és a teljes lánc módszer közötti átmenetet képviseli, megpróbálván az egyik hátrányait a másik előnyeivel kompenzálni. Két osztály távolságát az összes osztályközi, páronkénti távolság (az 5.5c rajz szakaszai) *aritmetikai átlagával* definiáljuk. Az osztályozás során a távolságok átszámításakor – ellentétben az előző két módszerrel – már figyelembe kell vennünk az osztályokban előzőleg egyesített objektumok számát is (vö. 5.1 táblázat). Mindez kiderül az alábbi egyszerű számításmenetből is, amely egy teljes klasszifikációs folyamatot illusztrál öt objektumra. A példa talán elősegíti annak megértését is, hogy miért elegendő a távolságmátrix és a rekurziós

formula, és miért nem kellenek az adatok a számoláshoz. Legyen a kiinduló távolságok félmátrixa a következő (az oszlopok ill. sorok elején az objektumokat is megszámozva):

	1	2	3	4	5
1	0,000	0,632	0,683	0,730	0,775
2		0,000	0,856	0,894	1,000
3			0,000	0,440	0,516
4				0,000	0,447
5					0,000

A szemléletesség kedvéért csak egy fúziót hajtunk végre egy ciklusban. A mátrix legkisebb értéke d_{34}=0,440, azaz a 3 és 4. objektumot vonjuk először össze. A többi objektumnak a kapott új {3,4} osztállyal vett távolságértékei a 3. és a 4. objektumokkal adott értékeik átlagai lesznek (vö. a Lance-Williams formulával):



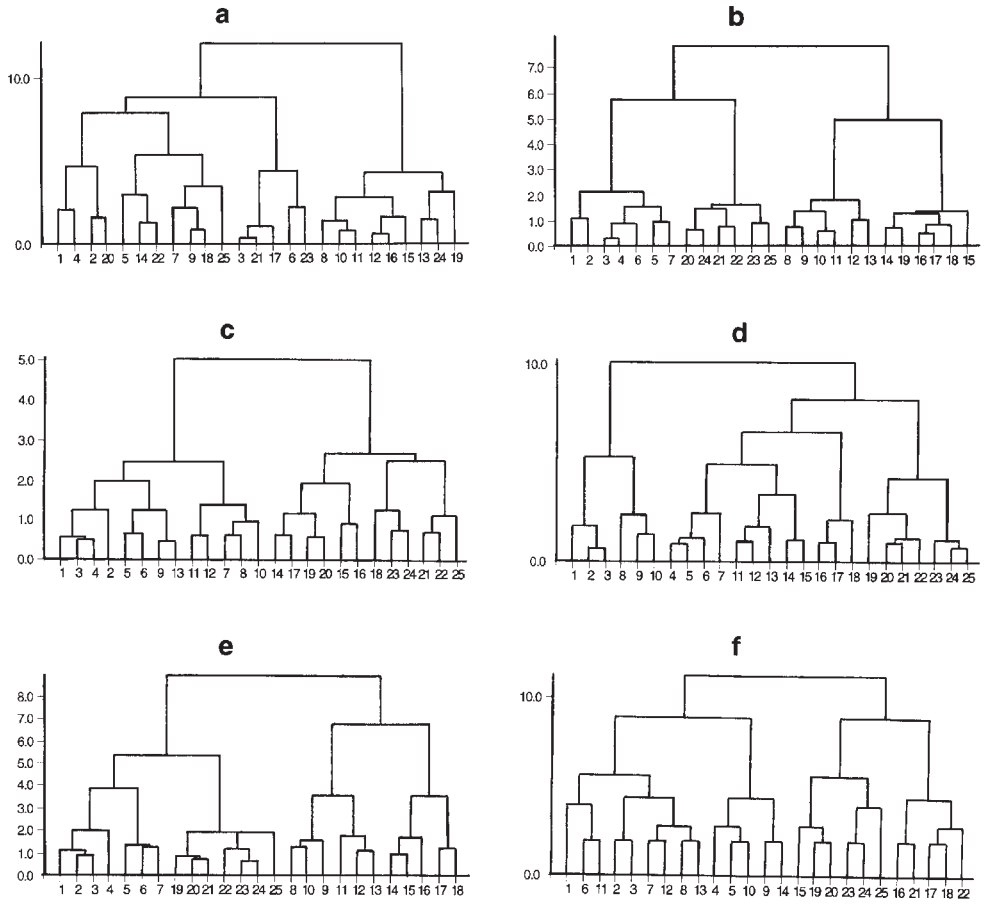
5.6 ábra. A 4.3 ábra mesterséges pontmintázatai az egyszerű lánc módszerrel értékelve.

$$d_{1\{3,4\}} = 1/2 \cdot 0{,}683 + 1/2 \cdot 0{,}730 = 0{,}706$$
$$d_{2\{3,4\}} = 1/2 \cdot 0{,}856 + 1/2 \cdot 0{,}894 = 0{,}875$$
$$d_{\{3,4\}5} = 1/2 \cdot 0{,}516 + 1/2 \cdot 0{,}447 = 0{,}481$$

Ezeket beírva (vastag betűvel) az egy sorral redukált új mátrixba kapjuk:

	1	2	{3,4}	5
1	0,000	0,632	**0,706**	0,775
2		0,000	**0,875**	1,000
{3,4}			0,000	**0,481**
5				0,000

A többi érték természetesen érintetlenül maradt. Az átszámított mátrixot megvizsgálva megállapíthatjuk, hogy a legkisebb érték a 0,481, azaz az előbb kapott csoporthoz hozzátehetjük az 5. objektumot is a 0,481-es szinten. A kapott {3,4,5} osztály távolságait az 1. és a 2. objektumoktól kell átszámolnunk az alábbiak szerint:



5.7 ábra. A 4.3 ábra példa adatainak osztályozása a teljes lánc módszerrel.

$$d_{1\{3,4,5\}} = 2/3 \cdot 0{,}706 + 1/3 \cdot 0{,}775 = 0{,}729$$
$$d_{2\{3,4,5\}} = 2/3 \cdot 0{,}875 + 1/3 \cdot 1{,}000 = 0{,}917$$

S itt látszik a lényeg: a távolságértékek fontossága az átlagolásnál arányos az illető csoportban már egyesített objektumok számával (2 ill. 1)! Az új mátrix ekkor:

	1	2	{3,4,5}
1	0,000	0,632	**0,729**
2		0,000	**0,917**
{3,4,5}			0,000

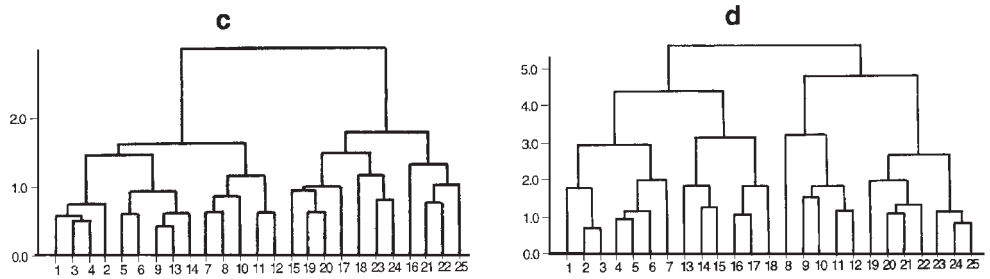
amelyben a d_{12} a legkisebb, tehát az 1. és 2. objektumok alkotják a következő csoportot. (S itt derül ki, hogy miért vonhattuk volna őket már előbb is össze: az eddigi számítások egyáltalán nem befolyásolták az 1. és 2. objektum relatív kölcsönös közelségét). Az összevonást követően a mátrix már teljesen leredukálódik, hiszen csak egy érdemleges távolságérték van benne, amit a következőképpen kapunk meg:

$$d_{\{1,2\}\{3,4,5\}} = 1/2 \cdot 0{,}729 + 1/2 \cdot 0{,}917 = 0{,}823.$$

A most kapott érték a dendrogram legfelső "gerendájához" tartozó szint. Ezek után már csak a dendrogram felrajzolása van hátra, s ezt az Olvasóra bízzuk.

Helykímélés végett a csoportátlag módszer eredményét nem mutatjuk be mindegyik esetre. Az **a** példában a dendrogram szerkezete lényegében véve megegyezik a teljes lánc módszerrel kapott dendrogrammal (5.7a), csak a kisméretű csoportokat tekintve, és a szintekben mutatkozik eltérés. A **b** esetben – s ez nem is lehet másként – ugyanúgy elválik a négy osztály, ahogy az 5.6b és 5.7b ábrákon. Bemutatjuk viszont az összeérő két csoportra kapott eredményt (**c** példa), mert ez mutatja legjobban a csoportátlag módszer átmeneti jellegét (5.8c ábra): a 6, 9, 13 és 14 objektumok a jobboldali osztályba kerültek, s ez a nem várt eredmény már "benne volt" az egyszerű lánc dendrogramban is. A **d** példa (5.8d ábra) három megnyúlt pontfelhőjének az esete pedig azt példázza, hogy egy általunk esetleg egységesnek tartott csoport (a középső) szabályosan kettészakad, azt a látszatot keltve, mintha az egyik fél a felső, a másik pedig a legalsó osztályhoz tartozna. Az **e** esetben a belső "mag" jól elvált, de az őt körülölelő csoport három egyforma részre szakadt. Az **f** példa szabályos ponteloszlására pedig majdnem olyan becsapós eredményt kaptunk, mint az előző módszerrel.

Egyszerű átlag módszer (WPGMA, McQuitty 1967). Ez az algoritmus ritkábban használatos, mert – első látásra kevéssé logikus módon – az átlagos távolságok kiszámításánál nem veszi figyelembe az osztályokban előzőleg egyesített objektumok számát. Ez azt jelenti, hogy az



5.8 ábra. A csoportátlag módszer eredménye két mesterséges példára (vö. 4.3c és d ábra).

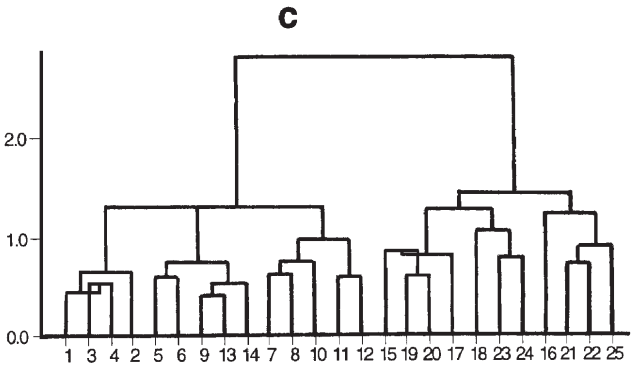
osztályozás folyamatában a kisebb méretű csoportok hangsúlyozottabban érvényesülnek. Geometriai illusztrációja kicsit nehézkesebb, de talán érdemes megpróbálkozni vele (5.5d ábra). Három osztályt kell most figyelembe vennünk, s tegyük fel: éppen most egyesítettük *i*-t és *j*-t, és eme új csoportnak keressük meg a távolságát *h*-val. Ez az objektumok közötti páronkénti távolságokból kapható meg olymódon, hogy először vesszük az *i* és *h* közötti távolságok (6 szaggatott vonal az ábrán) átlagát, majd a *j* és *h* közötti távolságok (9 folytonos vonal az ábrán) középértékét is kiszámítjuk. Ezután e két átlag átlaga adja az eredményt, s így már érthető, hogy ebben az *i*-*h* közötti távolságok nagyobb súlyt kapnak, mint a *j*-*h* közöttiek.

Sneath & Sokal (1973) szerint az ilyen típusú elemzések elsősorban akkor jöhetnek számításba. amikor a vizsgált objektumok különböző taxonokat képviselnek, de nagyon eltérő számban. Ekkor az esetlegesen kimutatott osztályok méretbeli különbségei eltűnnek.

Centroid (súlypont) módszer (UPGMC). Ha pontjainkat egy sokdimenziós euklidészi térben képzeljük el, akkor leginkább a centroid (azaz súlypont) módszer tűnik kézenfekvőnek. Az osztályokat a bennük levő objektumok adatainak átlagával definiáljuk s az osztályok közötti távolságot eme súlypontok távolságával mérjük (5.5e ábra). Első látásra úgy tűnik, hogy a számoláshoz mindvégig szükségünk van a nyers adatokra (vagyis az 5.3b ábra sémáját kell követnünk), de Lance & Williams (1967a) kimutatta, hogy a módszernek kombinatorikus megoldása is van (tehát elegendő a távolságmátrix tárolása). Pontosabban, a kiinduló mátrixba a távolságnégyzeteket kell beírnunk (vö. 5.1 táblázat). Két osztály összevonása után azonban az új súlypont közelebb kerülhet egy másik súlyponthoz, mint a két osztály eredeti távolsága volt, és ez a dendrogramon visszafordulásokat eredményez (nem teljesül az ultrametrika feltétele, ez egy helyen látszik az 5.9 ábra dendrogramján). A zavaró visszafordulásoktól eltekintve a módszert érdemes kipróbálni s másokkal is összehasonlítani, de csak akkor, ha a súlypont fogalmának értelme van a kiválasztott koefficiens és adattípus mellett (pl. euklidészi távolság esetén intervallum típusú adatokra).

A kétdimenziós pontmintázatokra számolt eredményeket, a **c** eset kivételével (5.9 ábra) nem mutatjuk be, mert most nem sok újat mondanak az egyszerű v. a teljes lánc módszer dendrogramjaihoz képest. Éppen a **c** példában viszont a csoportátlaggal erős az egyezés, kivéve az 5. objektum pozícióját.

Medián módszer (WPGMC). Ez az eljárás, amit Gowernek (1967) köszönhetünk, úgy viszonyul a centroid módszerhez, mint az egyszerű átlag a csoportátlaghoz. Ugyanis a súly-



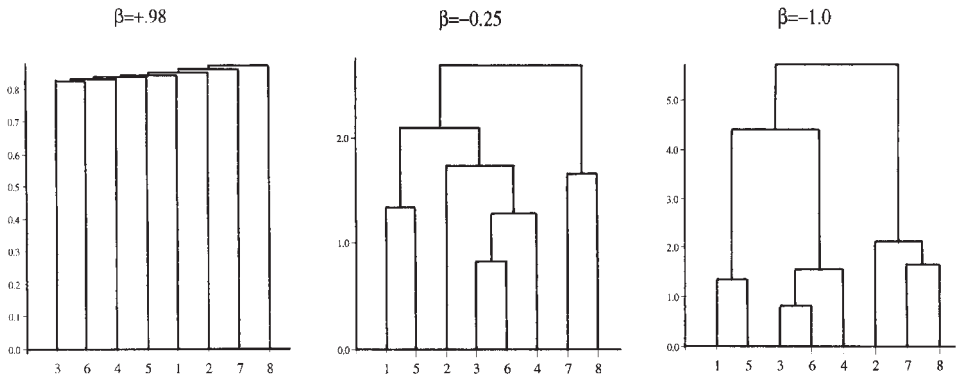
5.9 ábra. A centroid módszer eredménye gyakran visszafordulásokat mutat, mint ebben a példában is.

pont kiszámításánál nem vesszük figyelembe a már előzőleg egyesített objektumok számát. Az 5.5f ábra ezt illusztrálja számunkra. Tegyük fel, hogy az *i* és *j* osztályokat vontuk össze (2 ill. 4 elemből állanak). Először mindegyikük súlypontját kiszámítjuk, majd a súlypontok egyszerű átlagát vesszük, s ez lesz majd az összevont csoport új súlypontja (ami tehát közelebb kerül a kisebbik osztályhoz, mint a centroidnál). Ez a módszer is hangsúlyosan veszi figyelembe a kisebb osztályokat, s használata is akkor logikus, amikor az egyszerű átlag módszeré.

Flexibilis módszerek. Lance & Williams (1967a) vette észre, hogy ha a következő feltételek teljesülnek:

$$\alpha_i + \alpha_j + \beta = 1;\ \alpha_i = \alpha_j;\ \beta < 1;\ \text{és}\ \gamma = 0 \qquad (5.3)$$

akkor megadható egy – monoton növekedésű dendrogramokat eredményező – algoritmuscsalád, amely a β egyhez közeli értékeire erős lánchatást mutat (azaz emlékeztet az egyszerű láncra), míg a β=–1 esetben a teljes lánc módszerre jellemző csoportképzésre vezet (5.10 ábra). β változtatásával valójában egy – végtelen sok elemű – osztályozássorozat generálható, melynek révén sokkal többet tudunk meg az adatstruktúráról, mint bármelyik kitüntetett módszer alkalmazásával. A szerzők egyébként a β= –0,25 értéket tartják – tapasztalati alapon – a "legjobbnak" (Wiliams 1976). A β és γ értéke minden korlát nélkül is változtatható (DuBien & Warde 1979), de az így kapott eredmények értelmezhetősége már kérdéses, tehát DuBien & Warde propozíciójának inkább csak elméleti jeletősége van. Ismeretes viszont egy még újabb flexibilis módszer, melynek már nagyobb jövőt jósolhatunk (Belbin et al. 1992). Ez voltaképpen a csoportátlag módszer flexibilis verziója (5.1 táblázat utolsó sora), mint ahogy a β-flexibilis módszer az egyszerű átlag módszer változata. Szimulált adatsorokat kipróbálva a szerzők azt találták, hogy β –0,1 és 0,0 közötti értékeire a módszer jól reprodukálta az adatokban rejlő csoportosulásokat.



5.10 ábra. Osztályozási sor (β-flexibilis módszer, az A1 táblázat objektumai, a változók standardizálása terjedelem szerint). Figyeljük meg, hogy miképpen alakul át a klasszifikáció a β paraméter csökkenésével. Vizsgáljuk meg, jól illeszkedik-e a klasszifikáció a 2.2 ábra Chernoff-arcairól, ill. az adatok megvizsgálásával magunkban kialakítható képpel!

5.2.2 Homogenitás-optimalizáló kombinatorikus módszerek

A homogenitás-optimalizáló eljárások a kettő v. több objektum alkotta csoportok homogenitását maximalizálják (vagy, ami ezzel egyenlő állítás: heterogenitásukat minimalizálják). A homogenitás – ezt a gyűjtőfogalmat jobb híján használva – általánosságban az egy osztályba tartozó objektumok "egyformaságának" valamilyen mértéke. Az eltérésnégyzet-összeg, a variancia, és más mérőszámok alkalmazhatók ennek kifejezésére (l. a 3.105-3.115 függvényeket). Az összevonások során kétféleképpen dönthetünk: vagy a keletkező *új osztály* homogenitását optimalizáljuk (heterogenitását minimalizáljuk), vagy pedig az összevonást követő homogenitás *változását* optimalizáljuk (azaz: a heterogenitás növekedését minimalizáljuk). Az első stratégia kb. egyforma méretű osztályokat eredményez, s ezért ritkábban használatos (vö. Anderberg 1973). Kombinatorikus megoldása ismeretes azoknak a módszereknek, amelyek a variancia, az eltérésnégyzet-összeg vagy az osztályon belüli átlagos távolság minimalizálásán alapszanak. A kiinduló **Y** mátrix az összes lehetséges kételemű osztály heterogenitás értékeit tartalmazza, jele az i és j objektumra legyen y_{ij}. A kombinatorikus alapegyenlet (Jambu 1978, Podani 1979a) megadja, hogy az i és j összevonása után kapott új osztály, és egy harmadik h osztály összevonásával mekkora lenne a heterogenitás:

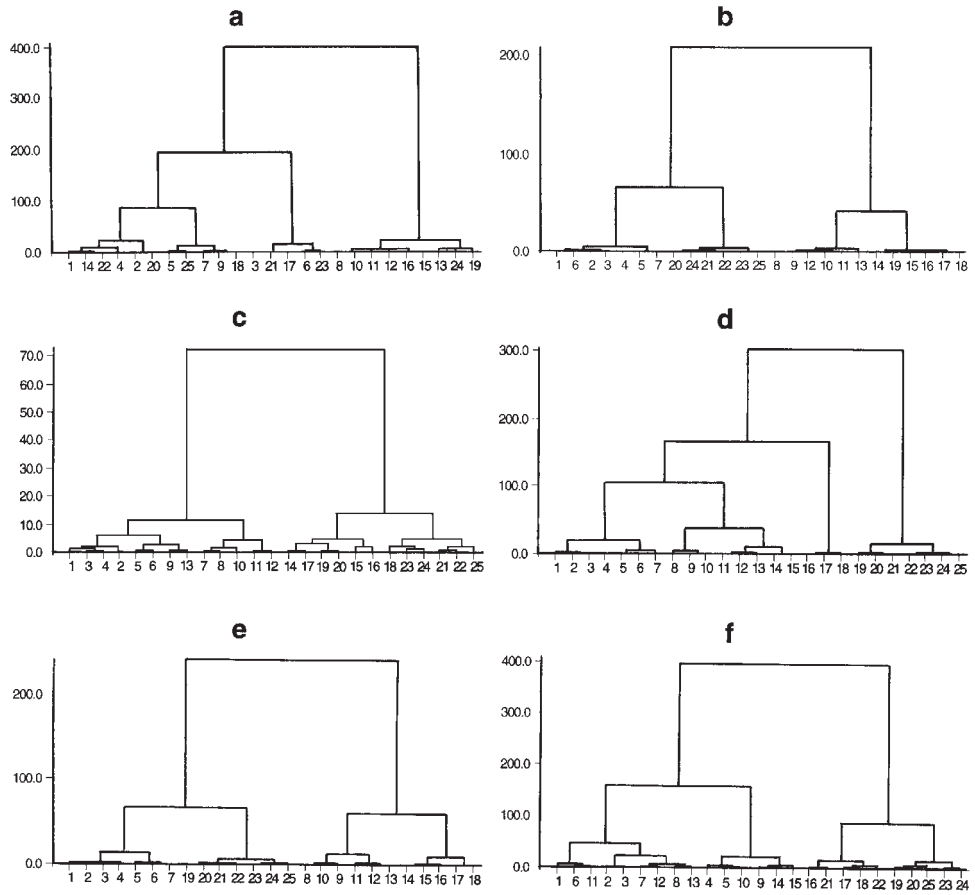
$$y_{h,ij} = \alpha_i y_{hi} + \alpha_j y_{hj} + \beta y_{ij} + \lambda_i y_i + \lambda_j y_j + \lambda_h y_h \qquad (5.4)$$

(a paraméterek konkrét értékeit az 5.2 táblázat tartalmazza). A másik egyenlethez képest itt fő újdonság, hogy minden egyes csoportnak, vagyis i-nek, j-nek és h-nak saját mérőszáma is van, ezek persze egy objektumra 0-val egyenlőek.

Eltérésnégyzetösszeg-növekedés minimalizálása (Ward 1963, Orlóci 1967, Wishart 1969). Ez a hierarchikus osztályozás talán legismertebb és legáltalánosabban alkalmazott módszere. A

5.2 táblázat. Egyes homogenitás-optimalizáló kombinatorikus módszerek főbb ismérvei és paraméterei. n_i=objektumok száma az i osztályban, $n_\cdot = n_i + n_j + n_h$, α_j, λ_h és λ_j nem szerepel külön a táblázatban, mert ezek α_i-hez, illetve λ_i-hez hasonló módon írhatók fel, $b_{hi} = \binom{n_h + n_i}{2}$, $b_i = \binom{n_i}{2}$ stb. Más módszerekről Podani (1989b) összefoglalásából informálódhatunk.

Módszer	α_i	β	λ_i	Kezdeti érték **Y**-ban	Reciprokpár algoritmus használható (+)
Eltérésnégyzetösszeg növekedés minimalizálása	$(n_h+n_i)/n_\cdot$	$-n_h/n_\cdot$	0	$d^2_{ij}/2$	+
Min. eltérésnégyzet az új osztályokban	$(n_h+n_i)/n_\cdot$	$(n_i+n_j)/n_\cdot$	$-n_i/n_\cdot$	$d^2_{ij}/2$	+
Variancia növekedés minimalizálása	$((n_h+n_i)/n_\cdot)^2$	$-n_h(n_i+n_j)/n_\cdot^2$	0	$d^2_{ij}/4$	–
Min. variancia az új osztályokban	$((n_h+n_i)/n_\cdot)^2$	$((n_i+n_j)/n_\cdot)^2$	$-(n_i/n_\cdot)^2$	$d^2_{ij}/4$	+
Minimum átlagos távolság az új osztályokban	$b_{hi}/b_\cdot$	$b_{ij}/b_\cdot$	$-b_i/b_\cdot$	d_{ij}	–



5.11 ábra. A pontmintázatok osztályozásai a eltérésnégyzetösszeg növekedését minimalizáló eljárással.

szakirodalom sokszor félrevezető cimkével illeti (pl. “minimum variance clustering”, ami semmiképpen sem lehet megfelelő név, s a variancia-alapon működő módszerek mindegyikét jelölhetné). Két csoport összevonásának az a feltétele, hogy az a lehető legkisebb eltérésnégyzetnövekedéssel járjon (az eltérésnégyzet-összeget a 3.105 formulával, de a 3.106-tal is kifejezhetjük). Formálisabban: A és B tehát összevonható, ha

$$\Delta SSQ_{(A+B)} = SSQ_{(A+B)} - SSQ_A - SSQ_B \qquad (5.5)$$

az összes lehetséges összevonásból a legkisebb. A stratégia megengedi, hogy egy ciklusban egyszerre több párt is összevonjunk.

Miután egy nagyon népszerű és könnyen megérthető módszerről van szó, a dendrogramot mind a hat példára bemutatjuk. Így látható igazán a fő különbség a távolság- ill. a homogenitás optimalizálása között. Az összehasonlítás nem ütközik akadályba, hiszen az eddigiekben – és itt is – a pontok közötti euklidészi távolság mérése adta a kiindulópontot. Ha az 5.11 ábra dendrogramjait megvizsgáljuk, akkor azonnal feltűnik, hogy a hierarchikus szintek növekedésének, a “lépcsők” emelkedésének szinte semmiféle jelentőséget nem szabad tulajdonítani.

A random esetben (**a**) és az igazán jól elváló négy osztály esetében (**b**) egyaránt tapasztalható ez a jelenség, amit inkább a módszer sajátjának, semmint az adatok szerkezetének kell betudnunk. (A mélyebb ok az, hogy az eltérésnégyzet-összeg igen gyorsan növekszik a pontok számának emelkedésével.) Persze a szintek változásának az "ütemében" van némi különbség, de ez csak egy összehasonlító elemzés során lesz nyilvánvalóvá. A **c** példában a két fő osztályt felismerjük, az elválás a 13. és 14. objektumok között van. Ugyanakkor a **d** esetben ez a módszer adta az eddigi "legjobb" eredményt, s majdnem egyértelműen kimutatta a három hosszú pontsereget. Az **e** példában viszont már nem kaptunk az előzőknél jobb hierarchiát. A szabályos esetre kapott **f** dendrogram pedig csupán megerősítheti a fentieket, miszerint a hirtelen növekedő szintek önmagukban semmit sem jelentenek e módszer esetében.

Fontos, hogy a dendrogramon a szintek ne a homogenitás megváltozását, hanem az összevonás után adódó új heterogenitás-értékeket jelentsék (még akkor is, ha az elemzés során a változást optimalizáljuk). Ez az ábrázolásmód jobban áttekinthető, s így a különböző módszerekkel kapott eredmények is összehasonlíthatóvá válnak.

Megemlítjük csupán, hogy a nem tárgyalt, de ide tartozó módszerek az eltérésnégyzet-összeg, a variancia és az átlagos osztályon belüli távolság kritériumainak és a két vizsgálati feltételnek (homogenitás-változás ill. új osztályok homogenitása) a kombinációi. Ezek közül az átlagos távolság (v. különbözőség) érdemel leginkább figyelmet, mert szinte bármilyen szimmetrikus függvényt használhatunk[5]. Az eltérésnégyzet-összeg és a variancia esetében némileg meg vagyunk kötve, hiszen itt feltétel az adatok átlagolhatósága (gondoljunk a formulákra). Itt is van egy flexibilis, változtatható paraméterű módszer, amely $\lambda=0$ értékeire lánchatást, míg λ egyre negatívabb értékeire a lépcsőhatást produkálja, s ilymódon osztályozás-sorozatok generálására alkalmas. Az 5.2 táblázat adja meg a rekurziós formula paramétereit és a mátrix kezdőértékeit (részletesebben l. Podani 1989b). Önmagukban mindenesetre ezeket a módszereket már kevéssé ajánljuk, mert tulajdonságaikat még nem ismerjük eléggé. Más, általánosan alkalmazott osztályozó módszerek mellett, azok kiegészítőjeként azonban számításba jöhetnek.

5.2.3 Homogenitás-optimalizáló nem kombinatorikus módszerek.

Ha az osztályok homogenitását információelméleti mérőszámokkal fejezzük ki, akkor az 5.3a ábra szerinti osztályozási stratégia alkalmazható csak (legalábbis mindeddig még nem ismeretes a kombinatorikus megoldás). A prezencia/abszencia esetre a 3.112 és a 3.115 formulák közül választhatunk. E függvények általánosításával a többállapotú nominális változókra leginkább alkalmas osztályozó módszerhez jutunk. A mérőszám és az algoritmus négy lehetséges kombinációjából a legismertebb a *súlyozott entrópia növekedését minimalizáló* eljárás, amelyben *A* és *B* összevonásának a feltétele, hogy a

$$\Delta H_{(A+B)} = H_{\{A+B\}} - H_A - H_B$$

mennyiség minimális legyen ("information analysis", Williams et al. 1966). A dendrogram szintjei célszerűen mindig az új osztályok homogenitásértékei. Akár az új osztályok homogenitását, akár annak változását optimalizáljuk, a hierarchikus szintek szükségszerűen monoton növekednek. E módszerek algoritmikus sajátságai sajnos kevéssé ismertek, pl. nem bizonyított

5 Eredetileg ezt Anderberg (1973) az adattárolást is igénylő módszerek közé sorolta, s később derült ki, hogy az osztályozó folyamatnak kombinatorikus változata is van (Podani 1979a).

(bár úgy tűnik), hogy a legközelebbi pár és a reciprok-pár algoritmusok azonos eredményre vezetnek (tehát az utóbbit célszerű használni).

5.2.4 Globális optimalizálás

Az eddigiekben ismertetett módszerek az objektumok (ill. osztályok) páronkénti összevonásához egy lokális kritériumot alkalmaztak, s nem voltak tekintettel arra, hogy mi történik az osztályozás egészével. Különösen igaz ez a távolság optimalizálásakor (az 5.1 táblázat módszerei), melyeknél az egymáshoz legközelebbi két osztály mindenképpen egy csoportba kerül. Ez a közelség (lokális optimum) nem biztos, hogy egybeesik az összes osztályra nézve kedvező megoldással (globális optimum). Ahhoz, hogy ezt a problémát jobban megérthessük, be kell vezetnünk valamilyen függvényt, ami az osztályozás egészének a "jóságát" méri s ezáltal alkalmas lesz a globális optimum megkeresésére. Számtalan lehetőség kínálkozik erre, de most csak egy függvényt vizsgáljunk meg, ami egyszerű, könnyen megérthető, és már szerepelt is az előző fejezetben. Az osztályokon belüli ill. az osztályok közötti távolságok átlagának a hányadosáról van szó (4.2-4.4 képletek), melynek előnyei a nem-hierarchikus osztályozásban – csak emlékeztetőül – 1) egyidejűleg figyelembe veszi a kohéziót és a szegregációt, 2) relatív mennyiség, így különböző osztályozások mérőszámai összehasonlíthatók egymással, és 3) gyakorlatilag bármilyen különbözőségi koefficiens használható. Ezek az előnyök a hierarchikus osztályozásban is érvényesülnek, ha az osztályozást a következőképpen folytatjuk le (Podani 1989a, vö. az 5.3c ábra sémájával):

1) Az **X** adatmátrixból előállítjuk az objektumok közötti távolságok (különbözőségek) **D** mátrixát.

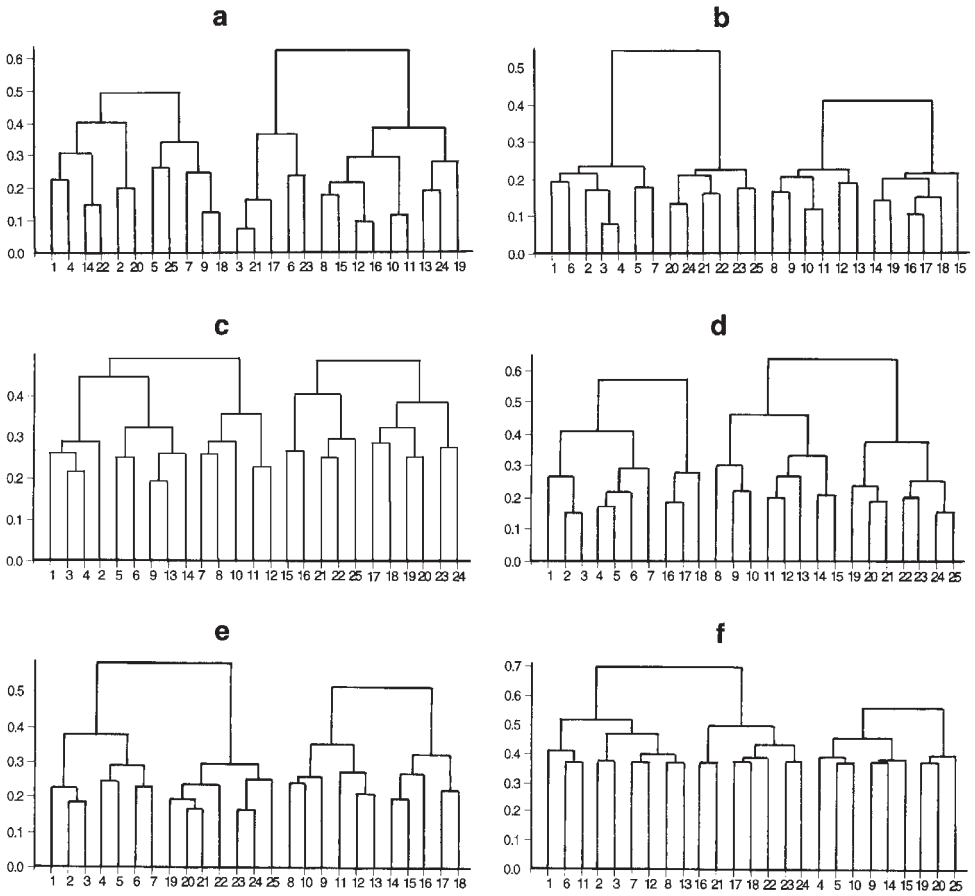
2) A **D** értékeiből kiszámítjuk, hogy az *i* és a *j* összevonásának hatására hogyan alakul a *G* mérőszám (4.4 formula) az egész osztályozásra nézve. Ezt minden párosításban meghatározzuk, s ezek az értékek kerülnek az 5.3c ábrán általánosan **G**-vel jelölt második szimmetrikus mátrixba.

3) Azt a párt vonjuk össze egy osztályba, melynek révén *G* értéke minimális lesz. A dendrogram ábrázolásakor ez a *G* érték jelenti a hierarchikus szintet. Csak a legközelebbi-pár algoritmus alkalmazható.

4) Ha kettőnél több az osztályok száma, a **G** mátrix értékeit kell átszámítanunk, s ehhez a **D** mátrix kell csupán (5.3c ábra). Ezután visszatérünk a 3. lépéshez. Ha csak két osztályunk maradt, további számításokra már nincs is szükség, mert ezek összevonása után a *G* nem számítható ki (nincsenek "külső" távolságok, tehát a nevezőnek nincs értelme *G*-ben). Emiatt a dendrogram "hiányos" lesz, a legfelső "gerenda" elmarad, ami – majd látni fogjuk – semmiben sem csökkenti az eredmény értékelhetőségét, hiszen a hierarchia így is teljes.

A többi agglomeratív módszerrel való összehasonlítás kedvéért ezt a módszert is alkalmaztuk a 4.3 ábra ponteloszlásaira. A dendrogramok az 5.12 ábrán láthatók. Most már érdemes lesz a függőleges tengelyen mért szintekre figyelni, mert ezek figyelembevételével is összevethetők az eredmények. Minél magasabb ugyanis a *G* értéke, annál rosszabbnak tekinthető az objektumok osztályozhatósága az adott számú csoportra. Egy dendrogramon belül pedig minél nagyobb az "ugrás" két szint között, annál kevésbé tekinthető az összevont két csoport értelmesnek[6]. Ezek után eredményeinket az alábbiakban összegezhetjük: Legjobb az osztályozhatósága (várakozásainknak megfelelően) a **b** eset négy csoportjának, a 0,2-nél

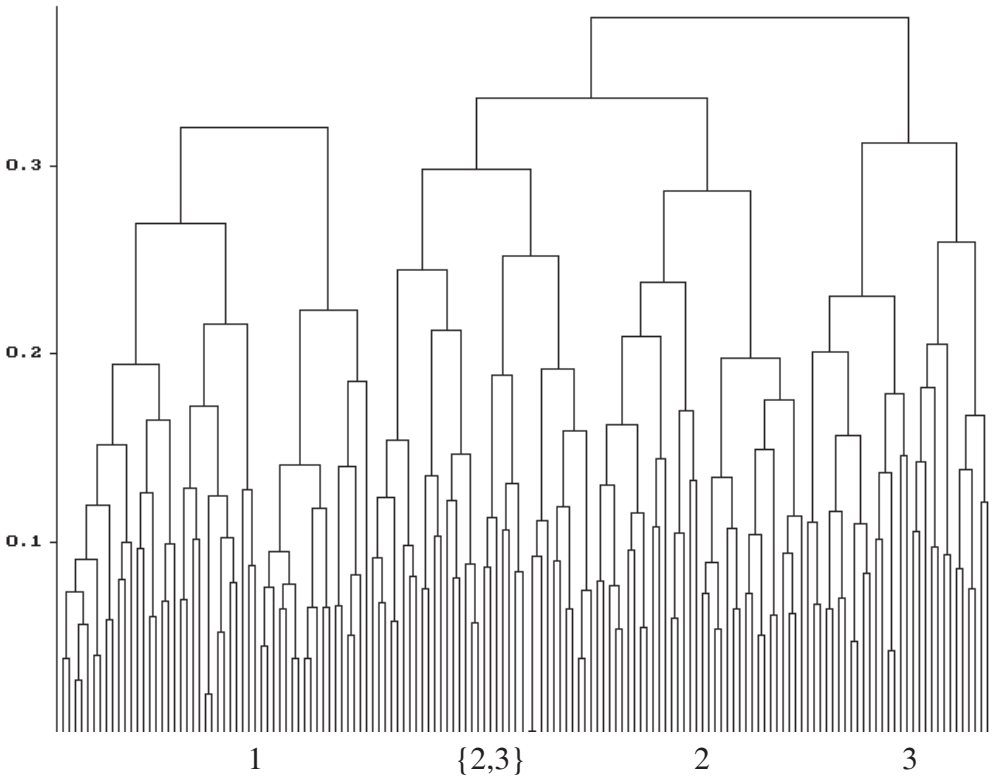
6 Ennél többet most nem mondhatunk. hiszen a hierarchiát értékelő eljárásokról csak később lesz szó (5.5 rész).



5.12 ábra. A belső és külső távolságok átlagának minimalizálása globális hierarchikus módszerrel a 4.3 ábra pontmintázataira.

alig magasabb szint és az ezt követő hirtelen emelkedés miatt. A **c** példa két osztálya is elválik egymástól (0,5 alatt), bár ez nem egyezik teljesen a nem-hierarchikus esettel (a 17. pont helyzetében, vö. 4.4 ábra). A **d** és **e** példákban a hosszú pontfelhők felismerése – hasonlóan sok más módszerhez – itt sem sikerült egyértelműen, s nem véletlen, hogy a kapott osztályok 0,5 körüli *G* értéket mutatnak. A random **a** eset jellemzője a szintek egyeneletesebb emelkedése, míg az abszolút osztályozhatatlannak tekintett **f** ponteloszlásra az első ill. az utolsó szint közötti kis intervallum és a relatíve magas első szintek a jellemzők.

A három *Iris* faj szétválását is megvizsgáljuk ezzel a módszerrel. Az A2 táblázat adataiból kiindulva, változónként a terjedelemmel standardizálva, az elemzés több órán át futott egy gyors PC-n is, ami mutatja az eljárás időigényességét (a távolság optimalizálás ugyanekkora adattömegre pár perc alatt lefut). Az eredmény részben igazolja csak az elválást. Kaptunk ugyan három teljesen "fajtiszta" osztályt, a negyedikben viszont keveredik a 2. és a 3. faj sok egyede (5.13 ábra). Ez igazolja a fuzzy osztályozással kapott eredményt, miszerint a két utóbbi faj nem határolható el élesen egymástól. Ugyanakkor a három faj relatív hasonlóságára utal a legfelső hierarchikus szint (0,37) alacsony volta és a hirtelen emelkedések hiánya. Az adatokat a későbbiek során más módszerekkel is megvizsgáljuk, s így alkalmunk lesz az összehasonlító értékelésre.



5.13 ábra. A három *Iris* fajhoz tartozó 150 egyed osztályozása a globális kritérium alapján. Az egyedek megszámozását – mivel úgysem látszana – elhagytuk, így csak a f•bb osztályokat jelöltük meg aszerint, hogy mely fajokat tartalmazzák (lásd a szöveget).

5.3 Divizív módszerek

E csoportba olyan eljárások tartoznak. amelyek az objektumhalmaz fokozatos kettéhasításával jutnak egyre kisebb méretű osztályokhoz. Számításigényük messze meghaladja az agglomeratív osztályozó eljárásokét, s ezért ritkábban használatosak. Egyesek a biológiai osztályozásban kiemelten fontosak, kifejezetten biológiai céllal fejlesztették ki őket a számítógépes adatelemzés "hőskorában" (60-as évek eleje) s ezért mindenképpen szólni kell róluk.

5.3.1 Politetikus módszerek

Tipikus – és egyben klasszikus – példa az Edwards & Cavalli-Sforza (1965) javasolta procedúra (lásd még Scott & Symons 1971), amely az A objektumhalmaz kettéválasztását $A1$ és $A2$ osztályra akkor hajtja végre, ha a

$$\Delta SSQ_A = SSQ_A - SSQ_{A1} - SSQ_{A2} \qquad (5.6)$$

mennyiség maximális. Szavakban megfogalmazva: a kapott új osztályok eltérésnégyzetösszegének a lehető legnagyobb mértékben kell csökkennie a felosztás során. A javaslat szerint minden lehetséges divíziót meg kell vizsgálni ahhoz, hogy a legkedvezőbbet kiválaszthassuk. Erre nagyon nagy objektumszám esetén gyakorlati lehetőségünk nincsen,

még a leggyorsabb számítógépek sem tudnak segíteni mert a kipróbálandó lehetőségek száma csillagászati (gondoljunk vissza, hányféle módon lehet *m* objektumot két osztályba sorolni, l. a 4.17 formula alatti részt). Az Edwards - Cavalli-Sforza-módszer legfeljebb 20-30 objektum esetén alkalmazható. A probléma megoldására egy – nagy adatmátrixokra azonban csupán megközelítőleg optimális eredményt adó – módszert ismertet Chandon et al. (1980) "branch and bound algorithm" néven. Ennek gyakorlati alkalmazása – úgy tűnik – még várat magára.

Osztályozás ordinációk alapján. A politetikus osztályozó módszerek egy csoportja nem közvetlenül osztályoz, hanem először egy ordinációt készít, s ennek alapján végzi el a divíziókat. E módszerek alapelvének megértéséhez előre kell lapoznunk tehát egy kicsit ebben a könyvben. Legismertebb az ún. TWINSPAN ("Two-Way INdicator SPecies ANalysis, Hill 1979a) módszer, melyre egy másik, a szerző által is említett – de el nem terjedt – néven is hivatkozhatnánk ("dichotomizált ordinációs elemzés"), ami jobban utal az ordinációs "alapra". A TWINSPAN elsősorban az növényökológiai-cönológiai osztályozások kedvelt módszere. A korreszpondencia-elemzés (7.3 alfejezet) során nyert első tengelyen, amely a variancia legnagyobb részét magyarázza, kiszámítjuk az objektumok súlypontját. Az ettől "jobbra" ill. "balra" eső objektumok alkotják a két fő csoportot, amelyek – különböző szempontok szerint – tovább finomíthatók. Az elemzés ezután a kapott csoportokban külön-külön folytatódik. Mivel a módszer célja a változók (rendszerint fajok) egyidejű osztályozása v. ordinációja is, és ezáltal egy átrendezett táblázat készítése, a későbbiek során visszatérünk rá (8. fejezet).

Az ordinációs tengelyek szerinti divizív algoritmusok közül időben több is megelőzte a TWINSPAN módszert, mégse vált igazán ismertté. Lefkovitch (1976) mutatott rá arra, hogyha az ultrametrikák **E** mátrixából (5.5.1 rész) főkoordináta módszerrel ordinációt állítunk elő (7.4.1 rész), akkor a – fontosságuk szerint sorrendbe állított – tengelyek mentén a negatív, ill. a pozitív régióba eső objektumok éppen az egyes hierarchikus szinten kialakuló dichotómiákat tükrözik. Nos, akkor miért is ne járhatnánk el fordítva, azaz a távolságok **D** mátrixát elemezzük először ezzel az ordinációs módszerrel, s az objektumok koordinátáinak előjele alapján állítsuk elő a hierarchiát: az első divíziós lépést az első tengely mentén, a következőt a második tengely mentén, és így tovább[7] . Williams (1976) számol be a POLYDIV eljárásról, amely sok tekintetben emlékeztet az előzőre, csak főkomponens elemzésen alapszik és a tengelyek mentén nem az előjelváltás szerint osztja ketté az objektumhalmazt, hanem azon a ponton, ahol az eltésnégyzet-összeg csökkenése maximális. A történeti hűség kedvéért megemlítendő, hogy ez az elképzelés korábban, Lambert et al. (1973) módszerében már megjelent.

5.3.2 Monotetikus divíziók

A monotetikus divizív módszerek – egy időben legalábbis – igen népszerűek voltak, és gyakorlati alkalmazásuk sem ütközött nagyobb korlátokba. Kiemelendő ezek közül a biológusok előtt is jól ismert *asszociáltság-analízis* nevű módszer(család), amely csak bináris (tehát prezencia/abszencia) típusú adatok feldolgozására alkalmas. Goodall (1953) munkásságát továbbfejlesztve a módszer első működőképes változatát Williams & Lambert (1959, 1960)

7 Lefkovitchnak kétségei voltak, hogy a módszer valójában nem is politetikus, hanem inkább monotetikus. Mivel azonban nem egy konkrét változón, hanem egy absztrakt, a változókat mintegy összesűrítő tengelyen alapul az elválasztás, a módszert nyugodtan sorolhatjuk a politetikusak közé.

dolgozta ki. Az asszociáltság-analízis alapgondolata az, hogy a változók közül kiválasztjuk azt, amelyik a legnagyobb mértékben "asszociálódik" a többivel (más szóval, amelynek megismerésével a legtöbbet tudjuk meg a többiről), Eme változó prezenciája és abszenciája szerint felosztjuk az objektumhalmazt, majd a részhalmazokban újabb divizív változók keresésével finomítjuk tovább – elméletileg az objektumokig lebontva – a klasszifikációt. Az elemzés kulcsa a kérdéses változó kiválasztását célzó függvény meghatározása. Kezdetben a változók között minden párosításban kiszámították a χ^2 értékét (lásd a a 3.14-15 formulákat), és ezeket összegezték minden változóra. A divízió kritériuma tehát következő volt:

$$\max_i \sum_{j=1}^{n} \chi_{ij}^2, \quad i \neq j. \tag{5.7}$$

E függvény hátránya, hogy a 2×2-es kontingencia-táblázat alacsony cellagyakoriságaira nem használható, s ezért elég sok változót ki kellett hagyni az elemzésből. A megoldás kétféle lehet, amelyek azonban nem feltétlenül vezetnek azonos eredményre. Az egyik szerint (Podani 1979b) a χ^2 függvényt egyszerűen a megfelelő információelméleti függvénnyel helyettesítjük, amely a 2×2-es kontingenciatábla jelöléseit alkalmazva a következő:

$$I = m \log m + a \log a + b \log b + c \log c + d \log d - (a+b) \log (a+b) - (a+c) \log (a+c) - (b+d) \log (b+d) - (c+d) \log (c+d) \tag{5.8}$$

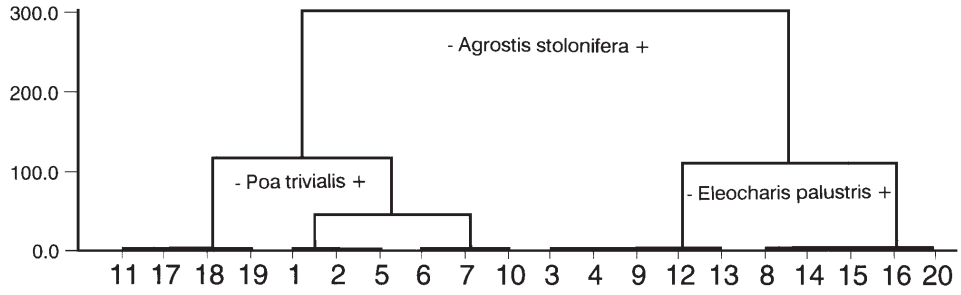
(változók "kölcsönös információja", $m=a+b+c+d$ objektum alapján). Ez minden korlátozás nélkül kiszámítható, és már csak a $2I \rightarrow \chi^2$ közelítés miatt (vö. Kullback 1959, Orlóci 1978) is jól pótolhatja a χ^2-et. Ezek után már minden a Lambert & Williams javasolta módon fut: a kölcsönös információ értékeket minden változóra összegezzük, s kikeressük a maximumot. Eszerint kettéválasztjuk az objektumhalmazt, majd újabb divizív változókat keresünk, a kapott két csoportban külön-külön. A felosztásokat célszerű egy T küszöbérték után abbahagyni, hiszen a hierarchia finomabb részletei amúgy is érdektelenek. Ritka esetben visszafordulások is jelentkezhetnek az eredményben, azaz a hierarchikus szintek nem csökkennek monoton módon.

A másik lehetőség a χ^2 kiváltására az, hogy megkeressük azt a változót, melynek prezenciája és abszenciája alapján kapott $A1$ és $A2$ osztályokra a

$$\Delta H_A = H_A - H_{A1} - H_{A2} \tag{5.9}$$

értéke, azaz az *entrópia csökkenése* maximális ("information fall", Williams et al. 1966, Lance & Williams 1968). H értéke a 3.112 formulával számítható ki. A dendrogram szintjei nem a változásokat, hanem célszerűen az osztályok entrópiaértékeit mutatják.

A monotetikus divizív módszerek kétségtelen előnye az eredmények közvetlen interpretálhatósága. A kiemelt változók segítségével az egyes csoportok közvetlenül jellemezhetők, akár egy "határozókulcs" is készíthető a csoportok elhatárolására, és új objektumok is beilleszthetők az osztályozásba (megfeledkezve most arról, hogy ekkor az új objektumok többé-kevésbé megváltoztathatják az asszociációs viszonyokat). A monotetikus felosztási elv hátránya ugyanakkor, hogy egy-egy objektum könnyen "félrecsúszik" az osztályozás során: bár minden tekintetben az $A1$ csoportra hasonlít, mégis az $A2$-be kerül, mert éppen a divizív változót tekintve tér el azoktól. A növénycönológiában például, amely az asszociáltság-elemzés legkiter-



5.14 ábra. A dünevegetáció (A4 táblázat) adataiból, az 5.9 kritériumra alapozott asszociáltság-analízis eredménye, a három legfontosabb divizív faj feltüntetésével. A függőleges tengelyen az entrópia-összeget mértük fel. Vessük össze a kapott klasszifikációt a 7.17 ábrával.

jedtebb alkalmazási területe, előfordulhat, hogy egy faj véletlenszerű okok miatt hiányzik egy adott mintaterületen, pedig a többi faj "jelzései" szerint ott kellene lennie. A monotetikus klasszifikációt ezért sokszor valamilyen utólagos relokációs procedúrával egészítik ki, amely kijavítja az ilyen félrecsúszásokat (Crawford & Wishart 1968, Weir 1970).

Az asszociáltság-analízis leginkább nagy objektumhalmazokra alkalmazható, hiszen ekkor jóval megbízhatóbbak a változók között számolt asszociáltság-értékek, mint kevés objektum esetén. A változók száma lehetőleg legyen nagyobb, mint az objektumok száma. A tengerparti dünevegetáció adatai nem felelnek ugyan meg ennek, de mégis jól szemléltetik a módszer lényegét. Mindkét információelméleti módszer ugyanarra az osztályozásra vezetett a legfelső szinteken, s így elegendő az egyiket bemutatni (5.14 ábra).

5.4 Speciális eljárások

Ebben az alfejezetben hierarchikus osztályozást eredményező alternatív, a fenti csoportosításba nem beilleszthető lehetőségekről szólunk. A következőkben leírtak csupán illusztrálják a tárgykör sokrétűségét: még ezzel együtt sem gondolhatunk arra, hogy a hierarchikus osztályozás témája akár közelítőleg is kimerült volna.

5.4.1. Kötött osztályozás

Bizonyos esetekben a kutatónak olyan osztályozásokra van szüksége, amikor az objektumok nem egészen a távolságviszonyok szerint kerülnek egy csoportba: a klasszifikációt valamilyen külső szempont korlátozza ("constrained classification"). Palinológusok és paleontológusok számára például fontos lehet, hogy csak időben egymást követő rétegek kerülhessenek együvé, s ilymódon a sztratigráfiai információ is kifejeződjék (vagyis: a rétegek sorrendje közvetlenül nem szerepel ugyan az adatokban, mégis hatással van az eredményre). Szerintük nem lenne értelme két távoleső réteget egy osztályba sorolni még akkor sem, ha azok fajösszetétele egyébként hasonló. De nemcsak az időbeliség, hanem a két- vagy háromdimenziós térbeli szomszédsági viszonyok is jelenthetnek ilyen megszorítást. A fentiekben tárgyalt módszerek módosításával elérhetjük, hogy az osztályozás megfeleljen a külső megkötésnek.

A módosítás lényege egy gráf bevezetése (legalábbis gondolatban), amelyben a szögpontok az osztályozott objektumok. Két szögpont között akkor van él, ha a megfelelő objektumok

időben (térben) érintkeznek vagy valamilyen más módon kapcsolódnak egymással. Az agglomeratív stratégiák alkalmazásakor e gráfot kell figyelnünk, amikor a távolságmátrixot vizsgáljuk. Csak azok az értékek jöhetnek számításba a mátrixban, amelyeknek él felel meg a gráfban, a többit egyszerűen figyelmen kívül hagyjuk. Két objektum összevonása után azok egy új szögpontként jelentkeznek a lassan zsugorodó gráfban, megtartva összes eredeti kapcsolódásaikat. Az egyszerű lánc (Gordon & Birks 1972) és az eltérésnégyzet-összeg növekedés minimalizálása (pl. Grimm 1987) a leggyakrabban alkalmazott agglomeratív módszer erre a célra.

A kötött osztályozás divizív módon is elvégezhető, bár ekkor csak a sorba állított (azaz időben vagy egydimenziós térben rendezett) objektumok osztályozása lesz problémamentes. Célunk az, hogy azt az élt távolítsuk el először a gráfból, amelyre a kapott két osztály eltérésnégyzet-összeg csökkenése maximális lesz. Ha az objektumok száma m, akkor esetünkben mindössze m–1 lehetőséget kell megvizsgálnunk (ennyi helyen vághatjuk ketté az objektumsorozatot). A kapott kisebb csoportokat ezután ugyanígy hasogathatjuk tovább, nyilván egyre kevesebb és kevesebb lehetőséggel számolva. Prezencia/abszencia adatokra szóba jöhet még a súlyozott entrópiaösszeg (3.112) csökkenésének a maximalizálása is (Gordon & Birks 1972). Az persze elképzelhető, hogy nem is a teljes hierarchia, hanem csupán egy adott k számú csoportra történő nem-hierarchikus felosztás az érdekes számunkra. Ekkor a hierarchiából kapott csoportokat egy nem-hierarchikus – ugyancsak kötött stratégiájú – osztályozó módszerrel optimalizáljuk utólag (Birks & Gordon 1985).

5.4.2 Adaptív osztályozás

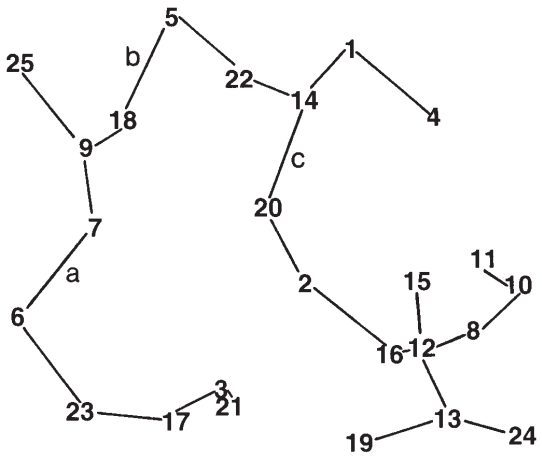
Az eddigiekben említett osztályozó módszerek mindenképpen produkálnak valamilyen végeredményt, függetlenül attól, hogy a módszer alkalmas-e voltaképpen az adatokban meglevő struktúra kimutatására. Mint a fenti példákban is láthattuk, a különféle módszerek az adatszerkezet más és más aspektusaira (más és más alakú csoportosulásokra) érzékenyek, s ettől függően egészen félrevezetőek is lehetnek, ha nem vagyunk eléggé óvatosak. Egy megoldás a módszerek párhuzamos alkalmazása és az eredmények összehasonlítása (erről még sok szó lesz), de vannak más elképzelések is. Miért ne lehetne egy olyan osztályozó algoritmust készíteni, amely valahogyan előzetesen feltárja a szóba jöhető tipikus eseteket, s ennek megfelelően, a felhasználó esetleges közbeavatkozásával alakítja át a klasszifikáció stratégiáját, azaz alkalmazkodik az adott szituációhoz? Számos próbálkozás ismeretes már eddig is, melyek közül Rohlf (1970) módszerét említjük meg először. A felhasználó előre megadhat bizonyos pontfelhő alakzatokat, amelyeket a programmal felismerhetünk. Az általánosított távolság (3.95) pl. az ellipszoid alakú csoportosulásoknak kedvez. Egy i pont és valamely j osztály távolságát úgy kapjuk meg, hogy a **W** mátrixot csak a j csoportban már benne lévő objektumok alapján számítjuk ki, ugyanakkor az l osztálytól való távolságát pedig az erre az osztályra számolt **W** mátrix figyelembevételével kapjuk meg. Más szóval, a már meglévBUBU osztályok belsBUBU struktúrája a döntBUBU, ebben az értelemben az osztályozás folyamata automatikusan "adaptálódik" a már meglévő csoportokhoz. Egy másik eljárás ("*mode analysis*", Wishart 1969) a zaj-elemek (vö. 4.6 ábra) kiszűrésével próbál előre nem rögzített alakú pontsűrűsödéseket felfedezni. Egy k egész számot és egy r sugarat kell megadnunk az elemzés elején, majd megvizsgálunk minden egyes pontot, hogy körülötte van-

e legalább *k* másik pont az *r* sugáron belül. Azokat a pontokat, amelyekre ez igaz, az egyszerű lánc módszerrel elemezzük, a többi, mint zaj-elem kikerül az elemzésből. *r* változtatásával egy osztályozássorozat generálható, amelyben az osztályok száma először növekszik, majd újra csökken (végül kellően nagy *r* esetén 1 lesz). Az adaptív osztályozásról további részleteket Sneath & Sokal (1973: 212-214) és Gordon (1981:137-139) könyvében találhatunk. Amint Gordon megjegyzi, előbb-utóbb lehetővé válhat majd olyan interaktív, az elemző ember döntéseit is igénylő módszer kialakítása, amely elszakadást jelenthet a teljesen automatizált klasszifikációs gyakorlattól.

5.4.3 Minimális feszítőfa

Az adatstruktúra feltárásában a dendrogramokon kívül más típusú gráfok is számításba jöhetnek. Először az ún. minimális feszítőfákat ("*minimum spanning tree*") kell megemlítenünk, melyek legszembetűnőbb eltérése a dendrogramoktól az, hogy minden szögpont megfelel egy objektumnak. Fa gráfról lévén szó körök persze nincsenek benne és – ami legalább ennyire lényeges – az objektumokat összekötő élek összhosszúsága minimális (Gower & Ross 1969, Rohlf 1973). A feszítőfa az objektumok távolságmátrixából állítható elő olymódon, hogy a minden lépésben az egymáshoz legközelebb eső két objektum között élt húzunk, ha ennek során nem keletkezik kör a gráfban. (A mátrix két legkisebb távolságértékéhez tehát eleve tartozik majd egy-egy él, a harmadikhoz már nem biztos.) A keresés az *m*–1-edik él behúzásával befejeződik, hiszen *m* pont összekötéséhez éppen ennyi élre van szükségünk. A minimális feszítőfának erőteljes a kapcsolata az egyszerű lánc módszerrel: a gráf alapján divizív módon ugyanazt a klasszifikációt lehet létrehozni, mint az egyszerű lánc módszerrel. Ha az éleket szukcesszív módon, mindig a leghosszabbat kiválasztva eltávolítjuk a gráfból, akkor a keletkezett részgráfok az egyszerű lánc módszer csoportjaival azonosíthatók (Gower & Ross 1969).

Az 5.15 ábra mutatja be a 4.3a pontmintázatra illesztett feszítőfát. A három leghosszabb élt (a, b és c) megjelöltük, s így könnyen ellenőrizhetjük, hogy ezek eltávolításával megkapjuk az 5.6a ábra dendrogramjának főbb osztályait. Amíg azonban a dendrogramon nem látjuk, hogy mely objektumpárok "felelősek" voltaképpen az osztályok között kialakult távolságokért, a



5.15 ábra. A 4.3a ábra pontjaira illeszthető minimális feszítőfa.

minimális feszítőfa ezt jól láttatja. Az egyszerű lánc módszer mellett tehát ez a gráf is szóba jöhet az osztályozások során.

A minimális feszítőfa azonban elsősorban nem osztályozásra való. Kiemelten jelentős a szerepe viszont a kétdimenziós ordinációs struktúrák "ellenőrzésében" (Digby & Kempton 1987:99, Gordon 1981:155, Dunn & Everitt 1982:75, és mások): a fát az objektumok ordinációjára rávetítve kimutatható, hogy mennyire hiteles az egyes objektumpárok esetleges relatív közelsége, vagyis elegendő-e a két dimenzió minden távolság-reláció közelítőleg hű ábrázolásához, vagy pedig torzítások is jelen vannak (lásd még a 9.5.2 részt). Rohlf (1975a) szerint a fa hatásosan alkalmazható az objektumhalmazba nem illeszkedő, "outlier" vagy zaj-elemek kimutatására is.

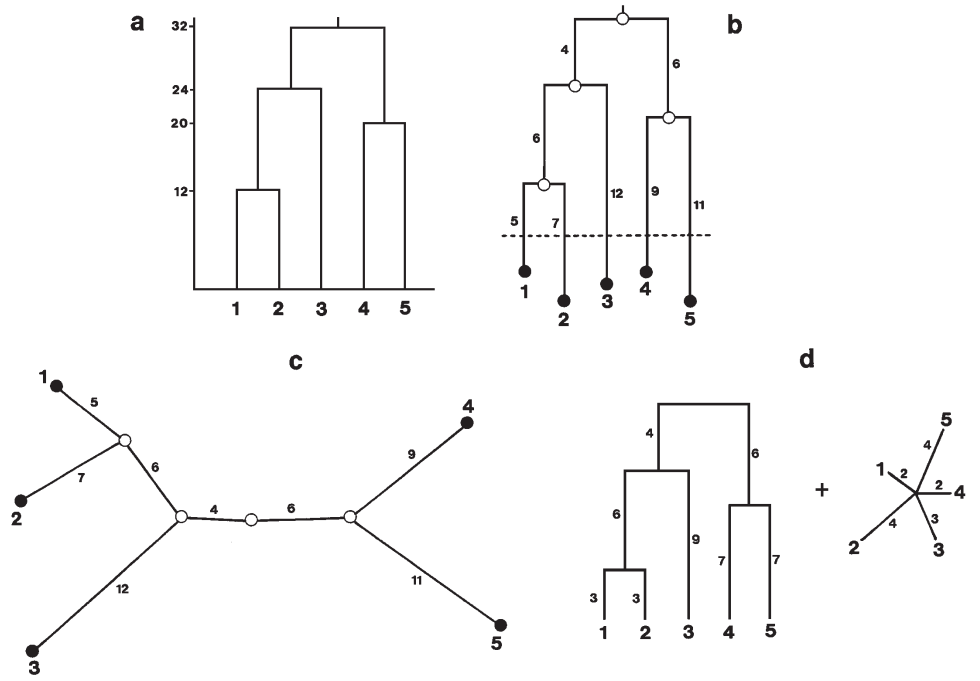
5.4.4 Additív fák

A dendrogramokat eredményező hierarchikus osztályozó módszerek és a minimális feszítőfa alkalmazása még nem merített ki minden lehetőséget a távolságmátrixok gráfokkal történő analízisére. Nem biztos például, hogy a dendrogramok szigorú hierarchiája jól tükrözi az objektumok távolságviszonyait is. Ennek megértéséhez tekintsük az alábbi távolság-félmátrixot 5 objektumra:

```
0,0 12,0 23,0 30,0 32,0
     0,0 25,0 32,0 34,0
          0,0 31,0 33,0
               0,0 20,0
                    0,0
```
(5.10)

Ebből kiindulva a csoportátlag módszerrel az 5.16a ábra dendrogramját kapjuk. Az 1. és 2. ill. a 4. és 5. objektumok közötti távolságokon ez a dendrogram még semmit sem torzít, de pl. az 1. és a 2. objektumok valamint az 1. és a 3. objektumok között már egyformán a 24,0-es távolságérték jelentkezik, holott ezek eredetileg 23,0 ill. 25,0 egységnyi távolságra voltak egymástól. Ezt a tényt persze elfogadtuk, amikor kifejezetten az osztályozás volt a cél. A dendrogram helyett – vagy inkább mellett – egy olyan *n*-fa is érdekes lehet azonban, amely a maximális mértékben megtartja az eredeti távolságértékeket minden objektumpárra. Nevén nevezve a dolgokat, az *additív fák*ról van szó, amelyek pl. a pszichológiai adatelemzésben régóta ismertek (Sattath & Tversky 1977, Shepard 1980, Pruzansky et al. 1982). A fenti mátrix az 5.16b ábra szerint rajzolható fel additív fa formájában, amely egy "egyenlőtlen lombozatú" dendrogramra emlékeztet. Egy kis vizsgálódással könnyen beláthatjuk, hogy bármely két pont távolsága hajszál pontosan kijön a közöttük lévő élek hosszának összeadásával (ez az összeg az ún. "*patrisztikus távolság*", Farris 1967). Mivel itt nem rajzolunk fel szinteket, voltaképpen a dendrogramszerű ábrázolást a szokványosabb gráf alak is helyettesítheti (5.16c ábra), amely már nem hangsúlyozza ki annyira az osztályozást. A következő fejezetben majd látni fogjuk, hogy az additív fák elsődleges funkciója valóban nem a klasszifikáció.

A fenti példa persze szándékosan sikeredett ilyen tökéletesre, a valóságban rendszerint a távolságmátrixok ábrázolása additív fa formájában nem valósítható meg maradéktalanul. Más szóval: a fán belüli távolságok egyáltalán nem biztos, hogy pontosan megfelelnek a távolságmátrix értékeinek, s valami "torzítás" itt is fellép, nemcsak a dendrogramban. Nem is könnyű feladat a fa megszerkesztése, így erre ehelyütt még fő vonalakban sem térünk ki. Az algoritmus iránt érdeklődőknek Sattath & Tversky (1977) cikkét ajánlhatjuk. Érdemes viszont az additív fák két tulajdonságáról egy kicsit többet szólnunk. Amíg a dendrogramokat leíró **E** mátrix – amint azt az 5.1 formula kapcsán elmondtuk – ultrametrikus jellegű volt, az additív



5.16 ábra. Az 5.10 mátrixból kapott csoportátlag-dendrogram (**a**) és additív fa (**b** ill. **c**), valamint az additív fa felbontása dendrogramra és "bokorra" (**d**).

fát leíró patrisztikus távolságok **A** mátrixa az ún. *négy-pont metrika* követelményeinek felel meg. Eszerint, az indexeléstől függetlenül, bármely négy pontra érvényes az alábbi összefüggés:

$$a_{hi} + a_{jk} \leq \max \{ a_{hj} + a_{ik} \ , a_{hk} + a_{ij} \} \qquad (5.11)$$

("*additív egyenlőtlenség*", Buneman 1971, Patrinos & Hakimi 1972, Sattath & Tversky 1977). A négy pontot egy tetraéder csúcsainak, a közöttük levő távolságokat (6 db) pedig a tetraéder éleinek tekintve a két-két szembenlévő él összegei egy egyenlőszárú háromszöget adnak. Ha egy **D** távolságmátrix eleve megfelel a fenti kritériumnak, akkor az közvetlenül egy patrisztikus mátrixnak tekinthető, egyébként viszont **D** csak közelíthető valamely **A** mátrixszal.

Az additív fák másik fontos tulajdonságát illusztrálja az 5.16d ábra. Minden additív fa felbontható egy dendrogram és egy csillag alakú fa (vagy "bokor") összegére: az additív fát az 5.16b ábrán szaggatott vonallal jelzett szinten elvágva dendrogramot kapunk, a levágott "végág-darabok" pedig egy bokorral ábrázolhatók (Carroll 1976). Ez valójában azt jelenti, hogy minden dendrogram additív fa, amelynek a bokor-komponense nulla, azaz az ultrametrikus egyenlőtlenség egy szigorúbb feltétel, mint az additív egyenlőtlenség. A különféle mértékekről ezek után a következő bennfoglalási sor állítható össze (Le Calvé 1985): *különbözőség* ← *metrika* ← *euklidészi távolság* ← *négy-pont metrika* ← *ultrametrika,* vagyis a legelső a legáltalánosabb, az utolsó pedig a legspeciálisabb mérték.

5.5 Hierarchikus osztályozások értékelése

A hierarchikus osztályozás során, mint azt sok példával is megpróbáltuk alátámasztani, nem elégedhetünk meg a puszta eredményekkel. További munkára van szükség: a kapott osztályozás finomabb elemzése, alternatív osztályozások összehasonlítása és az ordinációs módszerek alkalmazása ill. együttes értékelése a követendő stratégia. E fejezetben az első esetről, egy adott hierarchikus osztályozás elemzéséről, belső tulajdonságainak feltárásáról lesz csak szó. (Összehasonlításhoz legalább két eredmény kell, ezt a témát majd a 9. fejezetben részletezzük).

5.5.1 A torzítás mértéke

A hierarchikus osztályozások értékelésének több szempontja lehet. Először a fentiekben, az 5.4 részben már említett torzításra térünk ki. A dendrogramban felmért páronkénti távolságok, azaz az ultrametrikák, többé-kevésbé eltérnek az eredeti távolságértékektől. Egy osztályozó módszert ennek megfelelően annál jobbnak tekinthetünk, minél kisebb a változás a $\mathbf{D} \rightarrow \mathbf{E}$ irányban. E változás mérésére legrégebben a lineáris korrelációt (3.70 formula) ajánlották, mégpedig *kofenetikus korreláció* ("cophenetic correlation", Sokal & Rohlf 1962) néven. Ez egy speciális esete a később tárgyalandó mátrix összehasonlításoknak (9.2.1 rész), amelyekben két azonos méretű szimmetrikus mátrixot vetünk össze értékről-értékre. A **D** távolságmátrix és az abból hierarchikus osztályozással származtatott **E** ultrametrika esetében a korreláció a következőképpen írható fel:

$$COPH_{(\mathbf{D},\mathbf{E})} = \frac{\sum_{j=1}^{m-1}\sum_{k=j+1}^{m}(d_{jk}-\bar{d})(e_{jk}-\bar{e})}{\sqrt{\sum_{j=1}^{m-1}\sum_{k=j+1}^{m}(d_{jk}-\bar{d})^2 \sum_{j=1}^{m-1}\sum_{k=j+1}^{m}(e_{jk}-\bar{e})^2}} \qquad (5.12)$$

Az átló értékei tehát kiesnek az elemzésből, s a szimmetria miatt a számolást elegendő a félmátrix értékeire alapozni. A kofenetikus korreláció elsősorban a numerikus taxonómia kedvelt módszere, bár minden osztályozási célú vizsgálatban alkalmazható a távolságokat legkevésbé torzító stratégia kiválasztására. *COPH* értéke általában 0,6 és 0,95 közé esik, és rendszerint egy adott távolságmátrixra a csoportátlag módszerrel létrehozott ultrametrika adja a legjobb illeszkedést (Sneath 1966, Boyce 1969, Sokal & Rohlf 1970).

Mindezt ki is próbálhatjuk az egyszerű lánc, a teljes lánc és a csoportátlag módszer eredményeinek összehasonlításával, mondjuk a **b** és **c** példák alapján (az 5.6-8 ábrák megfelelő dendrogramjai). A **b** esetben, amikor is négy nyilvánvalóan jól elkülönülő osztályunk van, a legjobb eredményt a csoportátlag módszer adta (0,845), de nem sokban marad el mögötte az egyszerű lánc (0,839) és a teljes lánc módszer (0,830) sem. A két összeérő csoport – vagyis a **c** példa – esetében is ez a sorrend, hiszen a csoportátlag adta a legjobb illeszkedést (0,735), ennél kicsivel rosszabbat a teljes lánc (0,729) és sokkal rosszabbat az egyszerű lánc (0,426) módszer, mutatva ez utóbbi erős távolság-torzító hatását. Kimaradt az összehasonlításból a eltérésnégyzetösszeg-minimalizáló módszer, hiszen itt a hierarchikus szintek az eltérésnégyzet-összeget jelentik, s ezek korrelációja a távolságokkal – bár formailag kiszámítható ugyan – nem mérhető össze a fenti eredményekkel. A globális optimalizálásnál is fennáll az összehasonlíthatatlanság problémája, tetézve azzal, hogy a legfelső szint nincs meg, s emiatt még számolni sem tudnánk. A kofenetikus korreláció használatában tehát vigyáznunk kell az eredmények összehasonlíthatóságára, hogy ne jussunk teljesen megtévesztő következtetésekre.

A korreláció mellett más formulák is számításba jöhetnek, így például az euklidészi távolság négyzete (Hartigan 1967) vagy a Kruskal-féle (1964) *stressz-függvény* is. Ez utóbbit majd fő alkalmazási területe, a többdimenziós skálázás, ismertetésekor mutatjuk be (7.4.2 rész). A szintekkel kapcsolatos összes probléma pedig megoldható a *rang-korreláció* alkalmazásával (3.43 formula, Johnson 1967). Ez akkor jelez kis torzítást, ha az eredeti távolságok nagyság szerinti sorrendje közel áll az osztályozásbeli szintek sorrendjéhez. A korrelációkat, bármely formulával is számoltuk, nem szabad szokványos szignifikancia-próbának alávetni, hiszen a távolságmátrixon belül az egyes értékek nem függetlenek egymástól, és természetesen az ultrametrika sem független a távolságoktól (vö. 9. fejezet). Lényeges az is, hogy az értékek nem abszolút érvényűek, s leginkább csak egy adott vizsgálati kontextusban érvényesek. Azaz, egy 0,8-as korreláció nem biztos, hogy ugyanazt jelenti más és más adatstruktúrák esetében. További értékelési lehetőségeket ismertet Gower & Banfield (1975), ill. Gordon (1987).

5.5.2 Stabilitás és validitás

Egy hierarchikus osztályozástól elvárhatjuk, hogy az ne változzon meg jelentékenyen a kiindulási adatok kismértékű módosításának hatására ("stabilitás"). Ellenkező esetben erős kétségeink támadhatnak a kapott osztályozás érvényességét ("validitás") illetően. Ha a kapott eredmény stabilis, akkor joggal feltételezhetjük, hogy az osztályozás megfelelően összegzi az adatokban rejlő információt.

A stabilitás elemzésére számos lehetőség kínálkozik, ezek egy része a probléma matematikai, másik csoportja pedig a biológiai oldalát emeli ki. Megvizsgálható például, hogy az adatok vagy a távolságok mátrixának kismértékű véletlen megváltoztatása (random perturbációja) milyen változtatásokat eredményez a hierarchiában (pl. Rand 1971). Azt is érdemes lehet megnézni, hogy egy-egy objektum vagy változó elhagyása (vagy egy újnak a hozzáadása) megváltoztatja-e az osztályozást (pl. Jambu & Lebeaux 1983). A Smith & Dubes (1980) javasolta stratégia kissé komplikáltabb: az objektumhalmazt véletlenszerűen kettéosztják, és külön-külön is osztályozzák. Két formulát is javasoltak annak mérésére, hogy a "felezett" halmazban egy csoportba került objektumok mennyire maradtak együtt a komplett osztályozásban is. Tágabb értelemben stabilitási vizsgálat az is, ha ugyanazokat az adatokat alternatív módszerekkel osztályozzuk (bár itt nem az adatok módosulnak, hanem a módszer változik), és ha több módszer hasonló eredményt ad, akkor az osztályozás stabilisnak tekinthető (mint a **b** példa négy csoportja, vö. 5.6b, 5.7b, 5.11b és 5.12b ábrák). E példában persze viszonylag könnyű volt a dendrogramok hasonlóságát felmérni, általában viszont valamilyen kvantitatív módszert kell bevetnünk (lásd a 9. fejezetet).

Az osztályozások biológiai szempontok szerinti stabilitásáról Rohlf & Sokal (1981a) adja a legjobb összefoglalót. Egy taxonómiai klasszifikáció stabilitását értékelve igen fontos szempont, hogy történik-e változás ha különböző karaktercsoportokra alapozzuk az elemzést. Ilyen csoportok lehetnek pl. rovarok esetében a lárva és az imágóállapotra vonatkozó tulajdonságok csoportjai. Ilymódon ellenőrizhető az ún. "*non-specificity*" hipotézis (Sneath & Sokal 1973: 97), miszerint nincsenek elkülönült géncsoportok, amelyek a tulajdonságok egy elhatárolt csoportjára lennének kizárólagos hatással. (Más kérdés, hogy a legtöbb vizsgálatban ez nem bizonyult teljesen igaznak). Hasonló jellegű probléma az ökológiában a fajokra ill. a környezeti változókra alapozott klasszifikációk elemzése, bár itt nemcsak stabilitásról, hanem prediktabilitásról is beszélhetünk. Megvizsgálhatjuk azt is, hogy egy cönológiai osztályozás

megváltozik-e, ha a faji szintről áttérünk a genusz, család vagy esetleg a rend szintjére (Podani 1986).

5.5.3 Osztályok optimális száma

A particionáló módszerek túlnyomó többsége, amint az előző fejezetben is láthattuk, csak akkor használható, ha már van valamilyen elképzelésünk az adatokban rejlő osztályok számáról. A hierarchikus klasszifikáció algoritmusaitól – miután előzetesen semmilyen paramétert sem kell megadnunk – viszont már azt várjuk, hogy valamilyen erre vonatkozó információt is adnak. Más szóval: legalább sejtetni engedik azt, hogy melyik az a hierarchikus szint, ahol a dendrogramot "elvágva" optimális partíciót kapunk. Jelen fejezet példái azonban arra utaltak, hogy a dendrogram alakjából, a szintek növekedéséből stb. csak némi tapasztalattal, az algoritmikus sajátságok ismeretében következtethetünk az osztályok létére és számára; s igazándiból szinte sohasem lehetünk biztosak a dolgunkban. Valamilyen objektív eljárásra lenne tehát szükség, ami lehetővé teszi a dendrogramok alapján történő felosztást. Az osztályozási problémákkal foglalkozó kutatókat, csakúgy mint a biológus alkamazókat már meglehetősen régen izgatja ez a téma (vö. Dale 1988). Milligan & Cooper (1985) összefoglalójában 30- féle ilyen módszert dolgozott fel, s ez az áttekintés nem is mondható teljesnek. A legtöbb javaslat – amint azt előre is sejthetjük, hiszen szoros kapcsolatban állanak a klasszikus biometriai módszerekkel – az eltérésnégyzet-összegen alapszik. A fenti szerzők ismert tulajdonságú adatok osztályozásait kipróbálva azt találták, hogy a legtöbb esetben Calinski & Harabasz (1974) módszere jelezte az optimális osztályszámot. Jelölje SSQ_t az n változóval leírt m objektumot tartalmazó adathalmaz teljes eltérésnégyzet-összegét, amelyet a következő módon kaphatunk meg:

$$SSQ_t = \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \bar{x}_i)^2 \qquad (5.13)$$

ahol $\bar{x}_i$ az i változó átlaga. (Csak emlékeztetőül: egy A osztályra a 3.105 formula adta meg ugyanezt.) Az osztályokon belüli, azaz "belső" eltérésnégyzetek összegét is bemutattuk már a k-közép módszerrel kapcsolatosan (J, 4.1 képlet), amit most jelöljünk SSQ_b-vel. Az $SSQ_e = SSQ_t - SSQ_b$ különbség a teljes eltérésnégyzet-összegnek az osztályok közötti eltérésekre jutó része, ezt nevezhetjük "külső" eltérésnégyzet-összegnek. Minél nagyobb a külső eltérésnégyzet-összeg a belsők összegéhez képest, annál jobbnak tekinthető a felosztás, s rögtön adódik az ötlet, hogy képezzük az $SSQe$ és az SSQ_b mennyiségek hányadosát. Ez azonban még nem tenné lehetővé a k különböző értékeire kapott eredmények összehasonlítását, hiszen a megfelelő szabadsági fokokkal előzőleg osztanunk kell az eltérésnégyzet-összegeket. A Calinski & Harabasz javasolta kritérium ezek után k osztályra a következő alakot ölti:

$$CALHAR_k = \frac{SSQ_e}{(k-1)} \Big/ \frac{SSQ_b}{(n-k)} \qquad (5.14)$$

A szerzők szerint $CALHAR$ monoton növekedése k függvényében az osztályszerkezet teljes hiányára, monoton csökkenése pedig hierarchikus adatszerkezetre utal. Ha azonban maximuma van, akkor a maximumot adó k számú osztályt tekinthetjük az objektumhalmaz

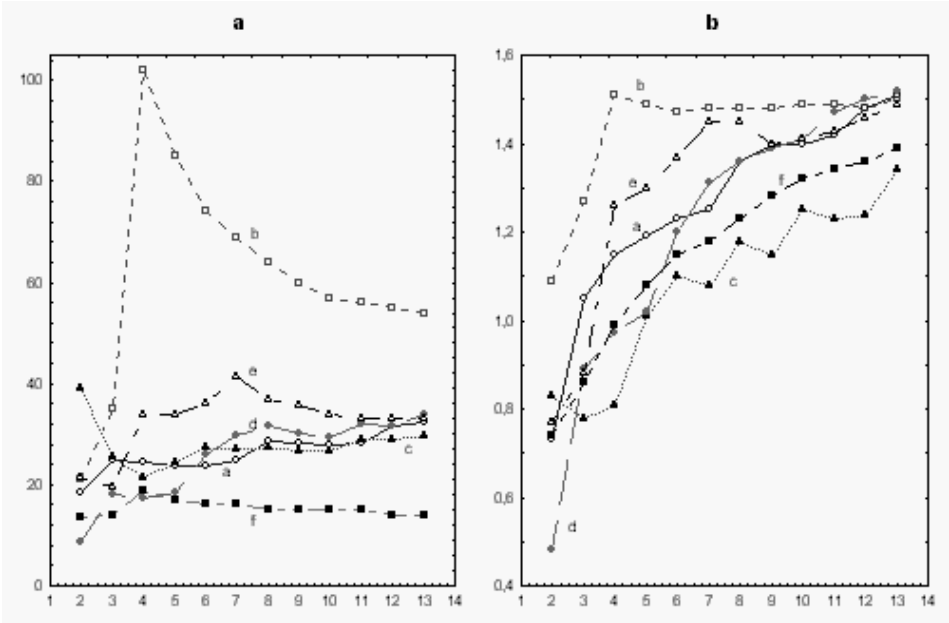
legoptimálisabb felosztásának (más szóval: *k*-t változtatva először az osztályszerkezet hiánya, az optimális értéket követően pedig már egy hierarchia megléte mutatható ki).

A 4.1.2 rész végén már alkalmazott mátrixokkal is felírhatjük a fenti kritériumot. Ha tehát **B** az osztályok közötti eltérésszorzat-összegek, **W** pedig az osztályokon belüli eltérésszorzat-összegek mátrixa, akkor nekünk csak az átlóba írt értékek kellenek, azaz

$$CALHAR_k = \frac{tr(\mathbf{B})}{(k-1)} / \frac{tr(\mathbf{W})}{(m-k)} \qquad (5.15)$$

(lásd a C függeléket). Vizsgáljuk meg a Calinski - Harabasz index "viselkedését", az interpretáció lehetőségeit a 4.3 ábra példamintázatainak teljes lánc módszerrel kapott osztályozásaira (5.7 ábra). Az osztályok száma 2-től 12-ig terjed, ennél tovább finomítani a felosztást már semmiképpen sem lett volna érdemes (5.17a ábra). Az **a** random esetre a függvény görbéje határozottan, de nem mindig monoton emelkedő. A hasonlóképpen osztály-nélküli **f** esetben éppen ellenkező a tendencia, enyhe csúccsal a *k*=4-nél, mutatva, hogy nem szabad bíznunk a kismértékben kiemelkedő maximumokban. A **b** példában annál inkább egyértelmű az eredmény, hiszen *k*=4-nél erősen kiugró maximumot kapunk, ami egyezik előzetes elvárásainkkal. Igen érdekes a függvény menete a két érintkező osztályt illusztráló **c** példában. Itt kifejezett maximummal "indulunk", majd hirtelen esés után a függvény értéke eléggé alacsony szint körül ingadozik. Eszerint az osztályozás a *k*=2 esetben a legjobb (ami igaznak is tűnik), majd minden további felbontás már csak lerontja az eredményt. A hosszú, illetve az ívelt pontfelhőket (**d** és **e** példák) a teljes lánc módszer egyáltalán nem ismerte fel, és így nem meglepő, hogy az egyébként nem kifejezetten erős maximumok *k*=8 ill. 9 körül mutatkoznak, azaz amikor a felosztás már eléggé finom ahhoz, hogy a kis csoportok ne fedjenek át az általunk feltételezett osztályok között.

Anélkül, hogy az Olvasót a többi, Milligan & Cooper (1985) áttekintésében szereplő (és nem szereplő) módszerrel terhelnénk, elmondhatjuk, hogy majdnem mindegyik interpretál-



5.17 ábra. Optimális felosztás keresése a 4.3 ábra példáira alkalmazott teljes lánc módszer elemzések alapján, **a**: Calinski-Harabasz index, **b:** rangsorolásos módszer.

ható geometriailag vagy legalább statisztikailag. Alkalmazásuk azonban olyan esetekre korlátozódik, amikor az eltérésnégyzet-összegnek értelme van. Sok esetben azonban az osztályozásra alkalmazott különbözőségi indexnek semmi köze az eltérésnégyzet-összeghez, így a Calinski-Harabasz-index használata értelmetlen. Egy általánosabb módszerre is szükség van tehát, amely nem is feltétlenül a geometriai értelmezhetőségre fekteti a hangsúlyt, hanem az értékelő módszer és az alkalmazott különbözőségi index összhangjára épít. Az alábbiakban egy ilyen eljárást mutatunk be röviden (Podani 1994). Ennek az az előnye, hogy minden egyes változóra megtudjuk, milyen mértékben magyarázza meg az adott osztályozást. Az alkalmazott rangsorolásos technika fontosnak találhat olyan változókat is, amelyek az osztályozás kialakulására eredetileg nem voltak nagy hatással. Intuitív előnye a módszernek talán az is, hogy a "többség dönt" elvre épül, azaz minél több változó "támogat" egy adott osztályozást, annál elfogadhatóbbnak tartjuk.

Az elemzés első lépésében a változók mennyiségi hozzájárulását kell kifejeznünk a k osztályon belüli ill. osztályok közötti távolságokhoz v. különbözőségekhez. Ennek kiszámítása szinte minden koefficiensnél más és más; az euklidészi távolság esetén például az i változó hozzájárulása a j és k objektum távolságához éppen $(x_{ij} - x_{ik})^2$-nel arányos; jelöljük ezt g_{ijk}-val. A következő feladat annak megvizsgálása, hogy e hozzájárulások miképpen oszlanak meg osztályokon belül ill. azok között. Az i változó teljes mértékben megmagyarázza a felosztást, ha minden osztályon belüli g_{ijk} kisebb, mint az osztály közöttiek, vagyis a g_{ijk} értékek nagyság szerinti sorrendjében az osztályon belüliek rangszámainak összege minimális. A változó teljesen közömbös is lehet az osztályozásra nézve, ekkor az osztályon belüliek ill. osztályközöttiek rangszámai véletlenszerűen oszlanak meg. Még az is előfordulhat, hogy egy változó kifejezetten ellentmond az osztályozásnak, s ekkor az osztályon belüli g_{ijk} értékek nagyobb rangszámot kapnak, mint az osztály közöttiek. Mindez egy ψ_{ik} mérőszám formájában összesíthető, amely 1-es értéket vesz fel, ha az i-edik változó teljes mértékben megmagyarázza a k osztályra történő felosztást. ψ_{ik}=0, ha a változó közömbös, és $\psi_{ik} < 0$, hogyha a változó ellentmondó. A ψ_{ik} értékek alapján a változók sorba rendezhetők, mutatva azok használhatóságát a k osztály értelmezésében.

Az eddig elmondottak voltaképpen a nem-hierarchikus osztályozásokra érvényesek, s csak a következő lépésben térünk rá a dendrogramok értékelésére. Az optimális osztályszám egy partíciósorozatban az lehet, amelyet a változók többsége támogat. Ez egyszerűen a ψ_{ik} értékek összegével mérhető, amelyet jelöljünk Ψ_k–val. Ennek maximális értéke nyilván n, vagyis a változók száma. Ψ_k változása felrajzolható k függvényében, s a görbe alakja informál bennünket a változók magyarázó erejéről. Bár e módszernek nyilvánvalóan nem geometrikus, hanem rendezési interpretációja van és elsősorban sok változó esetén hatékony, érdemes a Calinski-Harabasz-indexszel összehasonlítani (5.17b ábra). Azonos következtetésre juthatunk e két módszer alapján a **b** esetben, bár a csúcs kevéssé kifejezett az ezt követő kismértékű csökkenés miatt. Az **a** és **f** példákban végig monoton növekvő a függvény, ami az osztályozhatatlanság jele. Ugyanez figyelhető meg a **d** példára is, mutatva, hogy a teljes lánc módszerrel kapott osztályokkal valami "nincs rendben". Az **e** esetre itt is a k=7 mellett kapunk maximumot, csakúgy mint az 5.17a ábrán. A **c** példában viszont megmutatkozik, hogy a módszer kevés változóra nem mindig működik: elmarad a csúcs, tehát a Calinski-Harabasz-index adott jobban értelmezhető eredményt.

5.6 Irodalmi áttekintés

A hierarchikus osztályozás népszerű mind a biológiában, mind pedig számos humán tudományágban (pl. pszichológia, szociológia) és újabban a matematikusok is kiemelten foglalkoznak e témával. Ezt a sok országban megalakított klasszifikációs társaságok, valamint az őket egyesítő *International Federation of Classification Societies* léte is igazolja. (Nálunk

még nincs ilyen szerveződés, pedig ez jelentékenyen megkönnyítené a hazai kutatók közötti információcserét). A *Journal of Classification* című, 1984 óta megjelenő folyóirat is igazolja a témakör nem szűnő népszerűségét. Az irodalom óriási és egészében ma már teljességgel áttekinthetetlen (Blashfield & Aldenderfer 1978 adott az azelőtti időszakra érvényes összefoglalót, azóta ez nem nagyon megy). Ennek ellenére meg lehet adni azokat a fontosabb forrásokat, amelyeket minden biológus haszonnal vehet, ha a finomabb részletekre kíváncsi.

A téma matematikai vonatkozásairól kezdetben csak összefoglaló cikkek láttak napvilágot (Cormack 1971, Williams 1971), s lényeges áttörést Jardine & Sibson (1971), Anderberg (1973), Everitt (1974), majd Clifford & Stephenson (1975) könyvei jelentettek. Everitt műve már a harmadik átdolgozott kiadást is megérte. További alapmunkák: Späth (1980), Gordon (1981), Aldenderfer & Blashfield (1984), Romesburg (1984), Jambu & Lebeaux (1983). Feltűnő, hogy a többváltozós adatelemzés "hivatalos" szakirodalma az osztályozást egy kis – majdnem jelentéktelen – mellékterületként kezeli (pl. Mardia et al. 1979) vagy még úgy sem. Az osztályozó módszerek tulajdonságait elemző cikkek közül megemlítendő Diday & Simon (1976), Dubes & Jain (1976, 1979), Murtagh (1983), Day & Edelsbrunner (1984), Gordon (1987), Milligan (1989), és különösen érdekes Everitt (1979) problémafelvető cikke. Az újabb fejlemények tömör áttekintését Gordon (1996) adja.

A biológia két ága nevezhető kifejezetten osztályozó-"mániás"-nak, a rendszertan és a cönológia, mert itt sok esetben a vizsgálat végső célja az objektumok hierarchikus viszonyainak feltárása. Ha a jelen fejezetben taglalt számítógépes módszereket alkalmazzuk, akkor numerikus taxonómiáról ill. numerikus szüntaxonómiáról beszélünk. A numerikus taxonómia kezdetét nyugodtan számíthatjuk Sokal & Sneath (1963) alapművétől, melynek második kiadása (Sneath & Sokal 1973) mindmáig nagy haszonnal forgatható. Sajnos további kiadások nem jelentek meg, bár a leglényegesebb fejleményekről kisebb közleményekből tájékozódhatunk (pl. Sokal 1986). Miután a numerikus taxonómia alapvetően a fenetikai hasonlóság mérésén alapszik, s nem szándékozik ábrázolni a leszármazási viszonyokat, mára kissé háttérbe szorult a kladisztikához képest (következő fejezet). Stuessy (1990) növénytani alapműve ugyan még egyenrangúnak tekinti a kettőt, az állattanban iránymutató Mayr & Ashlock (1991) féle tankönyv viszont eléggé egyértelműen elveti azon hierarchikus módszerek alkalmazását, amelyek a leszármazás elemzésére nem alkalmasak. Erre a problémakörre részletesebben a következő fejezetben visszatérünk. A numerikus taxonómiát tárgyaló egyéb művek pl. Cole (1969) és Dunn & Everitt (1982). Pankhurst (1991) az osztályozó módszerek és az adatbázisok ill. a határozó-kulcsok összefüggéseit is részletesen megvizsgálja. A numerikus taxonómia immár több, mint 30 éves történetét Sneath (1995) foglalja össze röviden, kiemelve annak döntő szerepét a későbbi taxonómiai irányzatok (kladisztika, "új" morfometria) megalapozásában és kialakításában.

A numerikus szüntaxonómia és általában az ökológia osztályozási problémáit számos könyv elemzi kisebb-nagyobb terjedelemben. A rendszertan fenetika-kladisztika ellentétpárjának analógiája itt is megvan (klasszifikáció *versus* ordináció), de nem olyan kiélezett formában így az alábbi művek a hierarchikus osztályozás és az ordináció szempontjából egyaránt fontosak. Whittaker (1973), Williams (1976), Orlóci (1978), Gauch (1982), Greig-Smith (1983), Kershaw & Looney (1985), Legendre & Legendre (1983), Digby & Kempton (1987), Jongman et al. (1987), Ludwig & Reynolds (1988) és Kent & Coker (1992) csupán csak kiragadott példák a bőséges irodalomból. Green (1979) azoknak ajánlható, akik biológiai és környezeti adatok együttes értékelését tervezik. Klijn (1994) pedig az ökoszisztéma-szintű klasszifikációkhoz jelent kiindulási alapot. Az információelméleti osztályozás módszereiről Feoli et al. (1984) ad összefoglalót. Két megtekintésre ajánlható cikkgyűjtemény Mucina & Dale (1989) és Feoli & Orlóci (1991). A "review" cikkek közül kiemelhető Maarel (1979) és Gauch & Whittaker (1981).

5.3 táblázat. A hierarchikus klasszifikáció könyvünkben is tárgyalt legfontosabb módszerei különféle, személyi számítógépekre kidolgozott programcsomagokban.

Módszer	**Statistica**	**BMDP**	**NT-SYS**	**SYN-TAX**	**NuCoSA**
Egyszer•/teljes lánc	+	+	+	+	+
Csoportátlag	+	+	+	+	+
Centroid	+	+	+	+	+
β-Flexibilis			+	+	+
Eltérésnégyzet-összeg min.	+			+	+
Információelméleti módsz.				+	
Globális optimalizálás				+	
Monotetikus divizív m.				+	
Minimális feszít•fa			+	+	
Kofenetikus korreláció			+	+	
Optimális osztályszám				+	
Dendrogram grafika	+	+	+	+	+
Feszít•fa grafika			+	+	

5.6.1 Számítógépes programok

A hierarchikus osztályozás módszereivel számos programcsomagban találkozhatunk, bár többnyire csak a legfontosabb és a legismertebb módszerek szerepelnek bennük. Az 5.3 táblázat azokat a programcsomagokat tartalmazza, amelyek már megfelelő felhasználó-barát környezetet teremtenek személyi számítógépeken. E programcsomagok viszont alkalmatlanok lehetnek bizonyos speciális problémák megoldására. Ekkor fordulhatunk azokhoz a cikkekhez és könyvekhez, amelyek programlistákat is közölnek (Anderberg 1973, Späth 1980, Orlóci 1978), bár a listázott program adaptálása saját számítógépünkhöz igen fáradtságos lehet. Blashfield (1976) foglalta össze a korábban használatos legfontosabb számítógépes szoftvereket, amelyek természetesen nagy gépekre voltak alkalmasak. A **CONISS** program (Grimm 1987) az eltérésnégyzet-összeg növekedés minimalizáló eljárás kötött változata. A **TWINSPAN** program listáját Hill (1979a) közli.

5.7 Kérdezz - Válaszolok!

K: *A sok példából úgy tűnik számomra, hogy az osztályozó módszerek végül is – önmagukban – a legkevésbé sem arra valók ami a nevükből következne, azaz osztályozásra!*

V: Valóban, a helyzet kissé paradox. Egyetlen egy számítógépes osztályozó (pontosabban "clustering") módszer sem ajánlható egyes-egyedüli módszerként erre a célra. Az osztályozó algoritmusok fő funkciójának ma már az adatok struktúrájának a feltárását tekintjük. Sokféle ilyen információból áll azután össze a végső kép, amit osztályozásnak nevezhetünk. Arról van tehát szó, hogy – ellentétben a mondjuk 20 év előtti várakozásokkal – nincs kizárólagosan üdvözítő módszer ami helyettesítené saját óvatos és józan megítélésünket (Dunn & Everitt 1982). A taxonómus például "sajnos" nem elégedhet meg azzal, hogy adatait betáplálja egy számítógépbe, hogy az majd kiadja az általa vizsgált szervezetek "objektív" és megcáfolhatatlan osztályozását. Az az időszak is elmúlt, amikor a folyóiratok kapva-kaptak az olyan kéziratokon, amelyek hemzsegtek az ilyen módszerek – rendszerint kritika nélküli – alkalmazásaitól. Az osztályozásban voltaképpen legalább olyan szerepe van a csoportszer-

kezetet közvetlenül nem vizsgáló algoritmusoknak (pl. ordinációk, szeriálás, összehasonlítások, stb), amelyekkel majd csak ezután fogunk megismerkedni.

K: *Az egyes módszerek annyira eltérően viselkedtek, s oly sokszor adtak el nem fogadható eredményt, hogy szinte el is vetted a kedvemet a hierarchikus osztályozó algoritmusoktól!*

V: No, azért nem kell annyira elkedvetlenedni, könnyen így lehet ez a később tárgyalandó módszerek esetében is. Csupán arról van szó – még egyszer hangsúlyozom –, hogy nagyon sokféle módszer van, és nagyon sokféle adatstruktúra-típus is lehetséges.

K: *Tulajdonképpen érdekes-e egyáltalán maga a hierarchia? Mintha erről megfeledkeztél volna a példák során, s csak azt nézted, hogy a hierarchikus osztályozás alkalmas-e bizonyos nem-hierarchikus osztály-struktúrák detektálására. Nincs itt valami ellentmondás?*

V: Igazad van. Olyan példánk tényleg nem volt, amelyben hierarchikus csoportosulást rejtettünk el. Talán nem lett volna eléggé szemléletes, s ugyanakkor sokkal többet nem láttunk volna mert elhiheted: ha erőteljes csoportosulások több szinten is ismétlődnek, akkor azt a módszer nemcsak az egyik szinten képes felismerni.

K: *Írod is, hogy a kötött klasszifikációnál nem mindig a teljes hierarchia az érdekes, hanem csak egy adott partíció.*

V: Nemcsak ott, persze, mert rendszerint a teljes hierarchiára (egészen az objektumokig lebontva) igen ritkán van szükségünk. Majd' minden hierarchikus módszerhez található egy nem-hierarchikus megfelelő, amelyet arra használhatunk, hogy a hierarchikus osztályozást egy adott szinten elmetszve (hogy hol, azt már megbeszéltük) a kapott partíciót feljavítsuk. A *k*-közép módszer például az eltérésnégyzet-összeget optimalizáló hierarchikus módszerek jó kiegészítője lehet. Előfordulhat, hogy a hierarchikus módon kapott partíción az utólagos áthelyezések már nem is tudnak változtatni, de rendszerint mindig van egy kis javítási lehetőség.

K: *A partíciókról szólva beszéltél átfedő csoportokról. Vannak-e ilyenek a hierarchikus osztályozásban is?*

V: Igen, voltak bizonyos próbálkozások, hogy a hierarchikus osztályozásban rejlő bizonytalanságokat a dendrogramon is feltüntessék (pl. Dabinett & Wellman 1978 gomba-osztályozásaiban). Ez még úgy-ahogy áttekinthető, ha kicsiny az átfedés mértéke, de bonyolultabb esetben a dendrogram teljesen zavarossá válik. Ezért az átfedéses hierarchiákat ma már – tudomásom szerint – nem használják.

K: *Akkor hadd találgassak tovább: logikus lenne ezek után a fuzzy hierarchikus osztályozás is!*

V:Valóban van ilyen. Marsili-Libelli (1989) szerint a súlyok objektumonkénti maximumai alapján szerkeszthető egy dendrogram is, de ez rendszerint nem mentes a visszafordulásoktól.

K: *Úgy vélem, hogy nagy adattömegekkel itt is gondban lehetek, csakúgy mint a partícionáló módszerek esetében. Az 5.3b ábra szerinti algoritmusok a leggazdaságosabbak ugyan a memória szempontjából, de ha több ezer objektumom van, akkor még ezek is használhatatlannak tűnnek.*

V: Így van, de nagy adattömeg gyors hierarchikus osztályozására is születtek már különféle ötletek. Jambu (1981, programlistával) divizív jellegű módszere akár 5000 objektumra is alkalmas, persze nem építi fel a teljes hierarchiát, csak a legfelső max. 50 partíció készül el, de ez is bőven elegendő.

K: *Még egy provokatív kérdés, és szinte kitalálom a választ. Hogy állunk a klasszifikációs térsorokkal?*

V: Természetesen ezek is vannak, hiszen láthattad a különböző flexibilis módszereket és az 5.10 ábrát. Mi sem könnyebb, mint egy ilyen térsor előállítása, csak a megfelelő paramétereket kell szisztematikusan megváltoztatnunk.

K: *A különbözőségi indexek közötti tájékozódásban nagyon hasznosnak látszik a 3. fejezet végén megadott "határozókulcs". Tudsz-e valami ilyesmit megadni a hierarchikus osztályozásban még járatlan kezdő számára is?*

V: Azt hiszem itt nem nagyon van erre lehetőség. A különbözőségek használata valóban sok mindentől függ, alkalmazásuk köre értelmes módon leszűkíthető, így készülhetett el az a bizonyos döntési fa. A hierarchikus osztályozás módszereivel ez már nem megy. Néhány tanácsot azonban mindenképpen tudok adni. Az egyszerű-, a teljes lánc és a csoportátlag módszer semmiképpen se maradjon ki egy klasszifikációs vizsgálatból, mert ezek alapvető strukturális tulajdonságokat jelezhetnek számunkra. Ha adataink megengedik (az eltérésnégyzetösszegnek, az átlagnak van értelme), akkor az eltérésnégyzetösszeg-növekedés minimalizáló eljárást is futtassuk le. Ez a négy módszer annyira általánosan elterjedt, hogy eredményeink mások számára is azonnal értelmezhetőek lesznek. Emellett ízlésed szerint másokat is érdemes alkalmazni, mondhatnám – ki tudja hányadszor –, hogy minél többfélét. És a végső figyelmeztetés: még egyszer hangsúlyozom, hogy ne használjuk a hierarchikus módszereket önmagukban, hanem más eljárásokat is vegyünk be a vizsgálatba.

K: *Érdekelne, hogy m objektumnak hányféle hierarchikus osztályozása létezik? Csak sejtem, hogy jóval több, mint amennyi partíciója, mondjuk k csoportra.*

V: Igen, a *k*-ad szintű partíciók száma (4.17 formula) elenyésző a lehetséges hierarchiák számához képest, bármekkora is *k* értéke. Ha most csak a villás elágazások topológiája szerint különböző dendrogramokat vesszük figyelembe, akkor *m* objektumra éppen

$$V_m = \frac{(2m-3)!}{2^{m-2}(m-2)!} \tag{5.16}$$

különféle fa adható meg (Cavalli-Sforza & Edwards 1967, Phipps 1975). Ez m=10-re kb. 34 és fél milliót tesz ki, s ebből már érzékelheted, hogy egy jóval nagyobb objektumhalmaz összes dendrogramjának figyelembe vétele gyakorlatilag lehetetlen. Ha pedig a hierarchikus szintek sorrendiségét is fontosnak tartjuk, akkor a lehetőségek száma:

$$D_m = \frac{m!(m-1)!}{2^{m-1}} \tag{5.17}$$

(Frank & Svensson 1981). Ez 10 objektumra két és fél milliárdnál is több! A hierarchikus szintek konkrét értékeit tekintve természetesen már végtelen számú lehetőség adódik.