Tanulmányok

188/1986

MTA Számítástechnikai és Automatizálási Kutató Intézet Budapest

Magyar Tudományos Akadémia
Számitástechnikai és Automatizálási Kutató Intézete

A CLUSTER ANALÍZIS NÉHÁNY KOMBINATORIKAI ÉS VALÓSZÍNŰSÉGSZÁMÍTÁSI PROBLÉMÁJA

Lengyel Tamás

Tanulmányok 188/1986

A kiadásért felelős:

Dr. KEVICZKY LÁSZLÓ

Főosztályvezető:

DEMETROVICS JÁNOS

Jelen tanulmány eredetileg a szerző kandidátusi disszertációja.

ISBN 963 311 217 6

ISSN 0324-2951

Tartalomjegyzék

Bevezetés

A cluster analízis a többdimenziós statisztikai módszerek egy viszonylag újkeletű ága. Általában gyakorlati eszköznek tekintik, ami összefüggések feltárását segíti elő. Az irodalomban csak kevés példa található a feladatok matematikai igényességű tárgyalására (Anderberg [2], Sneath [102], Cormack [17], Everitt [26], stb.).

A matematikai érdeklődésre közvetlenül számot tartó eredmények általában valamilyen klasszikus, régóta kutatott területhez kapcsolódnak, mint például a valószínűségszámítás (MacQueen [80], Hartigan [42], Pollard [86], stb.), a véletlen gráfok elmélete (Ling [75], [76] stb.), a kombinatorika és gráfelmélet (Matula [82]). A műszaki alkalmazások közül kiemeljük a kvantizálási problémakört (Zador [112], stb.) és az ehhez kapcsolódó geometriai problémákat (Fejes Tóth [28], Heppes és Szüsz [43], Gray és Karnin [36]).

A cluster analízis mély matematikai kérdéseket is napvilágra hoz. A tényekhez tartozik, hogy ezeket gyakran nem közvetlenül clusterezési kérdéseknek tekintik, hiszen megfogalmazásuk más terület

terminológiáját használva általában egyszerűbb és tanulságosabb. A disszertáció a clusterező módszerek egyik legfontosabb osztályának, a hierarchikus cluster struktúráknak a leszámlálási problémáját oldja meg (aszimptotikus értelemben).

A disszertáció másik fontos vizsgálati iránya a gyakorlati szempontból nem elhanyagolható algoritmikus és bonyolultsági kérdéseket tárgyalja. Ezek közül első sorban a következő problémát vizsgáltuk: hogyan lehet az általános esetben lényegében csak teljes kereséssel megoldható problémákat közvetlen vagy közvetett módon ésszerű, de legalábbis polinom időben futtatható problémára egyszerűsíteni.

Az egyszerűség kedvéért a következő modellt használjuk: a clusterezendő mintát egy súlyozott élű teljes gráffal írjuk le. A mintaelemeknek a gráf csúcsai felelnek meg, az élek súlya (hossza) a megfigyelések közötti "különbözőség" nagyságát mutatja. A clusterozás célja az, hogy megadjuk a gráf csúcsainak valamilyen optimális particióját. Az egyik vizsgálati irány a kritériumfüggvény szerint vett optimális clusterezés keresését tűzi ki célul. Ekkor valamilyen, a clustereken értelemezett függvény értéket kívánjuk minimalizálni a megfelelő partició

keresésével.

A clusterekkel kapcsolatos homegenitási kritériumok közül a következő hármat tárgyaljuk: megadható-e, illetve ha igen, akkor hogyan adható meg olyan k-particionálás, hogy

1. minden osztályon belüli távolság kisebb a különböző osztályból vett pontpárok minimális távolságánál,
2. a különböző osztályból vett pontpárok minimális távolsága maximális legyen az összes k-particionálás között,
3. az azonos osztályból vett pontpárok maximális távolsága minimális legyen az összes k-particionálás között.

A 3. fejezetben a kombinatorika, az algebra, a véges topológikus terek elmélete és a cluster analízis egy közös kérdésére adunk választ: meghatározzuk a hierarchikus cluster struktúrák számát, illetve ennek szimptotikus nagyságrendjét. Ez a szám azonos a partícióháló nem feltétlenül maximális 0-1 láncai számával és az n elemen értelmezett lényegesen különböző ultrametrikák számával. Először a 3 feladat ekvivalenciáját bizonyítjuk, majd megoldjuk a leszámlálási feladatot.

Minden hierarchikus cluster struktúra az alaphalmazon értelmezett partícióhálóban egy láncot határoz meg és fordítva. A hierarchi-

kus cluster struktúrákat ún. dendrogrammal szokás reprezentálni. Ez egy speciális gyökeres, számozott levelű fa, amelyben a csúcsokhoz szintszámok vannak rendelve.

Tudomásunk szerint Schadertől származik az az észrevétel, hogy egy n-elemű halmazon értelmezhető különböző ultrametrikák száma megegyezik az Eq(n) partícióháló nem feltétlenül maximális láncainak a számával.

Ha a Kruskal-algoritmust úgy módosítjuk, hogy egy lépésnek tekintjük azokat a lépéseket, melyek során azonos hosszúságú éleket veszünk az erdőhöz, akkor jól látható ennek az eljárásnak az **univerzális** jellege: minden hierarchikus clusterezés reprodukálható a módosított Kruskal-algoritmussal.

Megjegyezzük, hogy fontos struktúrák számának meghatározásához máshol is szükség van valamilyen speciális hálóban a nem feltétlenül maximális láncok összeszámolására (pl. 3.1.7. tétel). Bender [8] 1974-ben általános módszert is megadott a nem feltétlenül maximális láncok száma aszimptotikus nagyságrendjének meghatározására ún. binomiális posetekben. A partícióháló azonban bizonyos szempontból a legbonyolultabb háló (pl. a Pudlak-Tuma tétel szerint minden véges háló beleágyazható) és nem

elégíti ki a "binomiális poset" kritériumát.

A leszámlálási feladatban jelölje Z(n) az n-elemű halmazon értelmezett Eq(n) partícióháló nem feltétlenül maximális 0-1 láncainak a számát. A Z(n) meghatározása a egy rekurzióra vezet (3.2.2. állítás). Ennek kiértékelésével kapjuk a fejezet legfontosabb eredményét (3.2.3. tétel):

Tétel. Létezik olyan C_1 és C_2 pozitív konstans, hogy

$$C_1 \leq Z(n)/f(n) \leq C_2,$$

ahol $f(n) = (n!)^2 (2 \ln 2)^{-n} n^{-1-(\ln 2)/3}$ (ln a természetes logaritmust jelöli).

A 3.2.2. állításban Z(n)-re megadott rekurziót nem tudjuk pontosan megoldani, de találunk egy olyan sorozatot, amely elegendően jó közelítéssel kielégíti. Ezen eljárás eredményessége a Babai Lászlótól származó általános segédtételen (3.2.6. lemma), illetve a 3.2.10. következményen múlik.

Ezen általános eredmény alkalmazhatóságához a másodfajú Stirling-számokat elég tág határok között elég pontosan kell becsülnünk. Hsu [45] korlátos k-ra aszimptotikus sorfejtést adott S(n,n-k)-ra. Nekünk a $1 \leq k \leq n^{1/3-\varepsilon}$ (ahol ε tetszőleges kicsi pozitív

szám) tartományban $\forall$ k-ra kell az S(n,n-k) mennyiségre becslést adnunk. Ezt $O(k^6/n^2)$ relatív hibával oldjuk meg (3.2.8. lemma).

Az előző tétel természetesen veti fel azt a kérdést, hogy létezik-e a $\lim Z(n)/f(n)$ határérték. Ezt igazoljuk a 4. fejezetben (4.1.1. tétel)

Tétel. A következő határérték létezik

$$\lim_{n \to \infty} Z(n)/f(n) = C,$$

ahol a C egy pozitív konstans. (A C konstansra numerikus számítások a $C \approx 1.1$ becslést sugallják.)

A bizonyítás céljából egy önmagában is érdekes általános konvergencia kritériumot (4.1.2. lemma) vezetünk le.

Az 5.fejezetben a cluster analízis algoritmikus és bonyolultsági kérdéseit vizsgáljuk. Az általában nehezen kezelhető problémák inputosztályait igyekszünk különböző módon úgy szűkíteni, hogy az így adódó részproblémákat már polinomiális idő alatt lehessen megoldani.

Nem hierarchikus clusterezés során valamilyen rögzített k-ra keresünk egy optimális clusterezést. A fejezet fő eredménye egy (kritériumfüggvény szerinti optimum keresést kitűző) nem hierarchikus clusterezési problémáról mutatja meg annak polinomkorlátos voltát. (Ez a probléma tartalmazza az ún. legkisebb négyzetes clusterezési problémát.) A kritériumfüggvény alkalmas megválasztásával, implicit módon a partíciók geometriai struktúrájára kötünk ki feltételt.

Definíció. Az $S=(x_1,x_2,\ldots,x_n)\subset R^m$ pontok konvex k-partícióján olyan k-particionálást értünk, ahol az osztályok konvex burka diszjunkt.

Legyen az f egy szigorúan monoton növő, folytonos függvény. Az $S=(x_1,\ldots,x_r)\subset R^m$ halmaz **f-centrumának** azt a $q\in R^m$ pontot nevezzük, amelyre az **S-nek a q-ra vonatkozó f-nyomatéka,** azaz az

$$M(f,S,q)=\sum_{i=1}^{r} f(\|x_i-q\|)$$

összeg minimális. A megfelelő összeget az **S halmaz f-centrális nyomatékának** nevezzük.

Legyen $S\subset R^m$ egy véges halmaz és $S=S_1\cup\ldots\cup S_k$ az S egy partíciója. E partíció **f-nyomatékán** a partíció osztályain vett f-centrális nyomatékok $W(f,S,(S_i)_{i=1}^k)$ összegét értjük.

Ezt az értéket kívánjuk minimalizálni S összes k-particiója között, tehát keressük azt a k-particiót, melyre

$$W(f,S;k) = \min_{\substack{(S_i) \\ S\ k\text{-particiója}}} W(f,S,(S_i)_{i=1}^{k}).$$

Az ilyen k-particiókat nevezzük **f-optimális**nak.

Az m-dimenziós tér n pontját általános helyzetűnek nevezzük, ha bármelyik m, vagy kevesebb elemű részhalmaza lineárisan független rendszert alkot.

A fejezet fő eredményét fogalmazza meg a következő (5.3.2. tétel)

Tétel. Ha f szigorúan monoton növő, folytonos függvény és $S \subset R^m$ általános helyzetű pontok véges halmaza, akkor rögzített k és m esetén S egy f-optimális k-particiója polinom időben megtalálható. (Az így adódó f-optimális k-partició konvex lesz.)

A megadott algoritmus sajnos gyakorlatilag csak nagyon kicsi m és k értékek mellett futtatható.

A tétel bizonyítása azon az észrevételen (5.3.4. állítás) alapul, hogy n pont konvex k-particóinak száma lényegesen kevesebb az összes k-partició számánál: n-től csak polinomiálisan függ,

midőn a k és m rögzített.

Felvetődik az a kérdés is, hogy mi történik akkor, ha a probléma k vagy m, illetve mindkettő paraméterét az input részeként vizsgáljuk. Sajnos, egyik esetben sem tudjuk a választ, de azt sejtük, hogy mindegyik probléma NP-nehéz.

Az előző állítás [63] ismeretében Hardy és Rasson [38] 1982-ben a következő clusterezési kritériumot vezették be: keressük azt a konvex k-particiót, amelyre az osztályok által meghatározott konvex burkok térfogatösszege minimális. Egy dinamikus programozási módszert javasoltak az optimum keresésére, amely csak k és m rögzítése, illetve az egydimenziós esetben polinomkorlátos.

E cikk hatására 1985-ben ugyanezt a problémát vizsgálta Krivanek és Morávek [56], de a k-t az input részeként tekintették. Eredményük szerint a megfelelő döntési probléma NP-teljes minden rögzített $m \geq 2$ esetén.

A fejezet további részében a bevezetésben megfogalmazott 1. és 3. kérdésekkel foglalkozunk. A 3. probléma (a "leghomogénabb clusterezés" keresése) vizsgálatához először a $\Pi_5(k)$ problémával foglalkozunk. Lehet-e az $x \in V$ megfigyelési pontokat k osztály-

ba particionálni olymódon, hogy az azonos osztálybeli megfigyelések - a clusterezési probléma távolságfüggvényével mérve - egy előre adott konstanst nem meghaladó távolságra legyenek egymástól. Az osztály legtávolabb eső pontpárjának távolságát nevezzük a **cluster átmérő**jének. Az egyszerűség kedvéért feltesszük, hogy mind a konstans, mind a távolságok nemnegatív egészek. A "leghomogénabb clusterezés" keresésének általános problémája NP-nehéz. Hochbaum és Shmoys [44] közelítő algoritmust adtak meg 1984-ben. Mi a pontos megoldást keressük metrikák speciális osztályaira.

Könnyen látható, hogy ultrametrikákra ez a kérdés polinom időben megoldható. Ennek az észrevételnek a metrikák egy lényegesen tágabb osztályára való kiterjesztése a fejezet másik fő eredménye. Ennek bemutatására szükségünk van néhány definícióra.

Definíció. Egy súlyozott élű teljes gráf **szintgráfja**in azokat a súlyozatlan élű részgráfokat értjük, amelyek úgy keletkeznek az eredeti gráfból, hogy csak azokat az éleket hagyjuk meg, melyek hossza egy adott számnál nem nagyobb, pl. a λ-szintű részgráf élei pontosan azok lesznek, melyekre $d(i,j) \leq \lambda$.

Ezek szerint a $\Pi_5(k)$ probléma ekvivalens a következővel: jelölje

R az adott küszöb számot. Létezik-e az R-szinthez tartozó szintgráfban a csúcsoknak egy olyan $(V_1,V_2,\ldots,V_k)$ particiója, amelyre minden V_i $(i=1,2,\ldots,k)$ a gráf teljes részgráfját feszíti. Ha a távolságértékeket a (0,1) halmazra szűkítjük le, akkor a kérdezett klikk-partició létezése ekvivalens a megfelelő gráf komplementer gráfjának a k-színezhetőségével. Tehát a k)2 eset tartalmazza az NP-teljes 3-színezhetőség problémáját, így nyilván $\Pi_5(2) \in P$ (azaz polinom időben megválaszolható), míg $\Pi_5(k) \in$ NP-teljes, ha k)2.

Bevezetünk egy távolságfüggvény osztályt, az ún. fa-szerű metrikák osztályát (5.4.6. definíció), és megadunk egy polinom idejű algoritmust az ilyen távolságokra megszorított Π_5 probléma megoldására (5.4.7. tétel).

Tétel. Fa-szerű metrikára $\Pi_5 \in P$ (azaz polinom időben megválaszolható a $\Pi_5(k)$ olyan formában is, hogy a *k*-t előre nem rögzítjük, hanem az input részeként tekintjük).

Az algoritmus lényege az, hogy a problémát bizonyos perfekt gráfok kiszínezésére vezetjük vissza. Elegendő ugyanis a szintgráfokkal foglalkozni és fa-szerű metrikára a teljes gráf mindegyik szintgráfja perfekt gráf. A $\Pi_5(k)$-ben tekintett általános

gráfszínezési probléma perfekt gráfokra Grötschel, Lovász, Schrijver [37] egy az ellipszoid módszert felhasználó algoritmusa révén polinom időben megoldható. A nekünk szükséges speciális esetben elkerülhető az ellipszoid módszer használata, hiszen a szintgráfként adódó speciális perfekt gráf illetve komplementere egyszerűen színezhető.

Az 1. ("homogenitási") kérdést fogalmazza meg a $\Pi_8(k)$ döntési probléma: egy n-pontú súlyozott élű gráfnak létezik-e olyan k-osztályú particionálása, amelyben az azonos osztálybeli pontok közötti távolságok maximuma kisebb a különböző osztályokból vett pontok minimális távolságánál.

A kérdés szorosan kapcsolódik a clusterező eljárások összehasonlítását lehetővé tevő - az 5. fejezetben fő vonalaiban ismertetésre kerülő - ún. megengedettségi osztályozásához.

Az előbbi kérdésre az igen választ közvetlenül bizonyító k-particiót nevezzük **kompakt szeparáltnak.** Ha a külső és belső távolságok egyenlőségét is megengedjük, akkor beszélünk **jól struktúrált** k-particiókról. Nyilvánvaló, hogy egy kompakt szeparált k-partició azt jelenti, hogy az eredeti gráfnak létezik olyan szintgráfja, ami k csúcs- és él-diszjunkt teljesre esik

szét.

A single linkage eljárás kompakt szeparált megengedett (5.4.9. állítás), ezért a Π_8 probléma megválaszolására a Kruskal-algoritmus alkalmas. Az ultrametrikus tulajdonság már biztosítja jól struktúrált k-particionálások létezését (5.4.3. állítás), sőt ekkor jól jellemezhetők azok az inputok, amelyek (k-tól függetlenül) pozitív választ adnak a Π_8 problémára (5.4.10. állítás).

A 6. fejezetben kitérünk arra a kérdésre, hogyan lehet verifikálni egy gráf feszítőfájáról, hogy minimális feszítőfa (MFF); - hogyan tesztelhető a gráf éleinek egy részhalmazáról, hogy kiegészíthető-e MFF-vá, illetve a csúcshalmaz egy particója előáll-e a feszítőerdő komponenseként a Kruskal-algoritmus futtatása során. Megadunk egy egyszerű struktúrát, amelynek segítségével a MFF egy viszonylag nagy részgráfja megkonstruálható. Ez a struktúra lényegében az "all nearest neighbor" probléma (Shamos [100]) megoldásakor adódik. Tekintsük ugyanis az ún. legközelebbi szomszédsági vagy röviden **NN-gráfot**, azaz amelyikben minden csúcsot valamelyik legközelebbi szomszédjába mutató éllel kötünk össze. A 6.4.13. tétel biztosítja, hogy parallel számításokkal (bizonyos esetekben a Prim-Dijkstra és a Kruskal-algoritmusnál

hatékonyabban [1]) megadható a MFF éleinek legalább fele:

Tétel. Tetszőleges távolságmátrix esetén a $G^*=(V,E^*)$ irányított NN-gráf bármely irányítatlan, egyszerű és körmentes $G_1=(V,E_1)$ részgráfjához található a G_1-t tartalmazó minimális feszítőfa. Ha a G_1 gráf maximális abban az értelemben, hogy további él hozzávétele a körmentesség feltételét sértené, akkor a G_1 egy legalább $n/2$ élű erdő.

Végül a 6.4.7. állításban az ultrametrikák egy a szokásostól eltérő geometriai jellemzését adjuk meg.

A 6. fejezetben ismertetjük azokat az eredményeket (Shamos és Hoey [101], Toussaint [105], Brown [11]), amelyek a disszertációban részletesen tárgyalt két látszólag teljesen független módszer közös algoritmikus gyökerére mutatnak rá.

A záró fejezetben a clusterezésekkel kapcsolatos statisztikai problémákkal foglalkozunk. Az ilyen jellegű vizsgálatok célja az, hogy statisztikai módszerrel is alátámaszthassuk a clusterezés eredményét. A kérdéskörben született eredmények aszimptotikus jellegűek, így közvetlen hipotézisvizsgálatra nem alkalmasak.

A 7.3.3. tételben Lengyel és Ruda ([61], [62]) egy elég általános eloszláscsaládban lényegében megválaszolták a (7.3.4)-ben

definiált $V(k;P,2)$ speciális kvantizálási veszteségre a következő problémát: mekkora az optimális veszteség aszimptotikus nagyságrendje, midőn a csoportok k száma tart a végtelenhez? Zador [112] kezdte el vizsgálni azt, hogy ez hogyan függ az eloszlástól. Bucklew és Wise [12] 1982-ből származó cikke a korábbi eredményeket általánosította.

1. Cluster analízis

1.1. A cluster analízis feladata

A klasszifikáció a tudományos fogalomalkotás egyik fontos módszere: dolgok absztrakt fogalmakkal történő meghatározásának, megnevezésének és megkülönböztetésének, lényeges és lényegtelen szétválasztásának az eszköze, ami a dolgokkal kapcsolatos közléseket is megkönnyíti. Ezt tekintik a tudományos gondolkodás egyik legősibb elemének. A tudományágak szétválásánál és az egyes ágazatok fejlődésében is fontos szerepet játszott a klasszifikáció. Különösen jelentős az a hatás, amit a biológiára és a zoológiára fejtett ki és aminek egyik csúcspontja a Darwin-féle fejlődéstörténet. Megemlíthetjük a Mengyelejev-féle periódusos táblát is, ami forradalmi változást hozott az általános kémiai gondolkodásban.

A clusterezés a klasszifikáció egyik ága. Elsősorban bonyolult jelenségek megértését és megmagyarázását segítő módszernek tekintik, de időnként az új összefüggések feltárása nyomán a fogalomalkotást is támogatják vele.

A számítógépek kapacitás-növekedését követi az cluster analízis

alkalmazási lehetőségeinek a köre. Ennek nyomán az utóbbi évtizedben érdekes eredmények születtek pl. az orvostudomány (pl. Tusnády [108]), a mezőgazdaság (pl. Jansen, Bethlehem [48]), a társadalomtudomány területén (pl. Kolosi, Lengyel [54]), stb.

A jelenség leírásához megfigyeléseket végeznek. A megfigyelések kvantifikálása alatt bizonyos jellemzők valamilyen skála szerinti mérését (pl. súly, magasság, stb.), mások véges sok kategóriába sorolását (pl. szín, nem, stb.) értjük.

Többnyire feltesszük, hogy a vizsgált jelenségek leírhatók az s-dimenziós euklideszi térben, és a jelenségek realizációi közötti különbségek mérésére a - megfigyelt n mintaelemet reprezentáló - pontok között euklideszi metrikában illetve ennek egyszerű függvényeivel mért távolságok alkalmasak.

A koordinátarendszer dimenziója és a koordináták (skálák és kategóriák) megválasztásakor a teljességre való törekvésnek és a gyakorlati megoldhatóság szempontjából káros redundanciák kiküszöbölésének egymással ellentétes hatását kell összehangolni. A különböző szempontból történő elemzések során szükséges lehet egyes koordinátáknak a fontosságuk szerinti kiemelésére illetve elhagyására. Ha a megfigyelések valamilyen többdimenziós normális

eloszlást követnek, akkor az egyes komponens változóknak a többi-hez viszonyított "természetes fontosságát" illetve "redundanciáját" figyelembe vehetjük az eredeti változókra végrehajtandó főkomponens analízis segítségével. A faktortérbeli koordinátái euklideszi távolsága az eredeti adatok ún. Mahalanobis-féle távolságát adja. Ez invariáns az adatok tetszőleges nem szinguláris lineáris transzformációjára, így egy koordinátarendszer független távolsággal számolhatunk.

A redundanciák kiszűrésének egyéb lehetőségei közül megemlítjük még a - főkomponens analízis általánosításának tekinthető - kanonikus korrelációanalízis módszereit (Tusnády [107], Lengyel [64], [66], [72]).

1.2. A cluster analízis problémái és módszerei

A cluster analízis alkalmazása a gyakorlati problémát megfelelően leíró modell megalkotásával kezdődik. Ezt a modell elemzése, a kitűzött feladatot a gyakorlatban is megoldani képes algoritmusok keresése és ezek esetleges további vizsgálata (implementáció, konkrét tapasztalatok) követi. A modell elemzése főleg szakmai jellegű kérdések megválaszolását igényli, elsősorban azzal kapcsolatosan, hogy a modell mennyire alkalmas a nehezen megfogható

belső összefüggések - gyakorlati célnak tekintett - feltárására.

Az alapfeladat az, hogy valamilyen értelemben homogén csoportokba soroljuk a megfigyeléseket. Annak megfelelően, hogy a keletkező clusterek halmazelméleti, metrikus, esetleg gráfelméleti értelemben milyen tulajdonságokkal rendelkeznek, valamint, hogy az ezeket produkáló algoritmus milyen úton, milyen optimalizálási kritériumok segítségével vezet a csoportosításhoz, osztályozhatjuk a különböző clusterezési algoritmusokat. A módszerek osztályozásra több próbálkozás ismeretes pl. Sneath, Sokal [102], Fisher, Van Ness [30], Jardine, Sibson [46]). A leggyakrabban alkalmazott algoritmusokat részletesen ismerteti Anderberg [2] és Hartigan [40].

A továbbiakban mindig a megfigyelések egy k-osztályú particionálását értjük clusterezésen, azaz minden megfigyelést k szóbajövő (nem üres) cluster közül pontosan egybe sorolunk be. Az egyszerűség kedvéért az optimális k-particiót (esetleg ezek k szerint vett sorozatát) kereső módszereket nevezzük nem hierarchikusaknak. A k-particiók finomításával (durvításával) partició sorozatot képező módszerek pedig az agglomeratív (divizív) hierarchikus módszerek.

A clusterezéshez szorosan kapcsolódik az ún. kvantizálási prob-
lémakör. Ekkor feltételezzük, hogy a megfigyelések ismert elosz-
lásból származnak.

2. Két clusterezési módszer ismertetése

2.1. A legközelebbi szomszéd módszer

A legközelebbi szomszéd módszer (nearest neighbor, single linkage, egyszerű kapcsolat) a clusteranalízis egyik leggyakrabban alkalmazott, az agglomeratív hierarchikus clusterező eljárások közé sorolt eljárása. Az agglomeratív hierarchikus clusterezések során a megfigyeléseket reprezentáló n egyelemű clusterből indulunk ki és minden lépésben egy vagy több clustert egyesítünk, amíg a maximális, egyetlen n-elemű osztályból álló clusterezéshez nem jutunk.

A gyakorlati alkalmazások során a módszert nem futtatják végig, hanem egy megfelelő clusterszámnál megállnak és a kapott cluster struktúrát elemzik tovább. Valójában egy csonka agglomeratív hierarchikus eljárást hajtanak végre. Ennek a gyakorlatnak az az alapja, hogy feltételezhető, hogy a megfigyelések nem valamilyen homogén struktúrát írnak le. Ha k eloszlás keverékéből érkeznek a mintaelemek, akkor k körüli clusterszámnál érdemes megállni. Általában a k értéke nem ismert előre, és ez bizonytalanná teszi,

hogy valójában meddig van értelme az eljárást folytatni.

Módszerenként változik az kritérium, amely szerint az összevonásra kerülő clustereket kiválasztjuk. Az alkalmazás jellege szerint törekedhetünk például arra, hogy egy adott szinten két megfigyelést már akkor is ugyanabba a clusterba soroljunk, ha létezik közöttük egy lánc, amelyben az egymás utáni elemek már legfeljebb annyira különböznek (más szóhasználattal legfeljebb olyan "messze" vannak) egymástól, mint a szintnek megfelelő szám. A legközelebbi szomszéd módszer esetében pontosan ezt jelenti az egyszerű kapcsolat fogalma.

érdekessége ennek az eljárásnak, hogy valójában a megfigyelések minimális feszítőfájának mohó algoritmussal történő konstruálása során adódó komponensek felelnek meg a clustereknek. Ezért ezt a módszert Kruskal-algoritmusnak is hívjuk és részletesen tárgyaljuk a kapcsolódó algoritmikus kérdéseket is (6.3., 6.4., 6.5. pont). A 6.6. pontban két másik kritériumot is megemlítünk. Mindkét eset a Kruskal-algoritmus segítségével könnyen programozható. Még érdekesebb a 3.1. pont azon megállapítása, hogy valójában minden hierarchikus clusterezés egy Kruskal-algoritmus végrehajtásának felel meg. A 3.2.3. és 4.1.1. tétel az ilyen clusterezések folyamán épülő struktúrák, azaz az összes agglome-

ratív hierarchikus struktúra számának aszimptotikus nagyságrendjét adja meg. A 7.1.1. tétel [103] a teljes eljárás végén adódó fa élei - euklideszi metrikában mért - összhosszának aszimptotikus nagyságrendjét állapítja meg.

2.2. Egy nem hierarchikus clusterezési kritérium

Általában a nem hierarchikus clusterezések során a megfigyelések alaphalmazának particióira vonatkozó valamilyen kritérium szerint optimális clusterezést keresünk. Természetesen a clusterezés tényleges céljának megfelelő, adekvát kritériumfüggvény megtalálása általában nehéz és gyakorta nem matematikai jellegű feladat. A kritériumfüggvény szerinti optimumoktól elvárható matematikai tulajdonságok általános vizsgálata pl. megengedettségi szempontok alapján (5.fejezet) azonban sok esetben szükséges.

Egy n-elemű halmaznak összesen annyi particionálása van, mint amekkora a 3.fejezet elején definiált $\omega(n)$ ún. Bell-szám, sőt még a k-osztályú particionálások száma is a másodfajú S(n,k) Stirling-féle számmal egyenlő, ami rögzített k mellett is exponenciális gyorsan nő az n függvényében. Ez mutatja, hogy a krité-

rium függvénynek az összes lehetséges esetben történő kiértékelése már viszonylag kis n érték mellett is gyakorlatilag kivitelezhetetlen.

Áthidaló megoldásként megelégszünk azzal, hogy rögzített k pozitív egészek mellett, a kritérium szerint globális optimumot adó clusterezések helyett olyan k-particiókat keresünk, amelyek valamilyen értelemben "lokálisan" optimálisak.

Az alábbiakban egy klasszikusnak tekinthető eljárást ismertetünk, amellyel a "clustereken belüli, a clusterek geometriai középpontjától vett euklideszi távolságok négyzetösszege" (within-cluster sum of squares) minimalizálási kritériumra (vagy röviden "legkisebb négyzetes eltérések" vagy még rövidebben "legkisebb négyzetes" kritériumra) nézve a megfigyelések optimális particióit keresik. Ezzel a problémával - a jelen fejezeten kívül - részletesen foglalkozunk a 7.2., 7.3. és a 5.3. pontokban.

A fenti kritérium kvantizálási megfelelője az "osztályokon belüli szórások négyzetösszege" (within-class-variance) - röviden: legkisebb négyzetes kvantizálási - kritérium. Ilyenkor az $P^{(n)}$ egy teljes k-particióját keressük.

Először a legkisebb négyzetes clusterezésekre vonatkozó k-közép eljárást ismertetjük. Megjegyezzük, hogy ez a módszer alkalmas módosítással az r. hatvánnyal mért eltérések esetére is általánosítható.

A k-közép eljárás

A k-közép clusterező eljárás az $S=\{x_1,x_2,\ldots,x_n\}\subset R^m$ pontok legkisebb csoporton belüli átlagtól való eltérések négyzetösszegét kívánja minimalizálni, midőn a megfigyeléseket legfeljebb k osztályba csoportosítjuk. (Az 5.3. pont terminológiáját használva az S halmaznak egy, a x^2 függvényre nézve minimális nyomatékot adó ún. x^2-optimális k-particióját kell megadni.) Keressük ugyanis azokat az egymástól különböző $q_1,q_2,\ldots,q_k\in R^m$ pontokat, melyekkel az egyes clustereket - a következő kritérium alapján optimálisan - lehetséges reprezentálni (q_i reprezentálja a i. clustert)

$$(2.2.1)\qquad W_n(S)=\min_{(q_i)} W_n(S,(q_i)_{i=1}^k),\qquad \text{ahol}$$

$$(2.2.2) \qquad W_n(S,(q_i)_{i=1}^k) = \sum_{i=1}^n \min_{1\le j\le k} \|x_i - q_j\|^2.$$

(Megjegyezzük, hogy a $W_n(S,.):R^{m\times k} \to R$ függvényt a reprezentáns pontok halmazfüggvényeként is értelmezhetjük, bár ekkor fölösleges magyarázatra szorulnának olyan fogalmak, mint az argumentum euklideszi metrikában vett környezete. Az 5.3. pont definíciói szerint a q_i pontokat a megfelelő clusterek x^2-centrumának nevezzük.)

Nyilván úgy kell megválasztani a clustereket, hogy mindegyik x_i megfigyelést valamelyik hozzá legközelebbi q_j által reprezentált osztályba soroljuk. Ekkor viszont nem feltétlenül a valódi átlagtól számítjuk az eltérések négyzetösszegét, és így a q_j reprezentánst a mintaközépre cserélve csökkenthetünk a W_n értékén.

Az előbb vázolt két lépés egymás utáni váltogatása egy olyan particionáláshoz és $Q = (q_1,q_2,\ldots,q_k)$ reprezentáns rendszerhez vezet, amihez éppen a fent definiált $W_n(S,(q_i)_{i=1}^k)$ eltérések négyzetösszege tartozik.

A k-közép eljárás a következő:

1. Induljunk ki egy kezdeti $(q_1,q_2,\ldots,q_k)$ cluster reprezentáló rendszerből és mindegyik x_i pontot azon cluster

valamelyikébe soroljuk, amelyik reprezentánsához a legközelebb van.

2. Azon clusterek reprezentánsát, amelyek közepe nem egyezik meg a csoport átlaggal, helyettesítjük a csoportátlaggal. Ha ilyen cluster nincs, akkor befejezzük az algoritmust.

3. A k reprezentáns alapján újra besoroljuk a megfigyelési pontokat, mindegyiket valamelyik legközelebbi reprezentáns clusterébe. Szemléletesen azt is mondhatjuk, hogy a 3. lépés az olyan clustereket változtatja, amelynek centruma vagy valamelyik szomszédos clusterjának a centruma elmozdult.

A 3. lépés után a visszamegyünk a 2. lépéshez.

Amikor az eljárás a 2. lépésnél véget ér, akkor a kapott reprezentáns rendszer stacionárius pontja lesz a $W_n(S,.)$ függvénynek abban az értelemben, hogy az eljárás nem vezet ki belőle. A k-közép módszer azonban nem feltétlenül a (2.2.2) kifejezésben definiált $W_n(S,.)$ függvény lokális optimumánál áll le. Ha nem vagyunk kíváncsiak az összes globális optimumot adó rendszerre, akkor a fentiek szerint elegendő az előző értelemben stacionárius pontok között keresni egy globális optimumot adó megoldást. Ezek szerint megfogalmazható a következő

2.2.1. állítás. Létezik olyan $Q=(q_i)_{i=1}^{k}$ reprezentáns rendszer, amely által az S-en meghatározott $O(Q)=(S_1,S_2,\ldots,S_k)$ legkisebb távolság partició (6.4. pont) szolgáltatja a $W_n(S,.)$ globális optimumát, azaz

$$W_n(S) = \sum_{i=1}^{k} \int_{x_j \in S_i} \|x_j - q_i\|^2. \qquad (2.2.3)$$

A legkisebb távolság partició (LTP) legfontosabb tulajdonságaival a 6. fejezetben részletesen foglalkozunk.

Könnyen látható, hogy a (2.2.2)-ben definiált $W_n(S,.)$ függvény nem feltétlenül konvex $R^{m\times k}$-ben, ezért általában nem várható, hogy konvex programozás segítségével megtaláljuk a globális optimumot.

A 4.3. pontban azt vizsgáljuk meg, hogy - a 2.2.1. állításhoz hasonló módon - a globális optimumhely keresésére vonatkozóan hogyan lehet leszűkíteni a megvizsgálandó esetek számát. A lényeges észrevétel az, hogy elegendő lesz az Q halmaz pontjai által meghatározott legkisebb távolság particiókra szorítkozni. Minthogy az $q_1,q_2,\ldots,q_k \in R^m$ pontok tetszőleges elhelyezkedésűek, ezért a LTP geometriai tulajdonságát (ti. a LTP mindegyik osz-

tálya konvex) hívjuk segítségül az esetek tényleges megszo rításához (5.3.2. tétel).

Röviden ismertetjük Lloyd ún. első módszerét, amely egydimenziós legkisebb négyzetes kvantizálás optimális struktúrájának keresésére vonatkozik, de könnyen lehet magasabb dimenzióra és más kitevőre általánosítani.

Jelölje P illetve F a szóbanforgó egydimenziós valószínűségi változó által indukált mértéket illetve ismert eloszlásfüggvényét. Tegyük fel, hogy a valószínűségi változónak véges második momentuma van.

Jelölje $c_1 < c_2 < \ldots < c_{k-1}$ az elválasztó pontok és $q_1 < q_2 < \ldots < q_k$ a kvantizáló értékek kezdeti sorozatát.

Legyen $i=2,3,\ldots,k-1$ esetén $S_i = (x \mid c_{i-1} < x \leq c_i)$, valamint $S_1 = (x \mid -\infty < x \leq c_1)$ és $S_k = (x \mid c_{k-1} < x < \infty)$ azoknak az intervallumoknak a kezdeti sorozata, ahol a kvantizáló függvény konstans, azaz, ha $x \in S_i$, akkor $Q(x) = q_i$.

2.2.2. definíció. Azt mondjuk, hogy az $c_1 < c_2 < \ldots < c_{k-1}$ és a $q_1 < q_2 < \ldots < q_k$ sorozat az

$$L((c_i),(q_i)) = \sum_{i=1}^{k} \int_{S_i} (x-q_i)^2 \, dF(x)$$

legkisebb négyzetes kvantizálási probléma stacionárius pontja, ha

$$(2.2.4) \qquad q_i = \int_{S_i} x \, dF(x) \; / \; P(S_i)$$

és

$$(2.2.5) \qquad c_i = (q_i + q_{i+1}) / 2, \; i=1,2,\ldots,k-1.$$

Megjegyezzük, hogy a S_i az (c_i) sorozat által van meghatározva és az L képzésekor mindig a megfelelő (S_i) halmazrendszeren értjük az integrálást.

2.2.3. definíció. Azt mondjuk, hogy az $c_1(c_2(\ldots(c_{k-1}$ és a $q_1(q_2(\ldots(q_k$ sorozat lokális minimumhelye a legkisebb négyzetes kvantizálási problémának, ha létezik olyan ε pozitív szám, hogy minden olyan $c_1'(c_2'(\ldots(c_{k-1}'$ és $q_1'(q_2'(\ldots(q_k'$ sorozat mellett, amire a

$$\max \left(\max_{i=1,2,\ldots,k-1} |c_i - c_i'| \, , \; \max_{i=1,2,\ldots,k} |q_i - q_i'| \right) < \varepsilon$$

egyenlőtlenség teljesül fennáll, hogy

$$L((c_i),(q_i)) \leq L((c_i'),(q_i')).$$

2.2.4. definíció. Globális optimumhelynek a minimális L értéket adó lokális optimumhelyeket nevezzük.

Könnyen látható, hogy a lokális optimumhelyek c_i elválasztó pontjai az F eloszlásfüggvény folytonossági pontjai lesznek, valamint minden lokális optimumhelynek létezik olyan környezete, ahol az L függvény folytonos (Lloyd [77]).

Egyszerűen bizonyítható az is, hogy minden lokális optimumhely egyben stacionárius pont is. Ha nem az lenne, akkor létezne olyan i index, amire (2.2.4) vagy (2.2.5) nem teljesülne. Az előbbi esetben a q_i értékének a (2.2.4) szerinti megváltoztatásával, az utóbbiban az c_i értékének a (2.2.5) alapján történő újraszámolásával, majd az új c_i-nek megfelelő S_i és S_{i+1} szinthalmazokkal egy kisebb veszteség értéket adó kvantizáláshoz jutnánk. Egyszerű folytonossági meggondolásból következne, hogy a választott pont nem lehetett lokális optimumhely.

Lloyd eljárása az előbb javasolt lépések sorozatával csökkenti az L értékét.

Lloyd eljárása

1. induljunk ki egy kezdeti $(c_i^{(0)})$ sorozatból. Állítsuk be a j=0 kezdeti értéket.
2. Határozzuk meg az eljárás során adódó j. kvantizálás $S_i^{(j)}$ szinthalmazait. A (2.2.4) alapján számoljuk ki a $q_i^{(j)}$ kvan-

tizáló értékeket, azaz

$$q_i^{(j)} = \int_{S_i^{(j)}} x \, dF(x) \; / \; P(S_i^{(j)}), \quad i=1,2,\ldots,k.$$

3. A (2.2.5) alapján számoljuk ki az új $c_i^{(j+1)}$ elválasztó értékeket, azaz

$$c_i^{(j+1)} = (q_i^{(j)} + q_{i+1}^{(j+1)}) \; / \; 2, \quad i=1,2,\ldots,k-1.$$

Növeljük meg a j értékét: $j=j+1$.

Ha az $c_i^{(j+1)} = c_i^{(j)}$ egyenlőség minden $i=1,2,\ldots,k-1$ indexre fennáll, akkor az eljárás véges sok lépésben véget ért.

Ellenkező esetben folytassuk az algoritmust a 2. lépéstől.

Ha a 3. lépésnél nem áll le a módszer, akkor egy monoton csökkenő, nemnegatív, így konvergens $L((c_i^{(n)}),(q_i^{(n)}))$ veszteség sorozatot kapunk, midőn $n \to \infty$. A $((c_i^{(n)}),(q_i^{(n)}))$ sorozat (a R^{2k-1} euklideszi metrikájában vett) $(c_i^*),(q_i^*)$ torlódási pontjai közül biztosan stacionárius pontok lesznek azok, amelyekre a c_i^*-k folytonossági pontjai az F-nek (Lloyd [77]).

3. Leszámlálási problémák

Ebben a fejezetben a kombinatorika, az algebra, a véges topológikus terek elmélete és a cluster analízis egy közös kérdésével foglalkozunk: határozzuk meg

- a hierarchikus cluster struktúrák számát,
- a partícióháló nem feltétlenül maximális 0-1 láncainak a számát,
- az ultrametrikák számát,

és adjuk meg ezek aszimptotikus nagyságrendjét.

Először a 3 feladat ekvivalenciáját bizonyítjuk be. A partícióháló láncai és a hierarchikus cluster struktúrákat leíró ún. dendrogramok ugyanazt jelentik. A 3.1. pontban megadunk egy kölcsönösen egyértelmű megfeleltetést a partició láncok és az ultrametrikák között.

A 3.2. pontban térünk rá a leszámlálási feladat megoldására. A 3.2.2. állítás egy rekurziót fogalmaz meg a kérdezett számra. Ennek aszimptotikus nagyságrendjét adja meg a fejezet fő eredményét tartalmazó 3.2.3. tétel.

A 4. fejezetben vizsgáljuk a vizsgált mennyiség és becsült értékének aszimptotikus arányát.

A partícióháló nem rendelkezik a binomiális poset struktúrájával így Doubilet, Rota és Stanley [22] és Bender [8] módszerei nem alkalmasak az említett láncok összeszámlálására.

Mielőtt a tárgyalásba kezdenénk megemlítjük, hogy egy n-elemű halmaz k-osztályú particionálásainak a száma a másodfajú S(n,k) Stirling-féle számmal egyenlő, míg az összes lehetséges particionálások számát nevezzük Bell-számoknak, vagyis

$$\omega(n) = \sum_{k=1}^{n} S(n,k).$$

A partíciók és az ekvivalencia relációk között kölcsönösen egyértelmű megfeleltetés létesíthető.

3.1. A hierarchikus cluster struktúrák száma és az ultrametrikák száma

Az n mintaelem hierarchikus clusterezései során előforduló cluster struktúrák és a particiháló láncai ugyanazt jelentik. Mind a cluster struktúrákat, mind a partició láncokat speciális gyökeres, számozott végpontú fákkal szokás reprezentálni. Az előbbi

esetben ezeket a fákat dendrogramoknak is hívják.

Ebben a fejezetben megadunk egy kölcsönösen egyértelmű megfeleltetést az Eq(n) partícióháló láncai és az n ponton értelmezhető különböző ultrametrikák között.

Először definiáljuk, hogy mit értünk a partícióháló láncán.

3.1.1. definíció. Legyen x és y az (1,2,...,n) halmaz két partíciója. Azt mondjuk, hogy az x partíció a y finomítása ($y \preceq x$), ha az y minden osztályát tartalmazza az x valamelyik osztálya. Ha $y \preceq x$ és az x különbözik az y partíciótól, akkor azt mondjuk, hogy az x szigorúan finomabb az y-nél ($y \prec x$).

Ezzel a részben rendezéssel a partíciók halmaza hálót alkot, ezt szokás partícióhálónak (Eq(n)) nevezni. A háló minimális eleme az ((1,),(2),...,(n)), a maximális pedig az (1,2,...,n) partíció.

3.1.2. definíció. A minimális elemmel kezdődő és a maximális elemmel végződő szigorúan finomodó partíció sorozatot nevezzük a háló nem feltétlenül maximális láncának.

Most rátérünk arra az érdekes kérdésre, hogy hogyan lehet speciális fastruktúrák számát meghatározni.

A számozott n pontú fák számát Cayley adta meg 1889-ben: n^{n-2}.

A klasszikus eredmények közül még megemlítjük Rényi tételét [97] a számozott n pontú, r végpontú (levelű) fákról, mely szerint ezek száma: $n!S(n-2,n-r)/r!$, ahol $S(n,k)$ a másodfajú Stirling számot jelöli. Mindkét állítás viszonylag könnyen igazolható a Prüfer-kódok segítségével.

A továbbiakban csak olyan speciális gyökeres, számozott levelű fastruktúrákat fogunk vizsgálni, amelyek agglomeratív hierarchikus clusterezések során állnak elő. A cluster analízis terminológiáját használva ezeket a fákat dendrogramnak is hívjuk. Az agglomeratív hierarchikus clusterezések során a megfigyeléseket reprezentáló n egyelemű clusterből indulunk ki és minden lépésben egy vagy több clustert egyesítünk, amíg a maximális, egyetlen n-elemű osztályból álló clusterezéshez nem jutunk. Ha minden clustert egy-egy ponttal reprezentálunk, melyeket akkor kötünk össze éllel, amikor éppen egyesítjük a megfelelő clustereket, akkor egy n számozott végpontú, a maximális osztálynak megfelelő gyökérpontú fát kapunk. A levelek számozása a fa csúcsainak egy cimkézését indukálja a következő módon. Minden levelet a hozzárendelt számot tartalmazó egyelemű halmazzal cimkézzük meg. A fa éleinek a segítségével a többi csúcsot is egyér-

telműen cimkézzük a már cimkézett szomszédok cimkehalmazainak az unió halmazával.

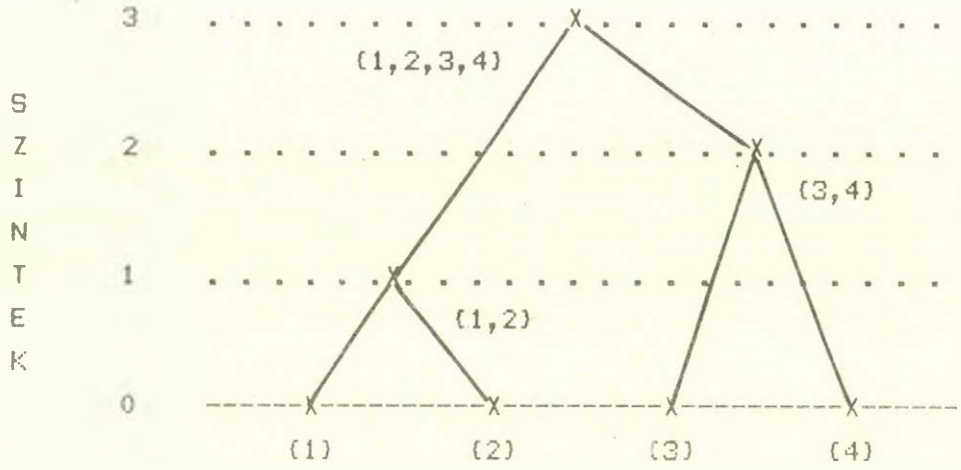
A végpontokból a gyökérhez vezető utak mentén a csúcsok cimkéi a tartalmazásra nézve szigorúan monoton növő halmazsorozatot alkotnak.

A csúcsokhoz szintszámokat is rendelhetünk. A leveleknek legyen 0 a szintszámuk. Minden egyesítésnél a keletkező cluster csúcsához az eddigi legnagyobb szintszámmal megegyező vagy eggyel nagyobb szintszámot rendelünk aszerint, hogy az egyesítő lépést az előző lépéssel egyszerre hajtjuk végre, vagy csak utána. Példa: a gyökérpont szintszáma 1, ha egy lépésben az összes egyelemű clustert egyesítjük, illetve n-1, ha minden lépésben ugyanazt a clustert növeljük egy egyelemű cluster beolvasztásával.

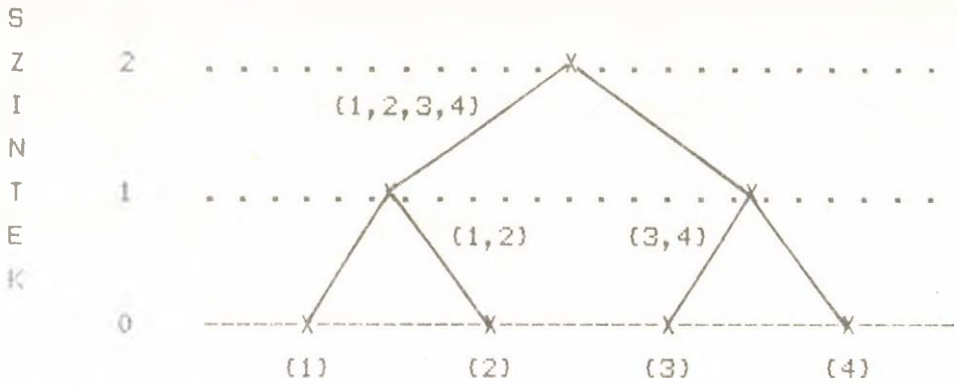
Két szint nélküli dendrogramot azonosnak tekintünk, ha megadható az egyik fa csúcsainak a másik csúcsaira való kölcsönösen egyértelmű, cimke és éltartó leképezése. Két szintezett dendrogramot azonosnak veszünk, ha a csúcsok között megadható kölcsönösen egyértelmű, cimke, él- és szintszámtartó leképezés.
A 3.1. és 3.2. ábrán látható dendrogramok szintezés nélkül azonosak, míg szintezéssel nem azok.

Egészítsük ki a szintezett dendrogramokat a fa éleire szükség szerint elhelyezett csúcsokkal úgy, hogy az azonos szinten levő csúcsok az eredeti halmaz egy particióját adják. Ezen partíciók sorozata a szintszám növekedő sorrendjében a partícióháló egy egyre finomodó láncát képezi. A partícióháló minden láncát reprezentálhatjuk egy-egy dendrogrammal, vagyis a két dolog között csak terminológiai különbség van.

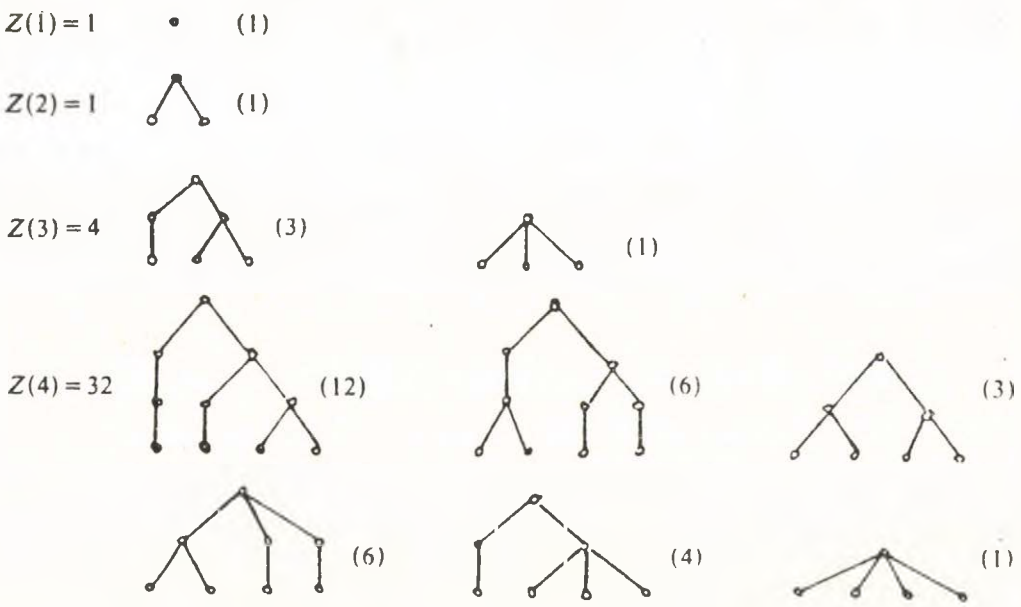
Murtagh cikkében [83] részletesen tárgyalja a különböző tipusú dendrogramok számára vonatkozó eredményeket. Murtagh különbséget tesz a számozott és számozatlan végpontú, bináris és nem bináris, szintezett és szint nélküli fák között. Bináris dendrogramhoz akkor jutunk, amikor pontosan n-1 egyesítési lépés után kaptuk meg a maximális clustert (vagyis amikor a fának 2n-1 csúcsa van). Schröder [99] 1870-ben vizsgálta azt a kérdést, hogy hány olyan cimkehalmaz rendszer van, aminek különböző (szint nélküli) dendrogramok felelnek meg.



3.1. ábra



3.2. ábra



3.3. ábra

Számunkra csak a szintezett dendrogramok érdekesek, ezért a továbbiakban dendrogram és fa alatt mindig ilyen fát értünk.

Kis n-ek esetén ($n \leq 4$) az összes ilyen fát felsoroljuk a 3.3. ábrán. A zárójelben álló számok a végpontok átszámozásából származó multiplicitásokat mutatják.

A cluster analízisben több helyen is szerephez jutnak az ultrametrikák. Egyfelől a single linkage eljárás (Kruskal-algoritmus) természetes módon definiál egy ultrametrikus távolságot a módszer során előforduló clusterek között. Másfelől ultrametrikus távolságok esetén egyszerűbb részproblémákhoz vezetnek bonyolult clusterezési problémák (5.2. és 5.4. pontok).

Most rátérünk az ultrametrikák leszámlálási feladatára.

3.1.3. definíció. Az X halmazon értelmezett kétváltozós, valós $d(x,y)$ $(x,y \in X)$ függvényt ultrametrikának nevezzük, ha metrika az X-en és $\forall x,y,z \in X$ hármasra a következő egyenlőtlenség teljesül

$$d(x,y) \leq \max\{d(x,z),d(z,y)\}. \qquad (3.1.1)$$

Az X halmazon értelmezett két ultrametrikát nem különböztetünk

meg egymástól és **ekvivalens**nek mondjuk őket, ha a single linkage eljárás során épített szintezett fák (dendrogramok) **azonos**ak.

Megjegyezzük, hogy erre a metrikára nézve minden háromszög egyenlőszárú. Ennél valamivel több is igaz: mindegyik háromszög vagy egyenlőoldalú vagy hosszabbik oldalai egyenlők. Nyilván a háromszögegyenlőtlenség következik a (3.1.1)-ből.

Példa. Egy összefüggő, súlyozott élű gráf akármelyik F feszítőfája segítségével ultrametrika definiálható a gráf csúcsain: legyen ugyanis $d_F(x,y)$ az x és y közötti egyetlen úton a leghosszabb (legnagyobb súlyú) él hossza.

Ha az F minimális (súlyú) feszítőfája a gráfnak, akkor a 6.3.2. állítás szerint $d_F(x,y) \leqslant d(x,y)$, ahol $d(x,y)$ jelöli az (x,y) él eredeti távolságát (súlyát). Ezt az ultrametrikát a d távolságfüggvényhez tartozó szubdomináns ultrametrikának is szokás nevezni, ti. bármely d^* ultrametrikára $d^* \leqslant d$ implikálja a $d^* \leqslant d_F$ egyenlőtlenséget. (Az utóbbi tulajdonság abból az egyszerűen igazolható szűk keresztmetszet (bottleneck) típusú eredményből következik, hogy a gráf tetszőleges két csúcsa között vezető utak közül éppen a minimális feszítőfában egyértelműen meghatározott út lesz az, amelyiken legkisebb a maximális él (súlya)).

Az ultrametrikák fontos tulajdonsága, hogy az n-elemű X halmazon értelmezett bármelyik ultrametrikus távolságfüggvény $\binom{n}{2}$ értéke "reprodukálható" a távolságfüggvény szerinti F minimális feszítőfa n-1 élhossza segítségével (Johnson [50]). Ekkor éppen az előző példa szerint definiált ultrametrika adja a megfelelő távolságértékeket, hiszen az ultrametrikus tulajdonságból $d_F(x,y) \geq d(x,y)$ következik. Az előzőek fontos következménye, hogy egy n-elemű halmazon értelmezett ultrametrikának legfeljebb n különböző értéke van (beleértve a d(x,x)=0 értéket is).

Tudomásunk szerint Schadertől [98] származik a következő

3.1.4. tétel. Egy n-elemű halmazon értelmezhető ultrametrikák ekvivalencia osztályainak száma megegyezik az Eq(n) partícióháló nem feltétlenül maximális láncainak a számával.

Az általánosság megszorítása nélkül választhatjuk az X=(1,2,...,n) halmazt a tételben mondott n-elemű halmaznak.

A 3.1.4. tétel bizonyítása. Azt kell mutatni, hogy minden ultrametrikához kölcsönösen egyértelműen megfeleltethetünk egy partíció láncot.

Először az ultrametrikához adjuk meg a láncot.

Legyen az ultrametrikának k különböző értéke és jelöljük ezeket $D_1(D_2(\ldots(D_k$ -val. Nyilván $D_1=0$, hiszen $x \in X$ esetén $d(x,x)=0$.

Az XxX halmaz $(C_1,C_2,\ldots,C_k)$ particióját a következőképpen definiáljuk: legyen $C_i = \{ (x,y) \mid d(x,y)=D_i \}$. Tekintsük a következő k relációt:

$\forall$ i=1,2,...,k esetén legyen

$$xS_iy \iff (x,y) \in C_1 \cup C_2 \cup \ldots \cup C_i \iff d(x,y) \leq D_i.$$

Az ultrametrika-tulajdonságból következik, hogy ezek ekvivalencia relációk az X-en.

Minden az X-en értelmezett ekvivalencia relációnak egyértelműen megfeleltethető az X egy particiója. Mivel xS_iy -ból következik, hogy $d(x,y) \leq D_i (D_{i+1}$, így $xS_{i+1}y$ is, tehát a megfelelő partíciók láncot alkotnak.

Fordított irányban rendeljük a partició láncot reprezentáló fához a csak a (számozott) végpontok között értelmezett következő ultrametrikus távolságot: legyen $d(i,j)$ az a legkisebb szintszám, ahol az i-vel és j-vel számozott levelek először kerülnek a dendrogram ugyanabba a részfájába. Definíció szerint legyen $d(i,i)=0$. (Könnyen látható, hogy teljesül a (3.1.1) egyenlőtlenség $\forall$ i,j,k=1,2,...,n számhármasra.)

A kapott ultrametrikának a kiindulási láncot felelteti meg a

bizonyítás elején tárgyalt leképezés.

A bizonyítás első részéről könnyen észrevehető a Kruskal-algoritmushoz fűződő kapcsolat: az 5.4.3. állítás és bizonyítása során látjuk majd, hogy a Kruskal-eljárás lépéseinek egy részsorozatát képeztük.

Ha a Kruskal-algoritmust úgy módosítjuk, hogy egy lépésnek tekintjük azokat a lépéseket, melyek során azonos hosszúságú éleket veszünk az erdőhöz, akkor jól látható ennek az eljárásnak az univerzális jellege: minden hierarchikus clusterezés reprodukálható vele (a 3.1.4. tétel bizonyításának második felében mondott módon). A 6.6. pontban illusztráljuk azokat az eseteket, amikor az ultrametrikus távolság közvetlen illetve közvetett kapcsolatban áll az eredeti távolságértékekkel.

Megjegyezzük, hogy fontos struktúrák számának meghatározása során máshol is előjön az a feladat, hogy számoljuk össze valamilyen hálóban a nem feltétlenül maximális láncokat. Például az n-elemű halmazon értelmezhető ún. különbözőségi mértékek száma egyenlő az n^2-elemű halmaz feletti részhalmaz háló nem feltétlenül maximális 0-1 láncainak a számával.

3.1.5. definíció. Az $X\times X$ halmazon értelmezett R bináris relációt különbözőségi mértéknek nevezzük az X-en, ha az R egy teljes, reflexív, tranzitív (ún. teljes preorder) reláció az $X\times X$-en.

Két különbözőségi mértéket azonosnak tekintünk, ha a megfelelő bináris relációk azonosak.

Példa: Minden d távolságfüggvényre a következő definíció egy R teljes preorder relációt határoz meg az $X\times X$ -en:

$$(x,y)R(u,v) \iff d(x,y)\leq d(u,v),\quad x,y,u,v\in X.$$

Könnyen bizonyítható a következő

3.1.6. állítás. Egy n-elemű halmaz részhalmaz hálójában a nem feltétlenül maximális 0-1 láncok $c(n)$ számára a következő összefüggések érvényesek.

$$c(0) = c(1) = 1,$$

$$c(n) = \sum_{k=1}^{n} \binom{n}{k} c(n-k) \qquad (n\geq 2), \tag{3.1.2}$$

$$c(n) = \sum_{k=1}^{n} k!S(n,k). \tag{3.1.3}$$

A háló minimális eleme a $\emptyset$ halmaz, a maximális pedig az X halmaz.

A 3.1.6. állítás bizonyítása. Legyen $\emptyset = S_1 \subset S_2 \subset \ldots \subset S_m = X$ a részhalmazok egy tetszőleges lánca.

Először a (3.1.2) rekurziót látjuk be. Jelölje k az S_2 halmaz elemeinek a számát. Ekkor a maradék n-k elemű halmaz c(n-k) láncával folytathatjuk az S_1, S_2 láncot, az S_2 halmaz elemeit pedig $\binom{n}{k}$-féleképpen választhatjuk.

A másik összefüggés bizonyításához rendeljük az $\emptyset = S_1 \subset S_2 \subset \ldots \subset S_k = X$ részhalmaz lánchoz az X következő partícióját:

$$Z_1 = S_1,$$

$$Z_{i+1} = S_{i+1} \setminus S_i, \quad i=1,2,\ldots,k-1.$$

összesen k! lánc adja ugyanezeket a partíció osztályokat és S(n,k) k-osztályú partíció van.

Tekintsünk egy tetszőleges $(x,y)R(u,v)$ $((x,y),(u,v) \in XxX)$ teljes preorder relációt az XxX -en. A 3.1.6. állítás bizonyításához hasonlóan particionáljuk az XxX halmazt, ti. legyen

$$C_{x,y} = \{ (u,v) \mid (x,y)R(u,v) \text{ és } (u,v)R(x,y) \},$$

$$\forall (x,y) \in XxX.$$

A definícióból azonnal látható, hogy az R ekvivalencia reláció a $C_{x,y} \subseteq XxX$ részhalmazon. Nyilván $(x,x) \in C_{x,x}$. A definícióból

adódik, hogy két ilyen halmaz azonos vagy diszjunkt, tehát jogosan beszélhetünk az XxX particiójáról. Jelöljе k a különböző $C_{x,y}$ halmazok számát. Ennyi $C_{x,y}$ halmazt $S(n^2,k)$ féleképpen választhatunk. Bárhogyan is veszünk ki egy-egy reprezentánst a különböző $C_{x,y}$ halmazokból, ezeken már - az antiszimmetria miatt - teljes rendezés lesz az R, és k elemen k! ilyen rendezés definiálható.

Felhasználva a 3.1.6. állítást, most már kimondható a

3.1.7. tétel. Egy n-elemű halmazon értelmezhető különbözőségi mértékek száma megegyezik az n^2-elemű halmaz részhalmaz hálójában a nem feltétlenül maximális 0-1 láncok számával.

A szokásos $\leq$ reláció a valós számokon egy teljes preorder reláció. Ezért a 3.1.7. tételhez hasonlóan bizonyítható a következő

3.1.8. tétel. n nem feltétlenül különböző kulcs összes lehetséges sorrendjének a száma megegyezik az n-elemű részhalmaz háló nem feltétlenül maximális 0-1 láncainak a számával.

3.1.9. megjegyzés. A (3.1.2) rekurzió segítségével a

$$(3.1.4) \qquad C(x) = \sum_{n=1}^{\infty} c(n) \frac{x^n}{n!}$$

exponenciális generátor függvényre

$$(3.1.5) \qquad C(x) = 1 / (2-e^x)$$

adódik. A Cauchy-formulával

$$c(n)/n! = \frac{1}{2\pi i} \int_{|z|=\varepsilon} (2-e^z)^{-1} z^{-n-1} dz$$

adódik, ahol ε egy tetszőlegesen kicsi pozitív szám.

Az integrandusnak a $z = 0$ és a $z = \ln 2 + 2k\pi i$ helyeken van pólusa. Alkalmasan megválasztva az integrálás görbéjét kapjuk a

$$(3.1.6) \qquad c(n) \sim \frac{1}{2} \frac{n!}{(\ln 2)^{n+1}}$$

aszimptotikus összefüggést (Lovász [78]).

Bender [8] az előbbi módszer általánosításával adott meg aszimptotikus összefüggést a nem feltétlenül maximális láncok számára binomiális posetekben.

Megjegyezzük, hogy Barthelemy [6] a teljes preorder relációk számára a (3.1.3) összefüggésből vezette le a (3.1.5) for-

mulát és adott meg a (3.1.6)-hez hasonló aszimptotikus becslést. A (3.1.2) rekurzióhoz is eljutott a (3.1.4) exponenciális generátor függvény deriválásával. A láncokkal való jellemzésre az irodalomban nem találtunk utalást.

3.2. Az Eq(n) partícióháló nem feltétlenül maximális 0-1 láncainak a száma

Ebben a pontban térünk rá a leszámlálási feladat megoldására. A továbbiakban jelölje Z(n) az Eq(n) partícióháló nem feltétlenül maximális 0-1 láncainak a számát. Először a Z(n) meghatározására alkalmas rekurzív összefüggést (3.2.2. állítás) adunk meg.

A rekurzió felhasználásával bizonyítjuk be a fejezet legfontosabb eredményét: a 3.2.3. tételt, amely a Z(n) aszimptotikus nagyságrendjét állapítja meg.

A 4. fejezetben tovább vizsgáljuk a Z(n) aszimptotikus viselkedését és bebizonyítjuk, hogy ez a mennyiség és a 3.2.3. tételben szereplő becslése aszimptotikusan arányos.

A kérdéses mennyiséghez kapcsolódóan megemlíthető a

3.2.1. megjegyzés. Az Eq(n) maximális 0-1 láncainak a száma: $n!(n-1)!/2^{n-1}$.

A nem feltétlenül maximális 0-1 láncok Z(n) számára a következő rekurzió írható fel.

3.2.2. állítás.

$$Z(n) = \sum_{k=1}^{n-1} S(n,k)\, Z(k), \qquad (n \geq 2). \tag{3.2.1}$$

Bizonyítás. Nyilvánvaló, hiszen az első szinten S(n,k) féleképpen lehet az n mintaelemet k osztályba particionálni és erről a szintről újraindítva Z(k) hierarchikus struktúra van.

A fejezet legfontosabb eredménye a

3.2.3. tétel. Létezik olyan C_1 és C_2 pozitív konstans, hogy

$$C_1 \leq Z(n)/f(n) \leq C_2,$$

ahol $f(n) = (n!)^2 (2 \ln 2)^{-n} n^{-1-(\ln 2)/3}$ (ln a természetes logaritmust jelöli).

3.2.4. megjegyzés. Ha bevezetjük a

$$G(x) = \sum_{n=1}^{\infty} Z(n) \frac{x^n}{n!}$$

(divergens) exponenciális generátor függvényt, akkor a következő

függvényegyenletet kapjuk:

$$2\,G(x) = G(e^{x}-1) + x.$$

A 3.2.3. tétel bizonyítása [68]. Mivel a (3.2.1) jobb oldalán azok a tagok fognak dominálni, amelyek az n-hez közel vannak, ezért a k indexet n-k -val helyettesítjük.

A túl gyorsan növő Z(n) helyett bevezetjük a

$$Z^{*}(n) = Z(n)2^{n}/(n!)^{2}$$

mennyiséget.

Ezzel az általakítással a következő rekurzióhoz jutunk

$$(3.2.2) \qquad Z^{*}(n) = \sum_{k=1}^{n-1} a(n,k)\, Z^{*}(n-k), \qquad (n \geq 2),$$

ahol

$$(3.2.3) \qquad a(n,k)=S(n,n-k)2^{k}/[n]_{k}^{2}.$$

(Itt az $[n]_k$ jelölést használjuk a $k!\binom{n}{k}$ kifejezés helyettesítésére).

Mint látni fogjuk, $a(n,k) \sim 1/k!$, ha $1 \leq k < n^{1/3-\varepsilon}$ (ahol ε tetszőlegesen kicsi pozitív szám) és (3.2.2) jobb oldalán a $k \geq n^{1/5}$ tagoknak a hozzájárulása elhanyagolható, sőt ez még $k)C'ln»$ esetén is igaz.

A $Z^*(n)$ aszimptotikus nagyságrendjének a meghatározásához az $a(n,k)$ finomabb közelítésére lesz szükségünk (3.2.8'. lemma), ti. olyanra, amelyik az első hibatagot is tartalmazza és a hibát $k^6 O(1/n^2)$ -ra redukálja.

Konkrétan megadunk egy olyan $y(n)$ függvényt, ami "majdnem" eleget tesz a (3.2.2) rekurziónak, legyen ugyanis

$$y(n) = (\ln 2)^{-n}\, n^{-1-(\ln 2)/3}. \tag{3.2.4}$$

3.2.5. lemma.

$$\sum_{k=1}^{n-1} a(n,k)\, y(n-k) = y(n)\cdot(1+O(1/n^2)). \tag{3.2.5}$$

Ez az eredmény a 3.2.6. lemma és a 3.2.7. következmény értelmében garantálja, hogy a $Z^*(n)$ és az $y(n)$ aránya két véges pozitív konstans között marad.

A lemma bizonyítására a 3.2.10. lemma bizonyítása után térünk vissza.

Az aszimptotikus nagyságrend meghatározásához szükségünk van a Babai Lászlótól származó következő, általános jellegű észrevételre:

3.2.6. lemma (Babai [4]). Legyen $x(n)$ és $y(n)$ valós számok egy-egy sorozata. Tegyük fel, hogy az $x(n)$ eleget tesz a követ-

kező rekurziónak

$$(3.2.6) \qquad x(n) = \sum_{k=1}^{n-1} c(n,k)\, x(n-k), \quad (n \geq 2),$$

ahol $c(n,k) \geq 0$, $\forall\ 1 \leq k \leq n-1$ esetén. Tegyük fel továbbá, hogy valamilyen N-re $\forall\ n \geq N$ esetén

$$(3.2.7) \qquad (1-\varepsilon_n)\, y(n) \leq \sum_{k=1}^{n-1} c(n,k)\, y(n-k) \leq (1+\varepsilon_n)\, y(n),$$

ahol $0 \leq \varepsilon_n < 1$. Ekkor $\forall\ n \geq N$-re

$$(3.2.8) \qquad \frac{A}{\prod_{j=N}^{n}(1+\varepsilon_j)} \leq \frac{y(n)}{x(n)} \leq \frac{B}{\prod_{j=N}^{n}(1-\varepsilon_j)},$$

ahol $A = \min_{1 \leq j \leq N-1} y(j)/x(j)$ és

$B = \max_{1 \leq j \leq N-1} y(j)/x(j)$.

Ezek után azonnal adódik a

3.2.7. következmény. Ha a 3.2.6. lemma feltételei teljesülnek és $\sum_{j=N}^{\infty} \varepsilon_j < \infty$, akkor $\exists\ C_1$ és C_2 pozitív valós, hogy $\forall n$-re

$$C_1 \leq x(n)/y(n) \leq C_2.$$

Ez a következmény és a 3.2.5. lemma már elegendő lesz olyan $0 < C_1 \leq C_2 < \infty$ konstansok létezésének a biztosítására, melyekre

$\forall n$ esetén

$$C_1 f(n) \leq Z(n) \leq C_2 f(n),$$

ami megadja a $Z(n)$ aszimptotikus nagyságrendjét. Ezzel a 3.2.3. tételt bizonyítottuk. Most rátérünk az említett lemmákra.

A másodfajú Stirling-számokra Hsu [45] a következő, tetszőleges fix k esetén alkalmazható becslést adta

$$S(n,n-k) = \frac{[n]_k^2}{2^k k!} \left(1 - \frac{k(k+2)}{3n} + O\left(\frac{1}{n^2}\right) \right).$$

Ezt a becslést a k lehetséges értékeire vonatkozóan terjeszti ki a

3.2.8. lemma. Létezik olyan C'' abszolút konstans, hogy $\forall k^3 < n/16$ esetén

$$(3.2.9) \quad \left| S(n,n-k) - \frac{[n]_k^2}{2^k k!} \left(1 - \frac{k(k+2)}{3n} \right) \right| \leq C'' \frac{[n]_k^2}{2^k k!} \frac{k^6}{n^2}.$$

Ezzel ekvivalens az $a(n,k)$-ra vonatkozó

3.2.8'. lemma. $\forall k^3 < n/16$ esetén

(3.2.10)

$$\left| a(n,k) - \frac{1}{k!}\left(1 - \frac{k(k+2)}{3n}\right)\right| \leq C'' \frac{1}{k!}\frac{k^6}{n^2}.$$

A 3.2.8. lemma által csak részben érintett k értékekre ad egy felső korlátot a

3.2.9. lemma. $\forall$ k: $3\ln n/\ln\ln n < k \leq n-1$ esetén

$$S(n,n-k) < [n]_k^2/(n^{2k}),$$

midőn az n elég nagy.

Az a(n,k) -ra vonatkozó ekvivalens alakot adja meg a

3.2.9'. lemma. $\forall$ k: $3\ln n/\ln\ln n < k \leq n-1$ esetén, minden elég nagy n-re

$$a(n,k) < 1/n^2.$$

Most rátérünk a 3.2.8. és 3.2.9. lemma bizonyítására.

A 3.2.8. lemma bizonyítása. Az S(n,n-k) számolja egy n-elemű halmaz (n-k) -particióinak a számát. Egy partició s_1 egyelemű osztályból, s_2 párból, ..., s_n n-elemű osztályból áll. Egy ilyen partició automorfizmus csoportjának a rendje

$$F(s_1,\ldots,s_n)=$$

$$= s_1!(1!)^{s_1} s_2!(2!)^{s_2} \ldots s_n!(n!)^{s_n}.$$

Minden $k,L>0$ esetén jelölje $\mathcal{L}(k,L)$ azon $(s_1,\ldots,s_n)$ nemnegatív szám n-esek halmazát, melyekre

$$(3.2.11) \qquad \sum_{i=1}^{n} is_i = n,$$

$$(3.2.12) \qquad \sum_{i=2}^{n} (i-1)s_i = k,$$

$$(3.2.13) \qquad \sum_{i=3}^{n} (i-2)s_i = L.$$

Világos, hogy $\mathcal{L}(k,L)=\emptyset$, ha $L<k$; és $s_{L+3}=\ldots=s_n=0$ $\forall (s_1,\ldots,s_n) \in \mathcal{L}(k,L)$ esetén. Könnyen látható, hogy

$$S(n,n-k) = \sum_{(s_1,\ldots,s_n)} \frac{1}{s_1!\ldots s_n!} \binom{n}{1^{s_1},\ldots,n^{s_n}},$$

ahol az összegzést a (3.2.11) és (3.2.12) egyenlőséget kielégítő összes nemnegatív szám n-esen értendő, és a zárójel egy multinomiális együtthatót tartalmaz. Csoportosítsuk azokat az n-eseket, amelyekre

$$\sum_{i=3}^{n} (i-2)s_i = L.$$

Ekkor

$$S(n,n-k) = \sum_{L=0}^{k} \left([n]_{2k-L} \cdot \sum_{\mathcal{L}(k,L)} \frac{s_1!}{F(s_1,\ldots,s_n)} \right).$$

Bebizonyítjuk, hogy $k^3=o(n)$ esetén a jobb oldalon álló kifejezések gyorsan csökkenek, amikor az L növekszik. Valójában már L=2 esetén is csak a (3.2.9) maradék tagjához járulnak hozzá. Nyilván $\mathcal{L}(k,0)=((n-2k,k,0,\ldots,0))$ és a megfelelő tag

$$\frac{[n]_{2k}}{k!\,2^k} = \frac{[n]^2_k}{k!2^k}\left(1 - \frac{k^2}{n} + k^4 O(1/n^2)\right),$$

hasonlóan $\mathcal{L}(k,1)=((n-2k+1,k-2,1,0,\ldots,0))$ és így a második tag

$$\frac{[n]_{2k-1}}{(k-2)!2^{k-2}3!} = \frac{[n]^2_k}{k!2^k}\left(\frac{2k(k-1)}{3n} + k^4 O(1/n^2)\right).$$

A további tagokra felső korlátot adunk meg. Könnyen látható, hogy

$$F(s_1,\ldots,s_n) \geq s_1!(k-2L)!2^{k-2L}.$$

Végül vegyük észre, hogy

$$|\mathcal{L}(k,L)| \leq (2k)^L.$$

Ez utóbbi abból következik, hogy (3.2.12) és (3.2.13) szerint a bal oldalon álló szám kisebb, mint az $L=x_1+\ldots+x_k$ egyenlet nemnegatív egész megoldásainak a száma. Végül

$$\sum_{L=2}^{k} [n]_{2k-L} \sum_{\mathcal{L}(k,L)} \frac{s_1!}{F(s_1,\ldots,s_n)} \leq \sum_{L=2}^{k} [n]_{2k-L} \frac{(2k)^L}{(k-2L)!\,2^{k-2L}} \leq$$

$$\leq \frac{[n]_k^2}{k!2^k} \sum_{L=2}^{k} \frac{1}{n^L} k^{2L} 2^{2L} (2k)^L = \frac{[n]_k^2}{k!2^k} \sum_{L=2}^{k} \left(\frac{8k^3}{n}\right)^L <$$

$$< \frac{[n]_k^2}{k!2^k} \frac{128k^6}{n^2},$$

amint $n \to \infty$. (Ha $k<2L$, akkor a $(k-2L)!$ kifejezés értékét 1-nek vesszük.)

A 3.2.9. lemma bizonyítása. Tetszőleges n és k mellett nyilván $S(n,k) \leq k^n/k!$. A k helyére $n-k$-t helyettesítve egyszerű számítások után elég nagy n-re

$$a(n,k) < 2^k e^n / [n]_k$$

adódik. Amennyiben $n/2 \leq k \leq n-1$, akkor a lemma állítása azonnal igazolható. Általában

$$\frac{a(n,k+1)}{a(n,k)} = \frac{S(n,n-k-1)}{S(n,n-k)} \frac{2}{(n-k)^2}.$$

A $S(n,1), S(n,2), \ldots, S(n,n)$ sorozat logaritmikusan konkáv [39], azaz

$$\frac{S(n,k)}{S(n,k+1)} \geq \frac{S(n,k-1)}{S(n,k)},$$

aminek ismételt alkalmazásával minden $m \leq k \leq n-1$ esetén

$$\frac{a(n,k+1)}{a(n,k)} \leq \frac{2}{(n-k)^2} \frac{S(n,n-m)}{S(n,n-m+1)}.$$

Ha $m^5 < n$, akkor a (3.2.9) felhasználásával

$$\frac{S(n,n-m)}{S(n,n-m+1)} = \frac{(n-m+1)^2}{2m}(1+o(1)).$$

Amennyiben ezen kívül még $k < n/2$ is teljesül, akkor az előző egyenlőtlenségből és egyenlőségből kapjuk, hogy

$$\frac{a(n,k+1)}{a(n,k)} < 2 \frac{2n^2}{(n/2)^2 2m} = \frac{8}{m},$$

ha n elég nagy.

A (3.2.10) egyenlőtlenségből azt kapjuk, hogy $8 \leq m < k < n/2$ és $m^5 < n$ esetén

$$a(n,k) < a(n,m) \leq 2/m!.$$

Végül legyen $m=[3\ln n/\ln\ln n]$. Ekkor $a(n,k) < 2/m! < 1/n^2$ adódik, amikor n elég nagy.

A 3.2.5. lemma bizonyításához szükségünk lesz még egy közelítő összegre.

3.2.10. lemma. Legyen f és i tetszőleges valós számok illetve q olyan valós, melyre $|q|<1$. Ekkor minden olyan m esetén, amelyre $m=o(n)$ és $m!>n^2$

$$\sum_{k=1}^{m} \frac{q^k}{k!}\left(1-\frac{k+i}{n}\right)=e^q-1-\frac{f(q+i)(e^q-1)+fq}{n}+O_{f,i,q}(1/n^2),$$

amint $n\to\infty$.

Bizonyítás. Az előző egyenlőség bal oldalán álló kifejezést a következőképpen fejtjük ki:

$$\sum_{k=1}^{m} \frac{q^k}{k!}\left(1-\frac{f(k+i)}{n}+\frac{k^2}{n^2}O_{f,i,q}(1)\right)=$$

$$=\sum_{k=1}^{m}\left(\frac{q^k}{k!}-\frac{fq^k}{n}-q\frac{q^{k-1}}{(k-1)!}-\frac{fiq^k}{n\,k!}\right)+$$

$$+O_{f,i,q}(1/n^2)\sum_{k=1}^{m}\frac{q^k}{k!}k^2=$$

$$=e^q-1-\frac{fq}{n}e^q-\frac{fi}{n}(e^q-1)+O_{f,i,q}(1/n^2)=$$

$$=e^q-1-\frac{f(q+i)(e^q-1)+fq}{n}+O_{f,i,q}(1/n^2),$$

midőn $n\to\infty$ (a $m!>n^2$ feltételt az utolsó előtti egyenlőségnél használtuk).

A 3.2.5. lemma bizonyítása. A $m=[n^{1/5}]$, $f=-1-(\ln 2)/3$, $q=\ln 2$ helyettesítéssel alkalmazzuk a 3.2.10. lemmát. Elég nagy n-re kapjuk, hogy

$$(3.2.14)\qquad \sum_{k=1}^{m} \frac{1}{k!}\left(1-\frac{k(k+2)}{3n}\right) y(n-k) =$$

$$= y(n)\sum_{k=1}^{m} \frac{(\ln 2)^k}{k!}\left(1-\frac{k(k-1)}{3n}-\frac{3k}{3n}\right)\left(1-\frac{k}{n}\right)^f =$$

$$= y(n)\cdot\left(1-\frac{1}{n}\left(2f\cdot\ln 2+\frac{2}{3}(\ln 2)^2+2\ln 2\right)+O(1/n^2)\right) =$$

$$= y(n)\cdot(1+O(1/n^2)).$$

A 3.2.9'. lemma felhasználásával kapjuk, hogy

$$\sum_{k=m+1}^{n-1} a(n,k)y(n-k) \ll \frac{1}{n^2}\sum_{k=m+1}^{n-1} y(n-k) \ll$$

$$\ll \frac{y(n)}{n^2}\, n(\ln 2)^{m+1} n^{-f} = y(n)\cdot O(1/n^2).$$

Az előbb éppen azt láttuk be, hogy a

$$\sum_{k=1}^{n-1} a(n,k)y(n-k) = \sum_{k=1}^{m} a(n,k)y(n-k) + \sum_{k=m+1}^{n-1} a(n,k)y(n-k)$$

összegben a második tag $y(n)\cdot O(1/n^2)$. A 3.2.8'. lemma szerint az első tag csak $y(n)\cdot O(1/n^2)$-val tér el a (3.2.14) bal oldalán

álló mennyiségtől. Ezzel a (3.2.5) egyenlőséget bizonyítottuk.

Ezek után a 3.2.5. lemma felhasználásával belátható, hogy a 3.2.6. lemma (3.2.6) és (3.2.7), valamint a 3.2.7. következmény feltételei az

$$x(n) = Z^{*}(n),$$

$$c(n,k) = a(n,k),$$

$$\varepsilon_n = O(1/n^2)$$

választás mellett teljesülnek. Ezzel a 3.2.3. tételt igazoltuk.

4. Egy általános konvergencia kritérium rekurzióval definiált sorozatokra

Ebben a fejezetben egy elég általános konvergencia kritériumot bizonyítunk be és azt alkalmazzuk a partició láncok leszámlálására. Ezzel a módszerrel sikerült megmutatni, hogy az előző mennyiség és becsült értéke aszimptotikusan arányos, bár az arány nagyságára csak numerikus számítások alapján tudtunk következtetni.

4.1. A kritérium alkalmazása a particióháló láncainak leszámlálására

A fejezet legfontosabb eredménye a

4.1.1. tétel. A $\lim_{n\to\infty} Z(n)/f(n) = C$ határérték létezik (a C egy pozitív konstans, melyre numerikus számítások a $C \approx 1.1$ becslést sugallják, vö.: 4.1. táblázat).

A következő lemma - a megfelelő helyettesítéssel alkalmazva, a 3.2.5. lemma felhasználásával - adja a 4.1.1. tétel bizonyítását.

4.1.2. lemma [74]. Tegyük fel, hogy az $x(n)$ valós számsorozat eleget tesz a (3.2.6) rekurziónak $\forall$ $n=N,N+1,N+2,\ldots$-ra. A $c(n,k)$ együtthatók legyenek nemnegatívak ($n=N,N+1,\ldots$ és

k=1,2,...), valamint a sorösszegek legyenek az egyhez közel

$$(4.1.1) \qquad \sum_{k=1}^{\infty} c(n,k) = 1+\gamma_n \qquad (n=N,N+1,\ldots),$$

ahol

$$(4.1.2) \qquad \gamma_n > -1$$

és

$$(4.1.3) \qquad \sum_{n=N}^{\infty} |\gamma_n| < \infty .$$

Tegyük fel továbbá, hogy az

(4.1.4)

$$\varepsilon_{n,k} = 1+\gamma_n - \sum_{i=1}^{k} c(n,i) \quad \left(= \sum_{i=k+1}^{\infty} c(n,i) \right),$$

mennyiségekhez létezik olyan egyenletes K_1, K_2 korlát, hogy

$$(4.1.5) \qquad 0 < K_1 \leq \sum_{k=1}^{\infty} \varepsilon_{m+k,k} \leq K_2 < \infty$$

$\forall$ m≥N -re teljesül.

A fenti feltételek fennállása esetén az x(n) sorozat egy véges pozitív számhoz konvergál.

4.1.1. tétel bizonyítása. Először az x(n) sorozat korlátosságát látjuk be. Legyen a^+ illetve a^- az a valós szám pozitív, illetve negatív része, azaz $a^+=(a+|a|)/2$ és $a^-=(a-|a|)/2$.

Legyen A=max(x(n): 1≤n(N). Ekkor indukcióval könnyen látható,

hogy $\forall\, n \geq N$

$$x(n) \leq A \prod_{i=N}^{n} (1 + \gamma_i^+),$$

és (4.1.3) miatt a jobb oldalon álló kifejezés korlátos.

Hasonlóan adódik, hogy $\forall n \geq N$

$$\underline{\lim}\ x(n) \geq B \prod_{i=N}^{n} (1 + \gamma_i^-),$$

ahol $B=\min(x(n):\ 1 \leq n(N)$. Legyen a továbbiakban M tetszőleges olyan pozitív valós, melyre

$$0 < M < \overline{\lim}\ x(n).$$

Olyan n_i és β_i sorozatokat definiálunk, hogy $n_i \to \infty$ és $\beta_i \to 1$ (midőn $i \to \infty$), valamint $\forall i \geq 0$ és $n \geq n_i$ esetén

$$x(n) \geq \beta_i M.$$

Ebből következik már, hogy

$$\underline{\lim}\ x(n) \geq M.$$

Mivel a fenti gondolatmenet bármilyen $M < \overline{\lim}\ x(n)$ esetén működik, ezért ezzel a tétel állítását kapjuk, azaz

$$\underline{\lim}\ x(n) = \overline{\lim}\ x(n) = \lim x(n).$$

A (4.1.3)-ból következik, hogy

$$(4.1.6) \qquad 1 \geq \gamma = \prod_{k=N}^{\infty} (1 + \gamma_k^-) > 0$$

és

$$(4.1.7) \qquad \infty > \delta = \prod_{k=N}^{\infty} (1 + \gamma_k^+).$$

A továbbiakban szükségünk lesz a következő mennyiségekre $\forall m \geq N$ esetén

$$(4.1.8) \qquad \alpha(m) = \prod_{k=1}^{\infty} \left(1 - \frac{\varepsilon_{m+k,k}}{1+\gamma_{m+k}}\right) < 1.$$

A (4.1.3) és (4.1.5) feltételek következtében

$$\lim_{n\to\infty} \gamma_n = 0$$

és $\forall m \geq N$-re

$$\lim_{k\to\infty} \varepsilon_{m+k,k} = 0.$$

Ha az m elég nagy, akkor

$$\exp\left(-2\sum_{k=1}^{\infty} \frac{\varepsilon_{m+k,k}}{1+\gamma_{m+k}}\right) \leq \alpha(m) \leq \exp\left(-\sum_{k=1}^{\infty} \frac{\varepsilon_{m+k,k}}{1+\gamma_{m+k}}\right).$$

A (4.1.5) szerint létezik olyan a és b valós szám, hogy

$$(4.1.9) \qquad 0 < a \leq \alpha(m) \leq b < 1$$

minden elég nagy $m \geq N$ esetén.

Legyen $N = m_1 < m_2 < \ldots$ olyan pozitív egészekből álló sorozat, hogy $\forall j \geq 1$ esetén

(4.1.10) $\qquad x(m_j) \geq M$

és $\alpha(m_j)$ konvergál:

(4.1.11) $\qquad \lim_{j\to\infty} \alpha(m_j) = \alpha .$

Világos, hogy $0<\alpha<1$. A (4.1.5) és (4.1.3) szerint $\forall n\geq N$ és $i\geq 1$ mellett létezik olyan $s_{n,i}\geq N$ egész, hogy

$$(4.1.12) \qquad \sum_{k=s_{n,i}}^{\infty} \varepsilon_{n+k,k} < \frac{1-\gamma b}{2\sqrt{(i+1)}} \qquad (i=1,2,\ldots)$$

és

$$(4.1.13) \qquad \sum_{k=s_{n,i}}^{\infty} |\gamma_k| < \frac{1-\gamma b}{2\sqrt{(i+1)}} \qquad (i=1,2,\ldots).$$

Legyen az $(n_i)_{i=1}^{\infty}$ az $(m_j)_{j=1}^{\infty}$ olyan részsorozata, hogy $i=1,2,\ldots$ esetén

$$(4.1.14) \qquad n_{i+1}-n_i \geq s_{n_i,i+1}.$$

Definiáljuk a következő β_i sorozatot $(i=0,1,2,\ldots)$:

$$(4.1.15) \qquad \beta_0 = 0$$

$$\beta_{i+1} = \beta_i\left(1-\gamma\alpha(n_{i+1})-\frac{1-\gamma b}{i+1}\right) + \gamma\alpha(n_{i+1}).$$

A (4.1.9) és a (4.1.11) kifejezésből könnyen látható, hogy

$$0 \leq \beta_i \leq 1,$$

valamint

$$\lim_{i\to\infty} \beta_i = 1.$$

A j-re vonatkozó indukcióval bizonyítjuk, hogy a

(4.1.16) $\qquad x(n) \geq \beta_j M$

egyenlőtlenség $\forall n \geq n_j$ $(j \geq 0)$ esetén fennáll.

Legyen $x(0)=0$. Az állítás nyilván igaz lesz $j=0$-ra, ha $n_0=0$-nak választjuk. Legyen $i \geq 0$ és tegyük fel, hogy (4.1.16) fennáll $j=0,1,\ldots,i$ esetén. Most belátjuk, hogy (4.1.16) akkor is teljesül, ha $j=i+1$.

Definiáljuk $m \geq n$ esetén az

$$\bar{x}_n^m = \min\{x(n),\ldots,x(m)\}$$

mennyiségeket.

A (3.2.6) rekurzióból $\forall n > n_{i+1}$ következik, hogy

$$x(n) \geq \{c(n,1)x(n-1)+\ldots+c(n,n-n_{i+1})x(n_{i+1})\} +$$

$$+ \{c(n,n-n_{i+1}+1)x(n_{i+1}-1)+\ldots+c(n,n-n_i)x(n_i)\},$$

és így

$$(4.1.17) \quad x(n) \geq \left(1+\gamma_n - \varepsilon_{n,n-n_{i+1}}\right)\bar{x}_{n_{i+1}}^{n-1} +$$

$$+\left(\varepsilon_{n,n-n_{i+1}} - \varepsilon_{n,n-n_i}\right)\beta_i M.$$

A (4.1.17)-ből következik, hogy

$$(4.1.18)\quad x(n)-\beta_i M \geq \left(\bar{x}_{n_{i+1}}^{n-1} - \beta_i M\right)\left(1+\gamma_n - \varepsilon_{n,n-n_{i+1}}\right) +$$

$$+ \beta_i M \gamma_n - \beta_i M \varepsilon_{n,n-n_i}.$$

Az n-re vonatkozó indukcióval $n>n_{i+1}$ kapjuk, hogy

$$(4.1.19)\quad x(n)-\beta_i M \geq$$

$$\geq \left(x(n_{i+1})-\beta_i M\right) \prod_{k=n_{i+1}+1}^{n} (1+\gamma_k^-) \prod_{k=1}^{n-n_{i+1}} \left(1-\frac{\varepsilon_{n_{i+1}+k,k}}{1+\gamma_{n_{i+1}+k}}\right) +$$

$$+ \beta_i M \left(\sum_{k=n_{i+1}+1}^{n} \gamma_k^-\right) \prod_{k=n_{i+1}+1}^{n} (1+\gamma_k^+) -$$

$$- \beta_i M \left(\sum_{k=n_{i+1}+1-n_i}^{n-n_i} \varepsilon_{n_i+k,k}\right) \prod_{k=n_{i+1}+1}^{n} (1+\gamma_k^+).$$

Végül a (4.1.6)-(4.1.8), (4.1.10) és a (4.1.12)-(4.1.14) miatt

$$x(n) \geq \beta_i M + \left(x(n_{i+1}) - \beta_i M\right)\gamma\alpha(n_{i+1}) - \beta_i M \frac{2\delta(1-\gamma b)}{2\delta(i+1)}$$

$$= M\left\{\beta_i\left(1-\gamma\alpha(n_{i+1}) - \frac{1-\gamma b}{i+1}\right) + \gamma\alpha(n_{i+1})\right\} =$$

$$= \beta_{i+1} M.$$

A következő lépésben a 4.1.2. lemmát alkalmazzuk a

$$x(n) = Z^{*}(n)/y(n)$$

sorozatra.

A (3.2.2) és a (3.2.4) alapján világos, hogy ($n \geq 2$)

$$(4.1.20) \qquad x(n) = \sum_{k=1}^{n-1} c(n,k)x(n-k),$$

ahol

$$(4.1.21) \qquad c(n,k) = a(n,k)\,\frac{y(n-k)}{y(n)}.$$

Ellenőriznünk kell még a 4.1.2. lemma feltételeit. Ezzel a 4.1.1. tétel bizonyítása teljessé válik.

A 3.2.5. lemma szerint

$$(4.1.22) \qquad \sum_{k=1}^{n-1} c(n,k) = 1 + \gamma_n,$$

ahol $\gamma_n = O(1/n^2)$ és $\gamma_n > -1$ $\forall n \geq 2$ esetén.

A (4.1.21)-ből

$$c(n,k) = \begin{cases} a(n,k)\,(\ln 2)^{k}\,(1-\frac{k}{n})^{f}, & \text{ha } k:\ 1 \leq k \leq n-1, \\ 0, & \text{ha } k \geq n, \end{cases}$$

ahol $f = -1 - (\ln 2)/3 < 0$.

Még meg kell mutatnunk, hogy (4.1.5) is teljesül.

Ehhez tekintsük az

$$(4.1.23)\qquad \varepsilon_{m+k,k} = \sum_{i=k+1}^{\infty} c(m+k,i) = \sum_{i=k+1}^{m+k-1} a(m+k,i)\cdot(\ln 2)^i\cdot\left(1-\frac{i}{m+k}\right)^f .$$

Legyen M_1 olyan nagy egész szám, hogy $\forall m \geq M_1$-re

$$(4.1.24)\qquad a(m,i) < 1/m^2, \text{ ha } i \geq m^{1/4}$$

és

$$(4.1.25)\qquad \frac{1}{2i!} < a(m,i) < \frac{2}{i!}, \text{ ha } i \leq m^{1/4} .$$

A 3.2.8'. és a 3.2.9'. lemma garantálja ilyen M_1 létezését.

Definiáljuk tetszőleges fix $m>0$ esetén a

$$(4.1.26)\qquad k_0(m) = \max\{k \text{ egész: } (m+k)^{1/4})k \geq 1\}$$

sorozatot. Nyilván, $\forall k$: $1 \leq k \leq k_0(m)$ esetén teljesül, hogy

$$(4.1.27)\qquad (m+k)^{1/4} > k .$$

Ezek szerint $\forall m \geq M_1$ esetén

$$(4.1.28)\qquad \sum_{k=1}^{\infty} \varepsilon_{m+k,k} - \sum_{k=1}^{k_0(m)} \left\{ \sum_{i=k+1}^{[(m+k)^{1/4}]} a(m+k,i)\cdot(\ln 2)^i\cdot\left(1-\frac{i}{m+k}\right)^f + \right.$$

$$\left. + \sum_{i=[(m+k)^{1/4}]+1}^{m+k-1} a(m+k,i)\cdot(\ln 2)^i\cdot\left(1-\frac{i}{m+k}\right)^f \right\} +$$

$$+ \sum_{k=k_0(m)+1}^{\infty} \sum_{i=k+1}^{m+k-1} a(m+k,i)\cdot(\ln 2)^i\cdot\left(1-\frac{i}{m+k}\right)^f .$$

A jobb oldalon álló összeghez az első tag hozzájárulása a legna-

gyobb.

Ha $i \leq (m+k)^{1/4}$, akkor $\forall m \geq M_1$ esetén (4.1.25) azt adja, hogy

$$\frac{1}{2i!} < a(m+k,i) < \frac{2}{i!}.$$

A második tagban $i > (m+k)^{1/4}$, így (4.1.24) alapján

$$a(m+k,i) < 1/(m+k)^2. \tag{4.1.29}$$

A harmadikban $k > k_0(m)$, ezért a (4.1.26) definíció szerint $\forall m \geq M_1$ esetén

$$(m+k)^{1/4} \leq k < i.$$

Tehát a (4.1.24) egyenlőtlenség ismét a (4.1.29)-hez vezet.

Könnyen látható, hogy $\exists$ olyan u_m sorozat, hogy $u_m = o(1)$, midőn $m \to \infty$, és $\forall m \geq M_1$-re

$$\sum_{k=1}^{k_0(m)} \sum_{i=k+1}^{[(m+k)^{1/4}]} \frac{1}{2i!} (\ln 2)^i \left(1 - \frac{i}{m+k}\right)^f \leq \sum_{k=1}^{\infty} \varepsilon_{m+k,k} \leq$$

$$\leq \sum_{k=1}^{k_0(m)} \sum_{i=k+1}^{[(m+k)^{1/4}]} \frac{2}{i!} (\ln 2)^i \left(1 - \frac{i}{m+k}\right)^f + u_m.$$

Ez garantálja, hogy elég nagy m esetén a (4.1.5) teljesül.

A (4.1.20), (4.1.22) és (4.1.28) biztosítja a 4.1.2. lemma alkal-

mazhatóságát, így a $\lim_{n\to\infty} Z^*(n)/y(n)$ létezését.

n	1	2	3	4
$Z(n)$	1	1	4	32
$Z(n)/f(n)$	1·386	1·128	1·145	1·131
n	5	6	7	8
$Z(n)$	436	9012	$2{\cdot}628\times10^{5}$	$1{\cdot}027\times10^{7}$
$Z(n)/f(n)$	1·124	1·120	1·117	1·115
n	9	10	11	12
$Z(n)$	$5{\cdot}183\times10^{8}$	$3{\cdot}280\times10^{10}$	$2{\cdot}543\times10^{12}$	$2{\cdot}371\times10^{14}$
$Z(n)/f(n)$	1·113	1·111	1·110	1·110
n	13	14	15	16
$Z(n)$	$2{\cdot}617\times10^{16}$	$3{\cdot}376\times10^{18}$	$5{\cdot}030\times10^{20}$	$8{\cdot}575\times10^{22}$
$Z(n)/f(n)$	1·109	1·108	1·107	1·107

4.1. táblázat

(a Z(n)/f(n) értékek 3 tizedesjegy pontossággal értendők)

5. A cluster analízis algoritmikus problémái

5.1. Az alkalmazott módszerek algoritmusairól

A gyakorlati alkalmazások szempontjából lényeges megvizsgálni a szóbajövő módszerek algoritmikus tulajdonságait. Az adott megfigyelés halmaz, a minta elemzése, kiértékelése alapján reméljük a jelenség megértését. Ahhoz, hogy az elemzést konkrét értelemben használhassuk, rögzítsük le azokat a szempontokat, melyek alapján a jelenséget elemezni kívánjuk. Az alkalmazott eljárások (pontosabban az ezeket megvalósító programok) futási idő és tárigényének előzetes ismerete meghatározhatja azoknak az eszközöknek és eljárásoknak az összességét, amelyben az adott méretű problémának, a rögzített szempontok alapján történő kiértékelése a választott módon megvalósítható.

Kiderülhet az is, hogy a probléma NP-nehéz, ilyenkor (legalább is pillanatnyi ismereteink szerint) a hatékony kiértékeléshez nem sok reményt fűzhetünk. A kiértékelés szempontjainak a megváltoztatásával "szoríthatjuk meg" az elemzést olyan polinomiális bonyolultságú problémára, amelynek megválaszolására lényegesen több remény van (bár a méret növekedése miatt a kiértékelési eljárás

megváltoztatása is szükségessé válhat).

Az algoritmusok bonyolultságával kapcsolatban Garey és Johnson könyvében [32] található terminológiát használjuk.

Konkrét algoritmusok tervezésénél (5.3.2. tétel és az egész 6. fejezet) egy egyszerűbb modellt használunk. Olyan RAM gépen képzeljük el az algoritmusok futtatatását, ami valós számokon végez aritmetikai műveleteket és az elérési és műveleti időt egyforma súllyal vesszük figyelembe. Ebben a modellben a feladat méretének nem a feladat lekódolásához szükséges bitek, hanem az input paraméterek számát vesszük. Az algoritmusok futási idejét csak nagyságrendileg vizsgáljuk, és optimális algoritmusok alatt is csak konstans szorzó erejéig optimális algoritmusokat értünk (azaz, amelyeknél bármely más, a feladatot megoldó algoritmus bármely input esetén legfeljebb konstansszor dolgozik gyorsabban [79]).

A gyakorlati problémák szempontjából különösen érdekesek azok az eljárások, amelyek adaptívak, vagyis amelyek a minta bővülésével nem igénylik az egész eljárás újraindítását, hanem az új mintaelemek hatékony feldolgozásával módosítják a korábbi eredményeket (konstrukciókat). Általában könnyebb nem adaptív algoritmust

találni. Elsősorban az utóbbi algoritmusokkal foglalkozunk, bár a 6. fejezetben - egy-egy vizsgált kérdés kapcsán - kitérünk adaptív (egyes szóhasználatban: on-line) algoritmusok bonyolultságára is.

A 5.2. pontban néhány kiértékelési problémáról bemutatjuk, hogy NP-nehéz.

A 5.3. pont új eredményeket tartalmaz: egy nem hierarchikus clusterezési problémáról bebizonyítjuk, hogy polinomiális bonyolultságú (5.3.2. tétel és 5.3.5. állítás). A bizonyításhoz felhasználjuk egy korábbi eredményünket (5.3.4. állítás).

Az 5.4. pontban bevezetünk egy olyan távolságosztályt (5.4.6. definíció), amelyre megszorítva egy természetes clusterezési probléma (a bevezetésben 3.-val jelölt "leghomogénabb" clusterezés keresése) inputjait, a megoldásra polinomkorlátos algoritmust adunk meg (5.4.7. tétel).

5.2. Algoritmikus bonyolultsági kérdések

A clusterezés általános optimalizálási problémája nagyon bonyolult, amint azt a 5.2.1. állítás mutatja.

Legyen adva egy egész számokból álló S alaphalmaz és egy f hibafüggvény, amely az alaphalmaz minden particionálásán értelmezve van. Az S halmaz, az f függvény, valamint a k természetes szám lesz a probléma inputja. Keressük S-nek azt a k-osztályú D particióját, amelyik minimalizálja az f-függvényt, azaz

$$\min_{D=(S_1,S_2,\ldots,S_k)} f(D).$$

5.2.1. állítás. Az előbb definiált Π_1 clusterezési probléma NP-nehéz, még abban az esetben is, ha az f egy polinom időben kiszámítható függvény és k=2.

Bizonyítás. A probléma inputjának a mérete az S halmaz elemeinek és az f függvény leírásához használt jelek száma.
Az állítás nyilvánvaló, hiszen a megfelelő Π_2 döntési probléma tartalmazza az NP-teljes partició problémát. Bővítsük ugyanis az eredeti clusterezési probléma inputját a t egésszel, azaz

5.2.2. definíció. Π_2 probléma: Létezik-e olyan D 2-partíciója S-nek, melyre $f(D) \leq t$?

Tekintsük a következő speciális esetet. Legyen az S egy egész számokból álló halmaz és jelölje (S_1,S_2) az S egy 2-particióját.

Rendelje ehhez az f veszteségfüggvény az azonos osztálybeli elemek összegéből képezett különbség abszolút értékét, azaz

$$f((S_1,S_2)) = \left| \sum_{i \in S_1} i - \sum_{j \in S_2} j \right|,$$

valamint legyen t=0.

Mivel Π_2 tartalmaz egy NP-teljes részproblémát, így $\Pi_1 \in$ NP-nehéz.

Ha $P \neq NP$, akkor ez azt jelenti, hogy a clusterezési problémát ebben az általános formában nem lehet polinom időben megoldani. Természetesen az f függvényre tett megszorítás, annak "gyors" kiszámolhatóságáról magától értetődő volt, hiszen különben nem is remélhetjük a probléma polinom idejű megoldását. Az f hibafüggvény speciális megválasztásával a clusterezési feladatok egy része polinomiális időben is megoldható. éppen ezért fontos kérdés a gyakorlati alkalmazások során, hogyan választjuk meg a hibafüggvényt.

Az alábbi problémáknál egy-egy gráfot feleltetünk meg a clusterezendő mintának. A gráf csúcsai reprezentálják a mintaelemeket és két csúcsot akkor és csak akkor kötünk össze éllel, ha a megfelelő mintaelemeken értelmezett távolság legfeljebb akkora, mint egy előre adott konstans. Ez utóbbi a gráf segítségével

történő clusterezés paramétere lesz.

Tekintsük a következő Π_3 gráf particionálási problémát. Adva van egy N csúcsú gráf. Keressük a csúcsok olyan 2-particionálását, amelyre minimális a halmazok között menő élek száma.

Ha a partíciók méretére nincs kikötés, akkor az N függvényében polinomiális időben megkonstruálható egy az előbbi értelemben optimális élszámmal rendelkező 2-partició.

Válasszunk ki ugyanis tetszőleges két csúcsot és tekintsük őket az egységnyi kapacitású élek által meghatározott hálózat forrásának illetve nyelőjének. A maximális folyam - minimális vágás tétel alapján a minimális vágást N-től függő polinomiális számú lépésben, a maximális folyam algoritmusa (pl. Lawler [60]); a csúcspárokat pedig $\binom{N}{2}$ féleképpen választhatjuk meg. (Megjegyezzük, hogy a Gomory-Hu [33] algoritmus segítségével elegendő N-1 maximális folyam feladatot megoldani az összes csúcspár közötti maximális folyamértékek meghatározásához.)

A probléma lényegesen bonyolultabb abban esetben, amikor az ún. kiegyensúlyozott 2-partíciók között keressük az optimálisat.

5.2.3. tétel. Ha N=2n és a Π_3 problémában csak azokat a 2-partíciókat engedjük meg, ahol mindkét halmaz n-elemű, akkor az

így kapott Π_4 probléma NP-nehéz.

Tekintsük a következő $\Pi_5(k)$ clusterezési problémát: lehet-e a $x \in V$ megfigyelési pontokat k osztályba particionálni olymódon, hogy bármely két $x,y \in V_i$ $(i=1,2,\ldots,k)$ megfigyelés a clusterezési probléma távolságfüggvényével mérve legfeljebb olyan messze van egymástól, mint egy előre adott konstans. Az osztály legtávolabb eső pontpárjának távolságát nevezzük a cluster átmérőjének. Az egyszerűség kedvéért feltesszük, hogy mind a konstans, mind a távolságok nemnegatív egészek.

Másképpen fogalmazva tehát a kérdést: lehet-e a mintaelemeket az előre adott konstanst nem meghaladó átmérőjű k clusterba particionálni.

Egy súlyozott élű teljes gráf szintgráfjain azokat a súlyozatlan élű részgráfokat értjük, amelyek úgy keletkeznek az eredeti gráfból, hogy csak azokat az éleket hagyjuk meg, melyek hossza egy adott számnál nem nagyobb, pl. a λ-szintű részgráf élei pontosan azok lesznek, melyekre $d(i,j) \leq \lambda$.

Ezek szerint a $\Pi_5(k)$ probléma ekvivalens a következővel: jelölje R az adott küszöb számot. A kérdés az, hogy az R-szinthez tartozó szintgráfban létezik-e a csúcsoknak egy olyan $(V_1,V_2,\ldots,V_k)$

particiója, amelyre minden V_i (i=1,2,...,k) a gráf teljes részgráfját feszíti.

5.2.4. Állítás. $\Pi_5(2) \in P$, míg $\Pi_5(k) \in$ NP-teljes, ha $k \geq 2$.

A megfelelő gráf klikk particionálási problémája mutatja, hogy miért igaz ez az állítás. Ez a $\Pi_5(k)$ problémának az a részesete, amikor a távolságértékeket a (0,1) halmazra szűkítjük le. A kérdezett gráf partició létezése ekvivalens a megfelelő gráf komplementer gráfjának a k-színezhetőségével. Világos, hogy még ilyen gráfokra megszorítva is $\forall k \geq 2 : \Pi_5(k)$ NP-teljes, mivel tartalmazza a 3-színezhetőség problémáját.

A $\Pi_5(k)$ problémával kapcsolatban felmerül az a kérdés, hogy hogyan lehetne olyan részproblémákat találni, ahol a színezési kérdés egyszerűbben válaszolható meg. Erre visszatérünk a 5.4. pontban.

A 2.2. pontban vizsgált nem hierarchikus clusterezési probléma (amikor az adott x_i (i=1,2,...,n) pontok legkisebb négyzetes eltérést adó k-osztályú clusterezését kerestük) következő természetes általánosítása NP-teljes.

5.2.5. tétel [32]. Legyen adva az $S=(x_1,x_2,\ldots,x_n)$ halmaz és egy $d(x_i,x_j)$ szimmetrikus, nemnegatív egész értékű távolságfüggvény, ami minden (x_i,x_j) pontpáron értelmezve van, valamint a k és B egész számok. Tekintsük a következő $\Pi_6(k)$ döntési problémát: létezik-e az x_i $(i=1,2,\ldots,n)$ pontoknak olyan $(S_1,S_2,\ldots,S_k)$ particiója, hogy

$$\sum_{r=1}^{k} \sum_{x_i,x_j \in S_r} d(x_i,x_j) \leq B. \qquad (5.2.1)$$

A $\Pi_6(k)$ probléma már a k=2 esetben is NP-teljes, következésképpen a (5.2.1) bal oldalán álló kifejezés minimalizálási problémája NP-nehéz.

Legyen a $w(S_i)^{-1}$ egész értékű súlyfüggvény értelmezve mindegyik S_i $(i=1,2,\ldots,k)$ halmazon. Ha a $\Pi_6(k)$ problémát kiegészítjük a $w(S_i)$ inputtal és (5.2.1) helyett a

$$\sum_{r=1}^{k} w(S_r) \cdot \sum_{x_i,x_j \in S_r} d(x_i,x_j) \leq B \qquad (5.2.2)$$

feltételre nézve megengedett particiókra vonatkozó minimalizálási problémát tekintjük, akkor nyilván az is NP-nehéz.

A legkisebb négyzetes kritérium alapján történő clusterezés során

az S az R^m - az egyszerűség kedvéért egész koordinátájú pontokból álló - n-elemű részhalmaza és egy speciális d távolságfüggvényt használunk, ti. a d-nek az euklideszi távolság négyzetét vesszük. Könnyen látható, hogy ekkor a (5.2.2) bal oldala a

(5.2.3) $$2 \sum_{r=1}^{k} w(S_r) \cdot |S_r| \sum_{x_i \in S_r} d(x_i, q_r)$$

kifejezéssel egyenlő, ahol q_r jelöli az S_r halmazbeli elemek átlagát. A $w(S_r)^{-1} = 2|S_r|$ választással a (2.2.3) átfogalmazását kapjuk. Láttuk, hogy a k-közép eljárás (2.2. pont) véges sok lépés után stacionárius ponthoz érkezik. Tehát minden olyan k esetén, amikor a fenti célfüggvénynek egyetlen lokális optimumhelye van, akkor a $\Pi_6(k)$ fenti speciális részproblémája P-beli.

A lokális optimumhelyek száma valójában nem érdekes. A 5.3. pontban megmutatjuk (5.3.5. állítás), hogy az előbb említetteknek megfelelő $\Pi_7(k)$ probléma P-beli.

5.3. Egy polinomiális bonyolultságú nem hierarchikus clusterezés

Ebben a részben egy olyan problémáról látjuk be, hogy P-beli, amelyben azzal egyszerűsítjük az optimális clusterezés

keresésésnek feladatát, hogy a particiók geometriai struktúrájára kötünk ki feltételt. Az itt közölt eredmények közül az 5.3.2. tétel, az ezt előkészítő 5.3.4. állítás új eredmény. Az 5.3.3. állítás egyszerűbb részére új bizonyítást adunk.

5.3.1. definíció. Az $S=(x_1,x_2,\ldots,x_n)\subset R^m$ pontok konvex k-particiójén olyan k-particionálást értünk, ahol az osztályok konvex burka diszjunkt.

Legyen az f egy szigorúan monoton növő, folytonos függvény. Az $S=(x_1,\ldots,x_r)\subset R^m$ halmaz **f-centrumának** azt a $q\in R^m$ pontot nevezzük, amelyre az **S-nek a q-ra vonatkozó f-nyomatéka,** azaz az

$$M(f,S,q) = \sum_{i=1}^{r} f(\|x_i-q\|)$$

összeg minimális. Ezt a q pontot $c_f(S)$-sel, míg a hozzá tartozó összeget - az **S halmaz f-centrális nyomatékát** - $W_f(S)$-vel jelöljük.

Megjegyezzük, hogy $f(x)=x^2$ esetén az f-centrum a súlyponttal azonos.

Legyen $S\subset R^m$ egy véges halmaz és $S=S_1\cup\ldots\cup S_t$ az S egy particiója. E partició **f-nyomatékán** a partició osztályain vett

f-centrális nyomatékok összegét értjük. Ez tehát a következő mennyiség

(5.3.1)

$$W(f,S,(S_i)_{i=1}^{k}) = \sum_{i=1}^{k} W_f(S_i) = \min_{(q_i)} \sum_{i=1}^{k} \sum_{x_j \in S_i} f(\|x_j - q_i\|).$$

Ezt az értéket kívánjuk minimalizálni S összes k-particiója között. Keressük tehát azt az $S=S_1 \cup \ldots \cup S_k$ particiót, amelyre $W(f,S,(S_i)_{i=1}^{k})$ minimális, azaz

(5.3.2)

$$W(f,S;k) = \min_{\substack{(S_i) \\ S \text{ } k\text{-particiója}}} W(f,S,(S_i)_{i=1}^{k}).$$

Az ilyen k-particiókat **f-optimálisnak** nevezzük.

Az m-dimenziós tér n pontját általános helyzetűnek nevezzük, ha bármelyik m, vagy kevesebb elemű részhalmaza lineárisan független rendszert alkot.

5.3.2. tétel. Ha f szigorúan monoton növő, folytonos függvény és $S \subset R^m$ általános helyzetű pontok véges halmaza, akkor rögzített k és m esetén S egy f-optimális k-particiója polinom időben megtalálható. Az így adódó f-optimális k-partició konvex lesz.

Bizonyítás. Az S_i osztályokat clustereknek nevezzük. Jelöljünk

ki minden osztály számára egy q_i reprezentáns pontot (f-centrum-jelöltet). Mivel az f szigorúan monoton növő függvény, ezért a x_j pontot a hozzá az euklideszi norma szerint legközelebb fekvő q_i reprezentáns pont osztályába kell sorolni. Abban az esetben, ha találunk olyan pontot, amely nem így lenne besorolva, akkor a legközelebbi q_i reprezentáns pont osztályába soroljuk át az összes érintett pontot. Az így keletkező új O' k-partició - megfelelő minimalizálás után adódó - f-centrális nyomatékára nyilván

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} f(\|x_j - q_i\|) > W(f,S,O') >$$

adódik. Ha ez az érték tovább már nem csökkenthető, akkor az összes x_j pontot a hozzá az euklideszi norma szerint (esetleg egyformán) legközelebb fekvő q_i reprezentáns pontok közül a legkisebb indexű pont osztályába soroljuk át.

Ezek szerint a legkisebb távolság partíciók (6.4. pont) között, így az LTP-k halmazánál bővebb konvex k-particionálások halmazában is található f-optimális megoldás. A tétel bizonyításhoz felhasználjuk, hogy a konvex k-partíciók száma lényegesen kisebb, mint az összes k-particionálás száma (5.3.3. és 5.3.4. állítás),

ti. n-től polinomiálisan függ, midőn a k és m rögzített. A következő állítás lényegében ekvivalens alakban megtalálható Covernál [18].

5.3.3. állítás. A síkon n pontnak legfeljebb $\binom{n}{2}$ olyan 2-particionálása van, ahol a két osztály pontjainak konvex burka diszjunkt. Tetszőleges rögzített m>2 mellett az m-dimenziós euklideszi térben n pontnak legfeljebb

$$2^{n^{\binom{m+1}{2}}}$$

konvex 2-particionálása van, midőn $n \to \infty$.

Általános helyzetű pontoknak

$$\binom{n}{m}\cdot(1+O(1/n))$$

ilyen 2-particionálása létezik, midőn $n \to \infty$.

Ez az eredmény egyszerűen általánosítható

5.3.4. állítás ([63]). Az m-dimenziós euklideszi térben n általános helyzetű pontnak legfeljebb $\left(2^{n^{\binom{k}{2}}}\right)^m$ konvex k-particiója van.

Először a 5.3.3. állítást bizonyítjuk.

Az 5.3.3. állítás bizonyítása. Megjegyezzük, hogy Winder [111], Cover [18] az m-dimenziós euklideszi tér n-elemű ponthalmazainak különböző felületekkel való két részre szeparálásainak számával kapcsolatban adott meg eredményeket. Két ponthalmaz konvex burka nyilván pontosan akkor diszjunkt, ha a halmazok lineárisan szeparálhatók. Az állítás első részét (m=2) a [18]-ban közölt módszertől eltérő módon igazoljuk. Bár további céljainkhoz elegendő lenne a [61]-ban megadott eredményünk is, magasabb dimenzióban egy másik bizonyítást ismertetünk.

Csak az $n\geq 2$ eset érdekes. Ha semelyik 3 pont sem fekszik egy egyenesen, akkor a síkon bármelyik két pont egyenese a konvex 2-particiók közül pontosan kettőt határoz meg, ti. a két osztály közös támaszegyeneseként. Ugyanakkor pontosan 2 ilyen támaszegyenes található minden 2-particióhoz, tehát a konvex 2-particiók száma $\binom{n}{2}$. Ha a pontok közül $n_i \geq 2$ kollineáris, akkor az ezen az egyenesen levő pontok csak $2(n_i-1)$ konvex 2-particiót határoznak meg a nem kollineáris eset $\binom{n_i}{2}$ partíciója helyett.

Magasabb dimenzióban, pl. m-dimenzióban n tetszőleges elhelyezkedésű pont konvex 2-particiói számára úgy adunk felső kor-

látot, hogy a konvex 2-particiókat beleképezzük az ún. irányí tott zászlók halmazába. (Amennyiben az n pont nem feszíti ki az m-dimenziós teret, akkor jelölje m a kifeszített tér dimenziójá t. Amint az egyszerűen látható, ez az átjelölés nem vezet bonyodalmakhoz.) Ezek számára adunk meg felső korlátot.

Irányított zászló alatt irányított affin alterek maximális láncát értjük.

Azt mondjuk, hogy egy hipersík 2 halmazt "elválaszt", ha a hipersík által határolt egyik zárt féltér tartalmazza az egyik halmazt, a másik zárt féltér a másikat. Tehát a két halmaz szeparálásához képest az a különbség, hogy ott a nyílt félterektől várjuk el, hogy tartalmazzák az osztályokat (vagyis úgyis mondhatjuk, hogy egy "elválasztó" hipersík a hipersíkra eső pontok kivételével szeparálja a 2 osztályt.)

Minden konvex 2-particióhoz megadható olyan "elválasztó" hipersík is, amit az n-elemű halmaz alkalmas m pontja határoz meg. A határlapra eső pontok száma akár n is lehet (általános helyzetű pontok esetén ez a szám pontosan m). A hipersíkon kívüli pontokról könnyen megállapítható, hogy melyik osztályhoz tartoznak, míg a határlapra eső pontokra újra meg kell nézni, hogy

hányféleképpen konvex 2-particionálhatók.

Az egyik osztály pontjait határozzuk meg, éspedig úgy, hogy kijelöljük az "elválasztó" hipersík által határolt megfelelő félteret, majd megkeressük a belső pontokat és ezt kiegészítjük a határlap valamelyik konvex 2-particiójának osztályával.

Általános helyzetű pontokra a határlapra eső szimplexnek 2^m-2 ilyen 2-particiója van, tehát a konvex 2-particionálások száma legfeljebb $\binom{n}{m}2^m$, ami minden m-re legfeljebb $2n^m$.

Tetszőleges elhelyezkedésű pontoknál az "elválasztó" hipersíkra eső pontok által definiált tér egy "elválasztó" (most már eggyel alacsonyabb dimenziós) hipersíkját (alterét) választjuk ki, majd a belső pontok megkeresése következik, végül az altér valamely konvex 2-particiójának egyik osztályával bővítjük az osztályt, stb. Ez az iránnyal bővített altér lánc alkotja az irányított zászlót, és így legfeljebb

$$\binom{n}{m}\binom{n}{m-1}\dots\binom{n}{1}\cdot 2^m \leq 2n^{\binom{m+1}{2}}$$

konvex 2-partició adható meg. Nem tudjuk, hogy az n kitevőjében szereplő érték mennyire éles.

Megjegyezzük, hogy általános helyzetű pontokra

$$\sum_{k=0}^{m} \binom{n-1}{k}$$

konvex 2-particionálás létezik [18]. Rögzített m és $n\to\infty$ esetén ez $\binom{n}{m}\cdot(1+O(1/n))$.

Az 5.3.4. állítás bizonyítása. Minden konvex k-particionáláshoz hozzárendeljük irányított zászlók $\binom{k}{2}$-hosszú sorozatát. Sorszámozzuk a k-partició osztályait. Az első zászló mutatja meg, hogy egy konkrét k-partició 1. és 2. osztályának unióját hogyan 2-particionáltuk. Ezek között lesz a két osztály konkrét 2-particiója is. A következő zászló az 1. és 3. osztály, az utolsó pedig a k-1. és k. osztály unióját reprezentálja a megfelelő 2-particiót. A bizonyítással készen vagyunk, hiszen az ilyen sorozatok száma legfeljebb

$$\left(2n^{\binom{m+1}{2}}\right)^{\binom{k}{2}}.$$

Ha az n pont általános helyzetű, akkor ez a mennyiség legfeljebb

$$2n^{m\binom{k}{2}}.$$

Végül a 5.3.2. tétel bizonyításához meg kell mondanunk, hogy hogyan lehet ezt az eredményt polinomiális idejű algoritmus készítésére felhasználni.

Egy pontokkal adott sík normálvektorának a kiszámítása illetve egy pont k hipersíkhoz viszonyított helyzetének a megállapítása k és m rögzítése esetén konstans lépést igényel.

Mivel az 5.3.4. állítás bizonyítása során, a lehetséges konvex k-partíciók generálásához (sőt annak a verifikációjához is, hogy egy generált k-partició konvex) csak az előbb említett lépéseket használtuk, így n függvényében polinomiális sok lépés elegendő a konvex k-partíciók generálásához.

Megjegyezzük, hogy a tétel akkor is igaz marad, ha az S halmaz pontjainak általános helyzetűségéről tett feltételt elhagyjuk, bár ekkor a javasolt módszer még lassúbb lesz.

Végül tekintsük a következő $\Pi_7(k)$ problémát: legyen a k egy rögzített pozitív egész szám.

Instancia: az R^m egy véges S részhalmaza, melynek minden x_i pontja egész koordinátájú; valamint egy B pozitív egész.

Kérdés: létezik-e az S-nek olyan $(S_1,S_2,\ldots,S_k)$ particiója, hogy

$$\sum_{r=1}^{k} \frac{1}{|S_r|} \sum_{x_i,x_j \in S_r} \|x_i - x_j\|^2 < B.$$

Az 5.2.5. tételhez kapcsolódó (5.2.2) és (5.2.3) formulák és az 5.3.2. tétel alapján most már kimondható a

<u>5.3.5. állítás</u>. $\Pi_7(k) \in P$.

5.4. A clusterező eljárások megengedettségi osztályozása és vizsgálatai

A clusterező eljárások egyik osztályozási módja az ún. megengedettségi vizsgálat. Ilyen vizsgálatok bevezetése és az algoritmusok egy részének ebből a szempontból való kiértékelése Fisher és Van Ness [30] nevéhez fűződik. Fő vonalaiban ismertetjük az osztályozás módszerét és néhány közvetlenül kapcsolódó eredményt. Az 5.5. pontban kitérünk a clusterezés néhány problémájának a dinamikus programozással való elvi kapcsolatára.

Az 5.4.7. tétel a $\Pi_5(k)$ döntési probléma (5.2. pont) egy az un. fa-szerű távolságok osztályára megszorított - részproblémájának polinomkorlátos megválaszolhatóságát mondja ki.

Az 5.4.9. állítás egy clusterezési problémának az input hosszában polinomiális idejű megoldhatóságát fogalmazza meg.
Az ultrametrikák esetében az 5.4.10. állítás bizonyítása során mondottak segítségével jól jellemezhetők azok az inputok, amelyekre az előző döntési problémában igen válasz adódik.
Az 5.4.3., 5.4.4. és az 5.4.10. állítás az ultrametrikák egy-egy érdekes tulajdonságára világít rá.

A megengedettségi feltételek két típusát különböztetjük meg. Az egyik közvetlenül a módszer tulajdonságaihoz kapcsolódik. A másik típusban előbb a clusterek bizonyos tulajdonságait definiáljuk. Ezek után akkor mondjuk azt, hogy egy módszer a tulajdonságra megengedett, amennyiben az outputként adódó clusterek az adott tulajdonsággal rendelkeznek.

Alább kiválasztottunk néhány tulajdonságot és módszert, és megadjuk ezek besorolását.
Az 5.4.11., 5.4.12. illetve 5.4.13. definícióban az első típusra adunk példákat, míg az 5.4.1., 5.4.2. definíció a második alternatívára vonatkozik.

5.4.1. definíció. Legyen az A a clusterezés osztályaira vonat-

kozó valamilyen ésszerű tulajdonság. Egy clusterező eljárásról azt mondjuk, hogy **A-megengedett**, ha tetszőleges input halmazra a módszer által adott clusterezés rendelkezik az A tulajdonsággal.

Ha az optimális clusterezés nem rendelkezik az A tulajdonsággal, akkor természetesen az optimumot sikerrel kereső eljárás sem lehet A-megengedett.

5.4.2. definíció. Egy k-osztályú clusterezést (k-csoport) **jól struktúrált**nak nevezünk egy adott távolságfüggvényre nézve, ha mind a k osztályban az azonos osztálybeli pontok közötti távolságok legfeljebb akkorák, mint a különböző clusterekből vett pontok közötti minimális távolság. Egy k-osztályú clusterezést **kompakt szeparált**nak nevezünk egy adott távolságfüggvényre nézve, ha jól struktúrált k-clusterezés és az osztályokon belüli távolságok határozottan kisebbek az osztályok közöttieknél.

Nyilvánvaló a

5.4.3. állítás. Egy n-elemű halmaznak bármely ultrametrikára nézve $\forall k=1,2,\ldots,n-1$ értékre van k-osztályu jól struktúrált clusterezése.

Bizonyítás. Ha ultrametrikából indulunk ki, akkor a megfelelő

MFF-ról - a 3.1. pont példája után említettek miatt - leolvasható az összes távolság. A Kruskal-algoritmus biztosítja, hogy minden komponens "átmérője" (azaz legtávolabbi pontjainak távolsága) az utoljára bevont élének hosszával legyen egyenlő. Ezért a komponensek belső élei minden lépésben legfeljebb olyan hosszúak lesznek, mint a komponensek közötti külső élek.

A fenti állítás vezet át a $\Pi_5(k)$ problémához

5.4.4. állítás. Annak szükséges és elegendő feltétele, hogy egy súlyozott élű teljes gráf összes szintgráfja csúcs- és él-diszjunkt klikkekre legyen partcionálható az, hogy a súlyok ultrametrikus távolságot alkossanak. Tehát $\Pi_5 \in P$, ha a szóbajövő d távolságfüggények osztályát az ultrametrikus távolságokra szorítjuk meg.

Bizonyítás. Ultrametrikus távolság esetén nyilvánvaló, hogy a szintgráf mindegyik komponense teljes részgráfot feszít a szintgráfban. Fordított irányban tegyük fel, hogy az eredeti gráfban létezik olyan háromszög, amelyben a leghosszabb élek nem egyenlők. Ha a szintszámot a háromszög második legnagyobb élhosszával egyenlőnek választjuk meg, akkor a szintgráfban a háromszög csúcsai egy nem teljes komponensbe kerülnének, ami ellentmondana

a feltevésnek.

A fentiek szerint ultrametrika esetében a szintgráf komponensekre bontásával egyszerűen megválaszolható Π_5. Az előbb említett szintgráfok perfektek.

Ennek a fejezetnek egyik fontos eredménye az, hogy a metrikák egy lényegesen bővebb osztályára terjesztjük ki az előző észrevételt. Ehhez néhány definícióra és kiegészítésre van szükségünk.

A $\Pi_5(k)$ problémával kapcsolatban felvethető a következő minimalizálási probléma is: adott k mellett adjuk meg azt a minimális átmérőt, amelyhez létezik a csúcsoknak olyan k-particiója, hogy egyik osztály átmérője sem haladja meg ezt a számot.
A minimalizálási feladatot átfogalmazhatjuk a szintgráfok (5.2. pont) segítségével is. Keressük azt a minimális λ számot, amelyre a λ-szintű részgráf k (csúcs-diszjunkt) klikkre bomlik. Jelölje I a probléma egy instanciáját, és OPT(I) a minimalizálási feladat megoldásaként adódó minimális átmérőt.

1984-ben Hochbaum és Shmoys [44] olyan polinomiális futási idejű közelítő eljárást adtak meg, amivel az OPT(I) minimális átmérő legfeljebb kétszerese (és egy ekkora legnagyobb átmérővel rendelkező k-partició) meghatározható. A módszer feltételezi,

hogy a távolságok kielégítik a háromszögegyenlőtlenséget. Ez a "közelítő" eljárás ultrametrikus távolság esetén alkalmas egy optimális struktúra megtalálására is, bár erre a cikkben nincs utalás. Fenti szerzők bizonyították, hogy $\alpha OPT(I)+\beta$ (ahol $\alpha(2$ és $\beta)0$) átmérőt garantáló módszer létezése azt vonná maga után, hogy P=NP. Mi a pontos megoldást keressük.

5.4.5. definíció. Távolság fán és az általa indukált távolságon egy súlyozott élű fát értünk, amiben bármely két csúcs távolságát a fában egyértelműen meghatározott összekötő útvonalukon érintett élek összhossza adja meg.

Most rátérünk annak a távolságosztálynak az ismertetésére, amire vonatkozóan a $\Pi_5(k)$ problémát polinom időben tudjuk megoldani.

5.4.6. definíció. Egy metrikát fa-szerűnek mondunk, ha megadható hozzá az eredeti ponthalmazt tartalmazó, az eredeti pontokon az eredeti távolságokat indukáló távolság fa.

Könnyen látható, hogy a távolság fa által indukált távolságok kielégítik az ún. négypont-feltételt. Ez a feltétel a csúcshalmazból kiválasztott minden pontnégyesre az általuk meghatározott négyszög kitérő élpárjainak hosszösszegére ró ki ultrametrikus

egyenlőtlenséget ([13]), ami az eredeti távolságokra a három-szögegyenlőtlenségnél erősebb, az ultrametrikus feltételnél gyengébb kikötést jelent.

Buneman [13] eredménye mutatja, hogy éppen a négypont-feltételt kielégítő távolságok fa-szerűek. Az 5.4.7. tétel fogalmazza meg a fejezet egyik fontos eredményét

5.4.7. tétel. Fa-szerű metrikára $\Pi_5 \in P$ (azaz polinom időben megválaszolható a $\Pi_5(k)$ olyan formában is, hogy a k-t előre nem rögzítjük, hanem az input részeként tekintjük).

A problémát perfekt gráfok kiszínezésére vezetjük vissza. Ez a színezési probléma perfekt gráfokra Grötschel, Lovász, Schrijver [37] egy az ellipszoid módszert felhasználó algoritmusa révén polinom időben megoldható. Mi elkerüljük az ellipszoid módszer használatát.

Az 5.4.7. tétel bizonyítása. Elegendő a szintgráfok komplementer gráfjának kiszínezésével foglalkozni. Fa-szerű metrikára a teljes gráf mindegyik szintgráfja perfekt gráf. Könnyen belátható ugyanis, hogy távolság fa által indukált távolságokra a teljes gráf szintgráfjaiban bármelyik legalább 4 hosszú kör tartalmaz átlót. Az ilyen gráfok perfektek ([78]). (Az "átló"-tulajdonság

jó karakterizációját adja az ún. részfa gráfoknak, [78] 9.29. feladat. Részfa gráfon egy irányítatlan fában a részfák metszetgráfját értjük.)

Speciálisan a távolság fából keletkező szintgráfok komplementerei (és a szintgráfok is) könnyen színezhetők. Rögzítsük ugyanis a távolság fa valamelyik csúcsát gyökérpontnak. Tekintsünk egy tetszőleges szintgráfot. Válasszunk ki egy olyan levelet a fában, ami a legmesszebb van a gyökértől. A négypont-feltétel biztosítja, hogy a szintgráfban az ezzel a ponttal éllel összekötött csúcsok egy teljes részgráfot feszítnek a pont komponensében. Válasszuk le ezt a teljest a szintgráfból. Ezek után a komplementer gráf kiszínezése a csúcsok számára vonatkozó indukcióval történhet. (Megjegyezzük, hogy a gráf egy csúcsának - az indukált távolság szerint vett - adott sugarú környezetét alkotó pontok a távolság fa valamelyik részfájának lesznek a csúcsai.)

Speciálisan egydimenzióban, az euklideszi távolságra tekintsük a következő intervallum gráfot: minden pontra egy R hosszúságú, a pontra nézve szimmetrikus elhelyezkedésű intervallumot illesztünk. Két intervallum pontosan akkor van éllel összekötve a gráfban, ha tartalmaznak közös pontot, azaz ha a meghatározó pontok legfeljebb R távolságra vannak egymástól. Ezek szerint

speciális esetként kimondható a

5.4.8. állítás. Egydimenzióban az euklideszi metrikára nézve a szintgráfok intervallum gráfokat alkotnak.

Természetesen minden potenciál jellegű távolságfüggvény esetén is használható az előbbi gondolat.

Végül megjegyezzük, hogy az ultrametrikus távolságok is fa-szerűek.

Most elevenítsük fel a (bevezetésben 1.-vel jelölt "szeparálási") kérdést, azaz egy adott input ponthalmazra létezik-e a pontok eredeti távolságára nézve kompakt szeparált k-clusterezés, és ha a válasz igen, akkor hogyan lehet ilyet megadni.

Tekintsük a következő $\Pi_8(k)$ problémát.

Instancia: az S alaphalmaz és a $d(x_i,x_j)$, x_i, $x_j \in S$ nemnegatív, szimmetrikus távolságságfüggvény ($d(x_i,x_i)=0$), valamint egy pozitív k szám; kérdés: létezik-e az S-nek olyan k-osztályú particionálása, amelyben az azonos osztálybeli pontok közötti távolságok maximuma kisebb a különböző osztályokból vett pontok minimális távolságánál. A kérdés ekvivalens alakja szintgráfokkal: létezik-e olyan szintgráfja az eredeti gráfnak, amelyik k

nem üres, csúcs- és él-diszjunkt klikkre esik szét.

Nyilván $\Pi_8 \in NP$.

Felvetődik a kérdés, hogy egyáltalában milyen távolságfüggvények esetén kaphatunk igen választ a $\Pi_8(k)$ kérdésre. Könnyen látható (Dunn [23]), hogy minden alaphalmazhoz és az elempárokon értelmezett tetszőleges távolságmátrixhoz legfeljebb egy kompakt szeparált k-clusterezés van. Az euklideszi távolság esetére szintén Dunn bizonyított elégséges feltételt arra vonatkozóan, hogy a k-clusterezés kompakt szeparáltsága implikálja azt is, hogy egyben stacionárius pont legyen a legkisebb négyzetes kritériumra nézve is. Bizonyos esetekben ez egy lehetőséget ad arra, hogy a ponthalmaz kompakt szeparált k-clusterezési lehetőségét észleljük. Azonban, sajnos, vannak olyan pont konfigurációk is, amikor - bár létezik kompakt szeparált k-clusterezés - az egyetlen stacionárius pontot adó k-clusterezés nem ilyen.

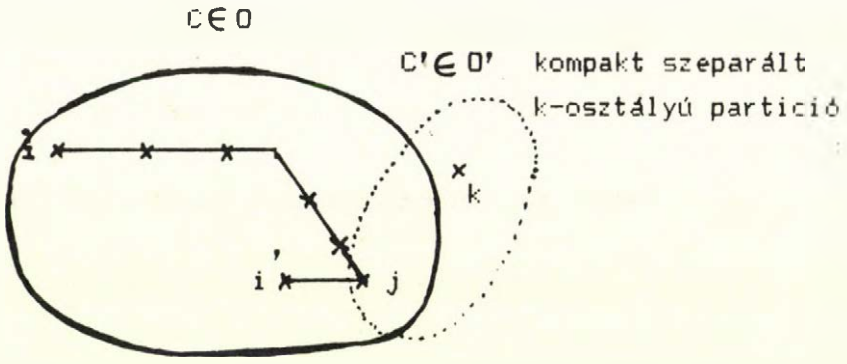
Fisher és Van Ness egydimenziós pontok olyan konfigurációját adták meg, hogy az (5.2.3) veszteségfüggvényre optimális clusterezés semmilyen nemnegatív w ($w \neq 0$) súlyfüggvény mellett sem jól struktúrált, tehát kompakt szeparált sem.

Az adott távolságokat szortolva nem nehéz ellenőrízni, hogy valamilyen alkalmas K-ra a legrövidebb K él által meghatározott

részgráf k (esetleg egyelemű) klikk-komponensből áll-e vagy sem. Ezek szerint a Π_8 kérdést polinomiális időben megválaszolhatjuk. A következő bizonyításban a Kruskal-algoritmus kompakt szeparált megengedettségét bizonyítjuk. Ez egyben az előbbinél egyszerűbb eljárást ad meg a Π_8 megválaszolására.

5.4.9. állítás. A single és complete linkage (6.6. pont) eljárások kompakt szeparált megengedettek.

Bizonyítás. A kompakt szeparált megengedett tulajdonságot indirekt módon bizonyíthatjuk a single linkage esetében: tegyük fel, hogy az eljárás által az n-k. lépésben adott O k-particionálás nem kompakt szeparált. Ekkor található 3 olyan pont, hogy $i,j \in C$, $k \notin C$, $j,k \in C'$ és $i \notin C'$, ahol C' illetve C az egyetlen kompakt szeparált O' illetve az O k-particionálás egy-egy osztályát jelöli. Jelölje i' az előző tulajdonságú i pontok közül azt, amelyik a j-hez legközelebb van és az O-nak megfelelő erdőben a j-vel él köti össze. A single linkage lépései miatt $d(i',j) \leq d(j,k)$, az O' kompakt szeparáltsága miatt viszont $d(i',j) > d(j,k)$, ami ellentmondás.



5.1. ábra

A kompakt szeparált k-clusterezés egyértelműsége és a single linkage eljárás kompakt szeparált megengedett tulajdonsága miatt amíg a k=n,n-1,n-2,... értékekre létezik kompakt szeparált k-clusterezés, addig a megfelelő clusterek finomodó halmazsorozatot alkotnak. Az is nyilvánvaló, hogy ha egy eljárás kompakt szeparált megengedett, akkor a megfelelő lépésekben ugyanazokhoz a clusterezésekhez vezet, mint a single linkage módszer. Az utóbbi gondolat segítségével bizonyítható pl. a complete linkage kompakt szeparált megengedettsége is. (Hiszen, ha nem ugyanabban a cluster párban lenne minimális a legközelebbi pontpárok távolsága, mint amiben minimális a legtávolabbiaké, akkor a single linkage nem lehetne kompakt szeparált megengedett.)

A fentiek egyben azt is jelentik, hogy minden n-elemű S halmaz

esetében pl. a single linkage eljárás segítségével nagyságrendileg $n^2 \log n$ lépésben az összes k értékre egyszerre megválaszolható a $\Pi_8(k)$ kérdés, ti. az eljárás által adott k-osztályú clusterezések tesztelésével. (Megjegyezzük, hogy a (6.6.1) egyenlőség a (6.6.2)-(6.6.4) clusterek közötti távolságfüggvények esetén nemcsak akkor áll fenn, amikor a legkisebb d_{ij} távolságot adó C_i és C_j clustereket egyesítjük a C_k clusterben. Ekkor minden lépésben a (6.6.1) képlet segítségével a keletkező cluster "átmérője" egyszerűen számolható.)

Az ultrametrikákra vonatkozóan jól jellemezhető a kompakt szeparáltság:

5.4.10. állítás. Egy ultrametrikára nézve akkor és csak akkor van $\forall k=1,2,\ldots,n-1$ értékre k-osztályú kompakt szeparált clusterezése egy n-elemű halmaznak, ha az ultrametrikának pontosan n különböző értéke van (a nullát is beleértve).

Bizonyítás. Az 5.4.3. állítás bizonyításához hasonlóan biztosítja a k-osztályú kompakt szeparáltságot a Kruskal-algoritmus abban az esetben, amikor n különböző értéke van az ultrametrikának. Másrészt, ha az ultrametrika n-nél kevesebb értéket vesz fel, akkor a MFF-jában legalább két él egyforma hosszú. Mivel az

5.4.9. állítás szerint a single linkage eljárás kompakt szeparált megengedett, így azokra a k értékekre nincs k-osztályú kompakt szeparált clusterezés, melyeknek megfelelő lépésekben egyforma hosszúságú élek közül választ a Kruskal-algoritmus.

Ha a clusterek geometriai struktúrájának feltárása fontosabb, mint a pontok eloszlásáé, akkor a pontok illetve a clusterek ismétlésére invariáns clusterező eljárások választása indokolt.

5.4.11. definíció. Egy clusterező eljárást pont ismétlés megengedettnek nevezünk, ha a pontok tetszőleges számú ismétlése után minden lépésben ugyanazokhoz a clusterhalmazokhoz vezet az eljárás, mint ismétlés nélkül.

5.4.12. definíció. Egy clusterező eljárást cluster ismétlés megengedettnek nevezünk a k. szintnél, ha tetszőleges input halmaz esetén az eljárás által adott $C_1, C_2, \ldots, C_k$ clusterek bármelyikének (ti. összes pontjának) tetszőleges számú ismétlése után az eljárás ugyanazokhoz a clusterhalmazokhoz vezet a k. lépésben.

Nyilvánvaló, hogy az ismétlés megengedett eljárások érzéketlenek a ponthalmaz eloszlásának bizonyos változtatásaira.

Statisztikai vizsgálatok végrehajtása szempontjából érdekesek azok a módszerek, ahol a clusterek elhagyása után nem változnak a megfelelő szinten keletkező clusterek.

5.4.13. definíció. Tetszőleges n-elemű alaphalmazra alkalmazva, a k. ($1 \leq k \leq n-1$) szinten keletkezett clusterek közül hagyjuk el valamelyik összes pontját és a maradék input halmazra futtassuk újra az eljárást. Ha a k-1. lépésben - az elhagyott kivételével - ugyanazokat a clustereket adja az eljárás, akkor cluster elhagyás megengedettnek nevezzük.

A single és a complete linkage eljárások az előbbi 3 követelményre nézve megengedettek, míg a legkisebb négyzetes kritériumra optimális clusterezést adó módszerek csak cluster elhagyás megengedettek lehetnek.

5.5. A megengedettségi vizsgálatokkal kapcsolatos egyéb megjegyzések

Ha egy clusterező eljárás valamilyen célfüggvény globális optimumát találja meg és cluster elhagyás megengedett, akkor ezt az optimumot elvileg a dinamikus programozás módszereivel is megkereshetjük.

Tekintsük az n-elemű S alaphalmaz pontjainak k-clusterezésein értelmezett $L_k(S)$ veszteségfüggvényt. Ha egy clusterező eljárás a függvény minimumát minden k érték esetén megadja és cluster elhagyás megengedett, akkor a

$$(5.5.1) \qquad L_k(S) = \min_{C_k \subset S} (L_1(C_k) + L_{k-1}(S \setminus C_k))$$

egyenlet fennáll. Általában a veszteség az alábbi módon van definiálva

$$(5.5.2) \qquad L_k(S) = \min_{\substack{(C_1,C_2,\ldots,C_k) \\ \text{az } S \text{ } k\text{-particiója}}} \sum_{i=1}^{k} f_i(C_i),$$

ahol az f_i -k tetszőleges halmazfüggvények. Speciálisan a legkisebb négyzetes kritérium esetében az $f_i(C)$ a $C \subset R^m$ ponthalmaz pontjaira a C-beli átlagtól való eltérések négyzetösszege.

Ha $f_i(C)$ függvénynek a C halmaz minimális feszítőfájában szereplő élek összsúlyát választjuk, akkor a Kruskal-algoritmus 6.4.14. tételben említendő tulajdonsága (a cluster elhagyás megengedettség révén) közvetve biztosítja az (5.5.1) összefüggés fennállását.

Az (5.5.1) felhasználásával a legkisebb négyzetes kritérium esetében egydimenzióban $O(kn^2)$ lépésszámú algoritmus adható meg (Fisher [29], Jensen [49]), az utóbbi esetben viszont a dinamikus programozásnál nyilván hatékonyabb a single linkage módszer.

5.5.1. definíció. Egy módszert hierarchikus megengedettnek nevezünk, ha tetszőleges alaphalmaz esetén az alaphalmaz egyre finomodó illetve egyre durvuló sorozatát adja a clusterezés outputjaként.

Az agglomeratív hierarchikus módszerek és az alaphalmaz egyre finomódó partició láncai közötti egy-egyértelmű megfeleltetést a 3.1. pontban ismertettük. Egy fontos negatív eredményt bizonyított Fisher és Van Ness [30]. Tekintsük az n-elemű halmaz clusterezéseinek (particionálásainak) következő sorozatát. Az első lépésben mind az n pont egy-egy különálló clustert definiál. Minden további lépésben a clusterező eljárás a legkisebb négyzetes kritérium

$$(5.5.3)\qquad \min_{\substack{(C_1,C_2,\ldots,C_k)\\ \text{partició}}} \sum_{r=1}^{k} \sum_{x_i \in C_r} f(d(x_i,q_r))$$

(ahol q_r a C_r-beli elemek átlagát, d az euklideszi távolságot jelöli, az f pedig egy tetszőleges nemnegatív függvény) általánosítására nézve optimális k-osztályú clusterezéseket (k=n-1, n-2,...,1) adja. Fisher és Van Ness eredménye a következő.

5.5.2. állítás [30]. Tetszőleges folytonos, monoton f függvényre, melyre f(0)=0 az előző eljárás nem hierarchikus, azaz

megadható olyan pontkonfiguráció, amelyre az eljárás által adott partíciók nem alkothatnak finomodó láncot.

A fenti állítás szerint a legkisebb négyzetes kritériumra nézve optimális partíció sorozatot képező eljárás nem hierarchikus megengedett, azaz nem fér bele a hierarchikus módszerek "sémájába".

6. A tárgyalt clusterező eljárások közös algoritmikus vonásai

Ennek a fejezetnek az a fő célja, hogy a clusterezéshez az alkalmazott vagy alkalmazható eljárásokon, illetve a keletkező struktúrákon keresztül kapcsolódó geometriai és kombinatorikus jellegű eredményeket legalább részben bemutassa.

Algoritmikus aspektusból érdekes, hogy a legkisebb négyzetes kritériumra optimális clusterezés, a legközelebbi szomszéd (nearest neighbor, single linkage, Kruskal) módszer, a minimális feszítőfa és a konvex burok keresés közös gyökerekre vezethető vissza.

A clusterezés során természetesen nem az a lényeg, hogy milyen algoritmust használva jutunk el egy alkalmasnak tűnő particióhoz, hanem az, hogy melyik ez a partició, illetve milyen a struktúrája.

Gower és Ross [34] vették észre, hogy a single linkage eljárás outputjai egy minimális feszítőfa (MFF) ismeretében hatékonyabban előállíthatók, mint ha magát a módszer futtatnánk. Ez az észrevétel teszi lehetővé azt is, hogy egyszerűen verifikálhas-

suk egy clusterezésről, hogy az single linkage clusterezés során előállhat-e (6.5. pont).

Ha előre nem ismert egy MFF, akkor vagy megkonstruáljuk, vagy ha legalább egy része ismert, akkor kiegészítjük MFF-vá. Ha a MFF részét csak tippeltük, akkor arra a kérdésre is választ kell adni, hogy általában hogyan tesztelhető egy gráf éleinek valamely részhalmazáról, hogy kiegészíthető-e MFF-vá (6.4.).

Speciálisan a 6.4.12. állítás és a 6.4.13. tétel a minden csúcs legközelebbi szomszédjához vezető élt tartalmazó, ún. NN-gráf (6.4.) körmentes részgráfjának minimális feszítőfává való bővíthetőségéről szólnak. A 6.4.13. tétel szerint a NN-gráffal jól lehet "tippelni" a MFF egy nagy részgráfjára, ti. segítségével meg is konstruálható a MFF éleinek legalább felét tartalmazó részgráf.

Visszatérve az algoritmikus kérdésekhez a konvex burok kereséssel kapcsolatos valószínűségszámítási és algoritmikus eredmények készítik elő a 6.3. és 6.4. pontokat. Utóbbiakban a teljesség igénye nélkül ismertetünk az n pont által definiált minimális feszítőfa (MFF) illetve legkisebb távolság partíció (LTP) konstruálására vonatkozó eredményeket.

A 6.4.-ben bemutatjuk, miként vezethető vissza az euklideszi metrika esetében a MFF és az LTP konstruálása konvex burok keresésre. A cluster analízis azonban nemcsak algoritmikus értelemben kapcsolódik a geometria klasszikus problémáihoz. Egyenletes eloszlás esetén, a legkisebb négyzetes átlagos eltérés kritériumra nézve aszimptotikusan optimális struktúrák (midőn a partició osztályainak száma, k elég nagy), szoros kapcsolatot mutatnak a legsűrűbb kitöltésekkel ([16], [5]).

A 6.5. pontban verifikációs kérdéseket tárgyalunk. A 6.6. pontban clusterező eljárások olyan általános osztályát ismertetjük, amelyek könnyen programozhatók a Kruskal-algoritmus segítségével.

A 6.4. pontban a vizsgált módszereknek az ún. "posta hivatal" és "összes legközelebbi szomszéd" problémákhoz való kapcsolatát is érintjük.

A fejezet eredményei közül kiemeljük a 6.4.7., 6.4.12., 6.5.2. állításokat és a 6.4.13. tételt. Ezek a clusterezésekhez kapcsolódó struktúrákra vonatkoznak.

A fejezetben megemlített ismert eredmények közül az ugyancsak a struktúrák tulajdonságára vonatkozó két karakterizációs tételt emeljük ki. A minimális feszítőfa éleinek egy (globális) jól

jellemzését fogalmazza meg a 6.3.2. állítás. Brown [11] adta meg (6.4.2. tétel) a LTP csúcsainak egy (lokális) jól jellemzését.

A 6.4.7. állításban az ultrametrikus tulajdonságnak egy a szokványos (v.ö.: 3.1.2.) definícióval ekvivalens, geometriai jellegű megfogalmazását adjuk meg az ún. relative neighborhood gráf (RNG) segítségével.

6.1. A konvex burokkal kapcsolatos valószínűségszámítási problémák

A síkbeli pontok által meghatározott konvex burok keresésének algoritmikus problémája illetve a konvex burokkal kapcsolatban felmerülő néhány kérdés a 60-as évektől kezdődően fokozódó érdeklődést keltett. Ezeknek a problémáknak közvetett kapcsolatuk van az eredeti problémakörrel (6.3). Röviden áttekintjük az idevonatkozó eredményeket ([93], [94], [91], [95]).

Tekintsük az m-dimenziós pontok egy n-elemű véletlen halmazát. Jelölje h illetve $E(h)$ a ponthalmaz konvex burkában az extremális pontok számát, illetve ezen pontok számának a várható értékét.

Könnyen lehet olyan síkbeli diszkrét eloszlást konstruálni, mely-

re

$$h \to 3, \quad 1 \text{ valószínűséggel}$$

és

$$E(h) \to 3,$$

midőn $n \to \infty$.

Jelölje l a konvex burok m-1-dimenziós lapjainak (azaz, határoló hipersíkjainak) a számát. Ha m=2, akkor nyilván $h=l$. Abszolút folytonos eloszlás esetén nulla annak a valószínűsége, hogy az n pont közül m+1 egy hipersíkra essen. Ekkor 1 valószínűséggel érvényes a következő összefüggés a konvex burok l lapszáma és h csúcsszáma között (Raynaud [91])

$$l = h(m-1) - m^2 + m + 2.$$

Ebből következik, hogy

$$E(h) \sim E(l)/(m-1), \quad \text{midőn } n \to \infty.$$

Rényi és Sulanke [93], [94] a síkban megadták az $E(l)$ aszimptotikus nagyságát. Raynaud [91] általánosította ezeket m)2 dimenzióban.

6.1.1. tétel. Válasszunk n független pontot a m- dimenziós egy-

séggömbben (B_m) értelmezett egyenletes eloszlás szerint. Ekkor

$$E(I) \sim C_1 n^{(m-1)/(m+1)} \quad (n \to \infty),$$

ahol C_1 egy (effektíven számolható) pozitív konstans.

6.1.2. tétel. Válasszunk n független pontot a m- dimenziós normális eloszlás szerint. Ekkor

$$E(I) \sim C_2 (\ln n)^{(m-1)/2} \quad (n \to \infty),$$

ahol $C_2 = 2^m \pi^{(m-1)/2} / \sqrt{m}$.

Bentley és Shamos [9] a maximális vektorok számának várható értékére vonatkozó eredmények segítségével speciális eloszlások esetén adtak felső korlátot a konvex burok extremális pontjainak a számára. Ez már az E_m egységkockában egyenletes eloszlás esetére is alkalmazható.

6.1.3. tétel (Bentley, Shamos [9]). Válasszunk n független pontot valamely független komponenső m-dimenziós eloszlásból. Ekkor

$$E(h) \leq 2^m \cdot (\ln n)^{m-1} \quad (\text{ha } n \to \infty).$$

6.2. A konvex burok keresés algoritmikus problémái

Konvex burok keresésre vezethető vissza a geometriai problémák egy meglepően széles osztálya. Ezeket a problémákat ismerteti Shamos [100], Toussaint [105], Lengyel [73]. Számunkra a legkisebb távolság partició és a minimális feszítőfa konstruálása szempontjából lesz érdekes a visszavezetés. A LTF (Delaunay gráf) konstruálása eggyel magasabb dimenziós konvex burok kereséssel megoldható, míg a MFF a Delaunay gráf részgráfja, így elég ez utóbbi minimális feszítőfáját megkonstruálni (6.4. pont). Ez utóbbi tulajdonságnak inkább elvi jelentősége van, mert az n csúcsú teljes gráf MFF-jának konstruálására hatékony eljárások ismeretesek.

A továbbiakban általában off-line eljárások legrosszabb esetre vonatkozó algoritmikus bonyolultságával foglalkozunk.

Könnyen megadható a síkban on-line értelemben is optimális, $O(n \log n)$ futási idejű konvex burok konstruáló algoritmus. Ha a ponthalmazt az egyik koordinátája szerint előre szortoltuk, akkor optimális lineáris eljárás is van. Az m=3 esetben Preparata és Hong [88] $O(n \log n)$ optimális algoritmust adtak meg. A térben a burok megadásán a határoló síklapok konvex poligonjainak felso-

rolását értjük és a poligonokat a csúcsok sorozatával adjuk meg.

Sajnos, m)3 dimenzióban keveset tudunk a konvex burok keresésről.

Ha a pontok eloszlásáról valamilyen a priori információ áll rendelkezésre, akkor ezt felhasználhatjuk hatékony eljárások készítésére. Kereshetjük nemcsak a legrosszabb esetben, hanem átlagos értelemben (ti. a futási időnek az összes lehetséges esetben, az eloszlásból adódó, megfelelő súlyozásával vett átlagára nézve) optimális algoritmusokat is.

Ezen kívül beszélhetünk még majdnem minden esetben (1 valószínűséggel) optimális eljárásokról is: ezektől azt várjuk el, hogy optimálisak legyenek minden inputra, kivéve az inputok valamilyen az n növekedésével nullához tartó százalékában.

érdemes megemlíteni, hogy Jarvis [47] a síkban olyan módszert adott meg, aminek a legrosszabb esetre vonatkozó $O(hn)$ futási ideje az outputként megadandó konvex burok extrém pontjainak h számától függ. Ezt speciális eloszlások esetén érdemes használni. A futási idő várható értékére $nO(E(h))$ adódik.

(Megjegyezzük azonban, hogy Kirkpatrick és Seidel 1982-ben [52] Jarvis módszerénél hatékonyabb, a legrosszabb esetben $O(n \log h)$ ideig futó algoritmust adtak meg. Ez a legnagyobb $h=n$ esetén

optimális marad és megőrzi a Jarvis módszernek az output méretére vonatkozó érzékenységét.)

A 6.1.1. tétel kimondása előtt említett eloszlásra átlagban és 1 valószínűséggel a legrosszabb esetben is lineáris ideig fut a Jarvis-eljárás. A 6.1.2. tétel a kétdimenziós normális eloszlásra is $n \cdot \log n$ -nél rövidebb átlagos futási időt biztosít. Kiemelhetjük még Bentley és Shamos [9] eredményét, akik a Jarvis-algoritmust oszdd meg és uralkodj elvvel kombinálva lineáris átlagos futási idejű módszert adtak meg, abban az esetben, ha $E(h)=O(n^p)$, ahol $p<1$.

Ez az eredmény világít rá a 6.1. pontban vizsgált kérdések fontosságára: a 6.1.1. tétel szerint a m-dimenziós egységgömbben egyenletes eloszlásra teljesül az imént említett feltétel.

6.3. A minimális feszítőfa keresésének algoritmikus problémái

Adott n pont és a pontpárokon értelmezett valamilyen távolságfüggvény, amit általában egy olyan szimmetrikus, nemnegatív távolságmátrix alakjában adunk meg, amelynek átlójában csupa nulla áll. A továbbiakban távolságmátrixon mindig ilyen mátrixot

értünk.

Keressük az n csúcson értelmezett megfelelő teljes gráfban azt a feszítőfát, amelyikben minimális az élekhez rendelt távolságértékek összege. Az irodalomban az él költségének vagy súlyának is nevezik az általunk az élt alkotó pontpár távolságaként említett mennyiséget. Ezt a fát szokás minimális feszítőfának (minimális súlyú vagy költségű feszítőfának) nevezni és MFF-vel rövidíteni. Speciális esetként megemlítjük azt, amikor a pontok az m-dimenziós euklideszi tér pontjai és távolságukat az euklideszi távolsággal mérjük.

Általában is értelmes beszélni egy összefüggő gráf minimális feszítőfájáról, azonban számunkra csak a teljes gráf MFF-ja lesz érdekes. Ha a távolságértékek nem mind különböznek, akkor egy gráfnak több különböző MFF-ja is lehetséges. A MFF keresés problémáján egy MFF megadását értjük. Az input tehát egy n*n -es mátrix, az output pedig n-1 él.

A MFF struktúrák alkalmazására a biológiában, az alakfelismerésben, a statisztikában találunk példákat. Az egyik első sikeres kísérlet minimális feszítőfák alkalmazásával történő clusterezésre a 40-es évek végéről való és wroclawi taxonometria

néven ismert (Steinhaus [104]).

A MFF konstruálásra hatékony eljárások ismeretesek. Prim [89] és Dijkstra [20] a pontpárok távolságát (azaz az élek hosszát, vagy más szóhasználattal költségét) reprezentáló tetszőleges, adott távolságmátrix mellett az n-pontú teljes gráfra $O(n^2)$; síkbarajzolható gráfokra Cheriton és Tarjan $O(n)$ ideig futó algoritmust adtak meg ([14]).

Ha az m-dimenziós euklideszi tér n pontjára, valamilyen konstans időben kiszámolható metrikus távolság mellett, a teljes gráfra kell a MFF keresést megoldani, akkor inputként a pontok koordinátáit, azaz m*n input paramétert szokás megadni a távolságmátrix helyett. Ebben az esetben $O(n^2)$-nél gyorsabb eljárások is ismertek: Shamos és Hoey [101] a síkon, az euklideszi metrikára vonatkozóan $O(n \log n)$; Yao [109] tetszőlegesen nagy véges dimenzióban az euklideszi, az L_1 és az L_∞ metrikára nézve $o(n^2)$ ideig futó algoritmusokat publikáltak. Utóbbi az adatok alkalmas szervezésével kerüli el az összes távolságpár túl költséges kiszámítását. Mindkét módszer a teljes gráf egy $O(n)$ élszámú részgráfjában való MFF keresésre redukálja az eredeti feladatot.

A MFF éleinek a meghatározása P-beli az élpárok hosszát egy

lépésben összehasonlító orákulum felett, igy abban az esetben is, amikor inputként csak az egész koordinátájú n pont m*n koordinátáját adjuk meg, a távolságot pedig az euklideszi távolsággal mérjük. Meglepő, hogy ebben az esetben az ún. megengedettségi problémáról (azaz arról, hogy létezik-e olyan feszítőfa, amiben az élek összege legfeljebb akkora, mint egy előre adott pozitív egész) még azt sem tudjuk, hogy NP-beli (Graham [35]) és ez még akkor is igaz marad, ha a problémát egy az élpárok hosszát egy lépésben összehasonlító és az élek hosszát tetszőleges pontosságig kiszámító orákulum felett tekintjük.

A MFF konstruálására használt eljárások nagy részének közös vonása, hogy lépésről lépésre eggyel kevesebb komponensű erdő építésével jutnak el a minimális feszítőfáig. A Kruskal-algoritmus során mindig az összes külső él közül keressük a legrövidebbet, de a 6.3.1. lemma következtében akkor is egy minimális feszítőfához jutunk, ha bármelyik komponens legkisebb kimenő élét vesszük hozzá az erdőhöz.

6.3.1. lemma ([1], 5.2 lemma). Legyen G=(V,E) összefüggő, irányítatlan gráf és legyen

$$((V_1,T_1),(V_2,T_2),\ldots,(V_k,T_k))$$

G egy tetszőleges k-komponensű feszítőerdője (k)1). Legyen

$T = \bigcup_{i=1}^{k} T_i$.

Tegyük fel, hogy az e=(v,w) él a lehető legkisebb költségű (hosszúságú) azon E-T -beli élek között, amelyekre $v \in V_1$ és $w \notin V_1$. Ekkor van G-nek olyan feszítőfája, amely tartalmazza a $T \cup (e)$ halmazt, és amelynek költsége (összhossza) a lehető legalacsonyabb (legkisebb) a T -t tartalmazó feszítőfák között.

A fenti lemmában természetesen nincs jelentősége a komponensek sorszámozásának.

Az előző lemma biztosítja a MFF egyértelműségét abban az esetben, ha az összes távolságérték különböző. Ez utóbbihoz hasonlóan bizonyítható a 6.3.2. állítás, ami a minimális feszítőfák egy jó karakterizációját adja.

6.3.2. állítás. Egy feszítőfa akkor és csak akkor MFF, ha a fához nem tartozó élek közül akármelyiket a fához véve a keletkező körben az élnél nincs hosszabb él.

A 6.3.2. állítás segítségével tetszőleges összefüggő gráfra Komlós [55] az élszám függvényében lineáris számú összehasonlító lépésben teszteli, hogy a gráf egy feszítőfája minimális-e. Ezek

szerint a tesztelési kérdés - legalább is az összehasonlítások száma tekintetében - "optimálisan" megválaszolható.

Ha a gráf éleinek száma nagyságrendileg legalább n log n (megjegyezzük, hogy ennél lényegesen kevesebb élszám is elegendő [1]), akkor a MFF megkonstruálásával ugyanilyen nagyságrendű időben többet is megtudhatunk, azonban ritka gráfokban a tesztelési kérdés az eddig ismert eljárások segítségével gyorsabban megválaszolható, mint ahogyan a konstruálás feladata megoldható.

6.4. A legkisebb távolság partició konstruálása és kapcsolata a minimális feszítőfa kereséshez

A legkisebb távolság partició az egyik alapvető geometriai struktúra, amelyet széles körben alkalmaznak a különböző tudományágakban. A földrajztudomány művelői Thiessen sokszögnek, a fizikusok Wigner-Seitz celláknak, a geometria és az alakfelismerés alkalmazói pedig Voronoi diagrammnak vagy Dirichlet tesszallációnak nevezik.

Tekintsük az R^m-beli különböző P_i $(i=1,2,\ldots,n)$ pontok S halmaza által az m-dimenziós euklideszi tér pontjain indukált $(S_1,S_2,\ldots,S_n)$ **legkisebb távolság particiót**, azaz legyen

$\forall$ i=1,2,...,n esetén

$$S_i = \{ x \in R^m \mid \forall j (i \text{ esetén } x \notin S_j$$

$$\text{és } \forall j \neq i \text{ esetén } d(x,P_i) \leq d(x,P_j) \},$$

ahol d az m-dimenziós euklideszi távolságfüggvényt jelöli.

Az S_i osztályok diszjunkt, nem üres, konvex poliéderek. Mindegyik P_i pontra az S_i osztály tartalmazza azokat a pontokat, amelyek közelebb vannak a P_i ponthoz, mint a többi n-1 ponthoz. A S_i poliéderek csúcsait szokás röviden a LTP csúcsainak, a P_i pontokat pedig a LTP meghatározó pontjainak is nevezni.

Azt mondjuk, hogy a P_i és a P_j pontok (illetve poliédereik) szomszédosak a LTP-ban, ha szakaszfelező merőleges hipersíkjuk a pontokhoz tartozó mindkét konvex poliédert határolja (ez alatt azt értjük, hogy a hipersíkban valamelyik pontnak egy nyílt környezete olyan, hogy az összes pont a P_i és P_j-hez közelebb van, mint a többi S-beli ponthoz).

A Voronoi diagram két lokális tulajdonságának fontos szerepe van az alkalmazásokban.

A később kimondásra kerülő 6.4.11. állításhoz hasonlóan igazolható a

6.4.1. állítás. Legyen a P_i pont legközelebbi szomszédja a P_s pont, azaz

$$d(P_i,P_s) = \min_{\substack{j \neq i \\ 1 \leq j \leq n}} d(P_i,P_j).$$

Ekkor a P_i és P_s pontok szomszédosak lesznek a LTP-ban.

Természetesen a P_i, P_s pontokra a (6.4.1) feltétel is teljesül, ha a P_j helyére a P_s pontot helyettesítjük.

A 6.4.12. állításban és a 6.4.13. tételben visszatérünk a legközelebbi szomszédok egy érdekes tulajdonságára.

A következő tétel segítségével jól lehet karakterizálni a LTP csúcsait ([11]).

6.4.2. tétel.

a/ A LTP poliédereinek bármelyik csúcsa körül rajzolható olyan gömb, amelynek határán (legalább) m+1 S halmazbeli pont van és a gömb belsejében nincs S-beli pont.

b/ Minden R^m-beli pont, amely körül hasonló gömb rajzolható a LTP valamely poliéderének lesz a csúcsa.

A diszkriminancia analízis egyik problémája is közvetve legkisebb távolság kereséshez vezet. Tegyük fel ugyanis, hogy k különböző várható értékű, azonos kovariancia-struktúrájú (kovariancia-mátrixú) m-dimenziós normális eloszláshoz keressük a legjobb lineáris diszkriminátor függvényeket. A maximum likelihood elv ilyen eloszlások esetén k lineáris függvényhez vezet. Egy klasszifikálandó x pontot abba az osztályba sorolunk, amelyikhez tartozó lineáris függvény a maximális behelyettesítési értéket adja, ugyanis az adott m_i várható értékek közül azt keressük, amire a

$$K \exp(-(x-m_i)^T \Sigma^{-1} (x-m_i)),$$

$$(K = (2\pi)^{-m/2} \det(\Sigma)^{-1/2})$$

kifejezés a maximumát veszi fel (Σ jelöli az eloszlások közös kovarianca mátrixát). Ez ekvivalens a

$$-2m_i^T \Sigma^{-1} x + m_i^T \Sigma^{-1} m_i$$

lineáris kifejezés minimalizálásával.

Főkomponens analízis végrehajtása után vizsgálva a problémát az ismeretlen várható értékek legkisebb távolság particiója adja meg azokat a halmazokat, amely pontjaira a fenti döntési elvvel ugyanazt az eloszlást választjuk.

A LTP-t mindig a meghatározó pontokkal adjuk meg. Ha egy

tetszőleges pontról azt szeretnénk megállapítani, hogy a k-osztályú LTP melyik clusterébe esik, akkor erre azonnal választ kapunk pl. a legközelebbi meghatározó pont megtalálásával. "Posta hivatal" problémáról ([53]) beszélünk akkor, amikor egy-egy pontról kell eldönteni, hogy rögzített k pont közül melyikhez van legközelebb. Először határozzuk meg a rögzített pontokhoz tartozó LTP határoló hipersíkjait. Ezek alkalmas preprocesszálása és tárolása után, a bináris keresés többdimenziós általánosításával ([21]) legfeljebb $C_1+C_2 \log k$ (ahol a C_1 és C_2 csak a dimenziószámtól függő pozitív konstansok) lépés is elegendő a probléma megválaszolásához.

Ha valamelyik meghatározó pontot elmozgatjuk, akkor a partició néhány clustere szintén elmozdul, a "szomszédosak" mindenféleképpen. A 6.4.1. állítás egy ilyen szomszédot megad.) Bevezetjük az ún. **Delaunay gráfot**, aminek a csúcsai a LTP meghatározó pontjai lesznek és két csúcsot pontosan akkor kötünk össze (szakasz) éllel, ha a csúcsok a LTP-ban szomszédosak (azaz a csúcsokat összekötő szakasz felezőmerőleges hipersíkja a pontokhoz tartozó mindkét konvex poliédert határolja). Ilyen módon az S ponthalmazhoz tartozó LTP határoló hipersíkjai és a Delaunay gráf élei **duális** fogalmak. Az utóbbi gráf teljes m+1 pontú gráfokat tartalmaz a LTP olyan csúcsai "körül", amelyekre nézve a

6.4.2. tétel a/ pontjában mondott módon jellemző gömbök pontosan m+1 pontot tartalmaznak. Ekkor az m+1 pont által meghatározott szimplex élei a Delaunay gráf élei lesznek. Ha a síkbeli esetben a pontok közül semelyik 4 sincs egy körön, akkor a Delaunay gráfról megmutatható, hogy háromszögekből álló síkgráf, ezért ezt az eredeti pontok Delaunay triangulációjának is szokás nevezni. A Delaunay gráf segítségével egyszerűen megadhatók a LTP csúcsai.

A 6.4.2. tételben megállapított módon jellemezve a LTP csúcsait, Brown [11] egy alkalmas inverziós leképezés segítségével vezette vissza az m-dimenziós Delaunay gráf konstruálásának problémáját egy speciális struktúrájú m+1 -dimenziós konvex burok keresésre.

Shamos és Hoey [101], valamint Toussaint [105] megadták a legkisebb távolság partició (LTP) és a minimális feszítőfa (MFF) közötti alapvető kapcsolatot, ti.

6.4.3. tétel. Tetszőleges m-dimenziós ponthalmaz esetében az euklideszi távolságra nézve MFF a ponthalmaz Delaunay gráfjának részgráfja.

Bizonyítás. Csak vázoljuk a lépéseket. A minimális feszítőfák egy fontos lokális tulajdonságát adta meg Toussaint [105]

(6.4.4. állítás). Ez a tulajdonság biztosítja, hogy a MFF a ponthalmaz ún. relative neighborhood gráfjának (6.4.6. definíció) részgráfja (6.4.8. állítás). Nyilvánvaló, hogy ez utóbbi az ún. Gabriel gráf (6.4.9. definíció) részgráfja (6.4.10. állítás). Ezek után a tételt úgy bizonyíthatjuk be, hogy megmutatjuk: az euklideszi metrika esetében a Gabriel gráf a Delaunay gráf részgráfja (6.4.11. állítás).

Megjegyezzük, hogy a 6.4.4., 6.4.7., 6.4.8., 6.4.10., 6.4.12. állítások és a 6.4.13. tétel lényegében tetszőleges szimmetrikus, nemnegatív és átlójában csupa nullát tartalmazó távolságmátrix esetén igazak.

6.4.4. állítás. A MFF-ban minden P_iP_j élre teljesül, hogy

$$(6.4.1) \qquad d(P_i,P_j) \leq \min_{\substack{1\leq k\leq n \\ k\neq i,j}} \max (d(P_i,P_k),d(P_j,P_k)).$$

6.4.5. megjegyzés. Az S halmazbeli pontokon értelmezett MFF P_iP_j élei a következő feltételnek tesznek eleget: a P_iP_j csak akkor éle a MFF-nak, ha az S-ben nincs olyan $P_iP_jP_k$ háromszög, amelynek a P_iP_j szigorú értelemben véve - a távolságfüggvényre vonatkozóan - leghosszabb oldala lenne.

Itt érdemes megjegyezni, hogy hogyan lehet megválaszolni azt a kérdést, hogy élek valamilyen (körmentes) halmazát lehet-e az eredeti gráf egy minimális feszítőfájává kiegészíteni. Ha az S halmazhoz egyetlen MFF létezik (pl. ha az összes távolságértél különbözik), akkor a MFF megkonstruálásával azonnal adódik a válasz. Ha több egyforma összhosszú MFF létezik, akkor a kérdéses élek hosszát átmenetileg a többi élnél kisebbnek választva a MFF megkontruálása segítségével ugyancsak $O(n^2)$ lépésben egyszerűen megválaszoljuk a kérdést.

6.4.6. definíció. Az S n-elemű halmaz és a pontpárjai halmazán értelmezett $(d(P_i,P_j))_{i,j=1}^{n}$ távolságmátrix relative neighborhood gráfján (RNG) azt a gráfot értjük, amiben pontosan akkor vezet él a P_i és P_j pontok között, ha a (6.4.1) feltétel teljesül.

Mivel a RNG csak olyan háromszöget tartalmazhat, aminek a távolságmátrix szerinti hosszabbik élei egyenlő hosszúak, így igaz a következő

6.4.7. állítás. A RNG akkor és csak akkor teljes gráf, ha a szóbanforgó távolság ultrametrika az S halmazon (vö. 3.1.3. definíció).

A 6.4.4. állításból adódik a

6.4.8. állítás. A minimális feszítőfa a RNG részgráfja.

6.4.9. definíció. Az S n-elemű halmaz és a pontpárjai halmazán értelmezett $(d(P_i,P_j))_{i,j=1}^{n}$ távolságmátrix Gabriel gráfján (GG) azt a gráfot értjük, amiben pontosan akkor vezet él a P_i és P_j pontok között, ha a

$$(6.4.2) \qquad d^2(P_i,P_j) < \min_{\substack{1\le k\le n \\ k\neq i,j}} d^2(P_i,P_k)+d^2(P_k,P_j)$$

feltétel teljesül.

A 6.4.6. és 6.4.9. definíciókban szereplő feltételek közül a (6.4.1) teljesülése maga után vonja a (6.4.2) egyenlőtlenség fennállását is. Pontosabban igaz a következő

6.4.10. állítás. A Gabriel gráf a RNG szupergráfja, ha $d(P_i,P_j)=0$ -ból következik, hogy $P_i \equiv P_j$.

Az előző feltétel teljesül minden metrikus távolság esetén. Speciálisan igaz a

6.4.11. állítás. Az euklideszi metrika mellett a P_i és P_j pontok

között akkor és csak akkor megy él a Gabriel gráfban, ha a LTP-ben a P_i és P_j pontok szomszédosak és poliédereik közös határlapja a lap belső pontjában metszi a két pont között haladó Delaunay gráfbeli (szakasz) élt.

Végül megadunk egy olyan struktúrát is, aminek segítségével a minimális feszítőfa éleinek legalább felét meghatározhatjuk.

Készítsük el az S halmaz pontjain értelmezett következő irányított gráfot: kössünk össze minden pontot a belőle induló és valamelyik legközelebbi szomszédjába mutató éllel. Az így kapott gráfot legközelebbi szomszédsági vagy röviden **NN-gráf**nak nevezzük. Az S halmaznak általában több NN-gráfja van. Mivel minden ilyen gráfban minden pontból pontosan egy él indul ki, így, ha a NN-gráf tartalmaz kört, akkor az csak irányított kör lehet. Tegyük fel, hogy az összes $d(P_i,P_j)$ távolságérték különbözik. Ekkor minden pontnak egyetlen legközelebbi szomszédja, így az S halmaznak egyetlen NN-gráfja van.

Nyilvánvaló, hogy a bemenő élek hosszabbak a kimenőnél, hacsak nem egyszerre legközelebbi szomszédja a P_i-nek a P_j és P_j-nek a P_i pont, amikor a két él egyforma hosszú. Könnyen látható, hogy ekkor a NN-gráfban nincs legalább 3 csúcson átmenő irányított

kör. A fentiek szerint a NN-gráf irányitás nélkül és a többszörös élek egyszeressé tétele után nézve is körmentes, tehát egy legalább n/2 élű erdő. (Megjegyezzük, hogy ha nem teljesül a távolságértékek különbözőségére tett feltétel, akkor minden kör egyenlő hosszú éleket tartalmaz.)

6.4.12. állítás. Tetszőleges olyan távolságmátrix esetén, melyre az összes $d(P_i,P_j)$ távolságérték különbözik, az irányitatlan egyszerű NN-gráf a MFF részgráfja.

Bizonyitás. Tegyük fel, hogy az állitással ellentétben a NN-gráf valamely P_iP_j éle nincs benne a G-vel jelölt MFF-ban. Legyen P_j ennek az élnek a végpontja. Ha a G-hez hozzávesszük ezt az élt, akkor lesz benne az élt tartalmazó kör. Hagyjuk el a keletkezett gráfból a kör azon élét, ami a P_iP_j élhez a P_i-nél csatlakozik. Ekkor ismét egy fához jutunk, azonban az élek összhosszát csökkentettük, mivel a P_i-hez csatlakozó élek közül a P_iP_j kimenő él a legrövidebb. Ezzel ellentmondáshoz jutottunk.

A 6.4.12. állitást bizonyithatjuk úgy is, hogy a minimális feszitőfát a Kruskal-féle mohó algoritmussal konstruáljuk meg. Ekkor minden lépésben az erdő komponensei között haladó legrövidebb külső él hozzávételével csökkentjük a komponensek számát. Az

eljárás során minden pont pontosan abban a lépésben lesz először egy többelemű komponens eleme, amikor a pont a legrövidebb külső él valamelyik végpontja. Az ilyen élek közül viszont a pontot a legközelebbi szomszédjával összekötő él a legrövidebb, tehát ez az él a MFF-nak is éle lesz.

Ha a MFF konstruálását úgy kezdjük, hogy először az egyelemű komponenseket számoljuk fel, akkor a 6.3.1. lemmából (ti. lépésről lépésre V_1-nek választva ezeket a komponenseket) a 6.4.12. állítás bizonyítása szintén azonnal adódik.

Az egyenlő távolságértékek esetére általánosítja a 6.4.12. állítást a

6.4.13. tétel. Tetszőleges távolságmátrix esetén a $G^*=(V,E^*)$ irányított NN-gráf bármely irányítatlan, egyszerű és körmentes $G_1=(V,E_1)$ részgráfjához található a G_1-t tartalmazó MFF.

Ha a 6.4.13. tételben tekintett G_1 gráf maximális abban az értelemben, hogy további él hozzávétele a körmentesség feltételét sértené, akkor a G_1 egy legalább n/2 élű erdő. Ez abból következik, hogy a NN-gráfban a körök csúcshalmazai páronként

diszjunkt halmazrendszert alkotnak. A 6.4.13. tétel biztosítja, hogy valamelyik MFF tartalmazza ezt az erdőt. Nemcsak a teljes, hanem tetszőleges összefüggő gráfban is bármely NN-gráf segítségével az élszámtól lineárisan függő lépésszámban valamelyik minimális feszítőfa éleinek legalább a felét megkaphatjuk. (A körök észleléséhez használhatjuk a kétszeresen összefüggő komponensek kijelölésére kis bővítéssel alkalmas, mélységi kereső fát (depth-first tree) konstruáló algoritmust is, pl. [1], 5.3. algoritmus)).

A maradék legfeljebb n/2 komponenst egy-egy szögponttá összehúzva új gráfot készítünk, amelyben az egyes pontok között futó élek hossza a megfelelő komponensek közötti minimális élhosszal egyenlő. Ilyen módon $O(n^2)$ futási idejű MFF konstruáló eljáráshoz juthatunk. Ennek a gondolatnak a továbbfejlesztésével nagy élsűrűségű gráfokra is készíthető az élek számában lineáris futási idejű algoritmus (Aho, Hopcroft, Ullman [1]). Ennek a módszernek az is előnye, hogy a NN-gráf élei parallel számításokkal adhatók meg.

Érdemes megemlíteni, hogy a MFF konstruáláshoz elegendő a 6.4.12. állítás is, hiszen apró súlyváltoztatásokkal elérhetjük, hogy minden távolság különböző legyen. Ezek után megkonstruáljuk

a NN-gráfot és az összehúzástól folytatjuk az előbbi gondolatmenetet.

"All nearest neighbor" problémáról beszélünk akkor, amikor egy adott súlyozott élű gráfban minden ponthoz egyszerre keressük a legközelebbi szomszédját. Tetszőleges olyan gráfban, amelyben nincs izolált pont egyszerű megadni olyan algoritmust, ami az élszámtól lineárisan függő lépésben megoldja a problémát. Speciálisan R^2-ben az euklideszi távolságra nézve optimális, nagyságrendileg $n \cdot \log n$ futási idejű eljárás ismert (a Delaunay gráf elkészítése után a legfeljebb 3n-6 él átvizsgálásához már csak $O(n)$ lépés kell, hiszen a 6.4.1. állítás szerint a NN-gráf az előbbi részgráfja).

A 6.4.13. tétel bizonyítása. A bizonyítás csak annyiban tér el a 6.4.12.-étől, hogy a NN-gráf helyett a G_1 éleiről kell belátni, hogy a G-vel jelölt MFF élei közül valók, vagy a MFF élei ezekre cserélhetők. Most a $P_i P_j \in E_1$ élnek a G gráfhoz vételekor olyan kör keletkezik, amelyben az élhez a P_i pontnál csatlakozó e' él nem lehet hosszabb a $P_i P_j$ élnél, hiszen a G-ről feltettük, hogy MFF. A fordított esetben a P_i-nek nem a P_j pont lenne a legközelebbi szomszédja. Ezek szerint a két él egyenlő hosszú. Ha $e' \notin E_1$, akkor a G minimális feszítőfában az e' él lecserélhe-

tő a P_iP_j-re. Ellenkező esetben a körben az e' él P_i-től különböző végpontjánál csatlakozó e'' élt vesszük sorra és a fentiekhez hasonlóan belátjuk, hogy a P_iP_j éllel egyenlő hosszú. Mivel a kör összes éle nem tartozhat a körmentes E_1-hez, ezért ezt az eljárást folytatva olyan élhez jutunk, ami lecserélhető lesz a P_iP_j élre. Ezzel a módszerrel megadható egy olyan MFF, ami a G_1 összes élét tartalmazza.

A Kruskal-algoritmus fontos tulajdonságát adja meg a

6.4.14. tétel. A Kruskal-algoritmus egyes lépései során keletkező erdő az azonos számú komponensből álló erdők közül minimális összhosszúságú.

Egy gráf körmentes élhalmazai matroidot (ún. grafikus matroidot) alkotnak. A 6.4.14. tétel következik abból az általánosabb állításból, hogy egy matroid valamelyik minimális súlyú bázisát (azaz a matroidelmélet szóhasználatával: a matroid egy minimális összsúlyú maximális "független" halmazát) adja meg az analóg mohó algoritmus. Ennek bizonyítása során (Lawler [60], 7. fejezet, 6.1. tétel) kiderül az is, hogy a mohó algoritmus k. lépésében választott báziselem legfeljebb akkora súlyú lehet, mint bármely

más "független" halmazban az elemek súlya szerinti növő sorrendben a k. elem.

Ebből egyrészt következik az előző tétel, másrészt az is, hogy ha egy gráfnak több különböző minimális feszítőfája van, akkor az élek között megadható hosszúságtartó megfeleltetés.

A fentiek legfontosabb következménye azonban az, hogy a single linkage eljárás lépései nemcsak ésszerűek a clusterezés szempontjából, hanem minden lépésben egy globális optimalizálási - ti. a bevezetésben 2.-vel jelölt "szeparálási" - feladatot is megoldanak.

6.4.15. tétel. A Kruskal-algoritmus egyes lépései során keletkező particionálás az azonos számú osztályból álló partíciók között maximalizálja a komponensek között haladó élek minimális hosszát.

Megjegyezzük, hogy az euklideszi távolság esetében Yao [109] egy olyan gráfot használ a MFF konstrukciójához, amit tulajdonképpen a legközelebbi szomszéd fogalom általánosításával kap. Ez a gráf a MFF szupergráfja és szintén $O(n)$ élt tartalmaz. A feladat ezek után a viszonylag kevés élű gráf MFF-jának keresésére vezethető vissza.

Mivel két dimenzióban a Delaunay gráf síkbarajzolható, így a

konvex burok keresés $O(n \log n)$ lépése után $O(n)$ időben megkonstruálható a MFF (hiszen a 6.4.3. tétel szerint a MFF szupergráfja az ebben esetben legfeljebb $3n-6$ élű Delaunay gráf). Magasabb m-dimenzióban a Delaunay gráf síkbarajzolhatósága megszűnik ([105]), sőt akár $m+1$-es teljes gráf is lehet [70].

Amíg az n-pontú S halmaz konvex burkának R^2 és R^3-ben nagyságrendileg n éle, addig R^4-ben n^2 éle és 3-dimenziós hiperlapja lehet. Így R^3-ben a LTP-nak nagyságrendileg n^2 csúcsa lehet, és ez alátámasztja Preparata [87] azt az eredményét, miszerint bármilyen algoritmus, amivel R^3-ban a LTP megkonstruálható a legrosszabb esetben nagyságrendileg legalább n^2 ideig kell, hogy fusson.

6.5. A single linkage eljárás során adódó clusterezések származtatása és tesztelése tetszőleges MFF segítségével

A clusterezéssel gyakorlati módon kapcsolatba kerülő szakemberek többféleképpen reagálnak egy-egy clusterezés eredményének a kiértékelésekor. Egy gyakori hozzáállás, amikor az outputot azzal a megjegyzéssel fogadják el, hogy azért jobb lenne, ha egyik-másik

pont a partició konkrétan megjelölt másik osztályába kerülne.

Felvetődik a kérdés, hogy egy adott módszerre vagy kritériumra lehetséges-e olyan bizonyítékot találni, ami amellett szólna, hogy a módszer biztosan nem szolgáltathat egy konkrét outputot.

Az irodalomban több olyan módszer található, aminek célja a clusterezés outputjainak az összehasonlítása, de hasonló verifikációs kérdéssel nem találkoztunk.

A következő problémára adunk választ a single linkage eljárás esetében: adott egy n-csúcsú, súlyozott élű teljes gráf és egy minimális feszítőfája:

adott k-particióról döntsük el, hogy az a Kruskal-algoritmus n-k. lépésében keletkezhetett-e.

Gower és Ross [34] mutattak rá arra, hogy a Kruskal-algoritmus lépései származtathatók egy MFF-ból. Hagyjuk el egymás után egyesével a gráf egy MFF-jának leghosszabb k-1 élét. Ha több egyenlő él is szóba jön, akkor valahogyan döntünk arról, hogy melyik élt hagyjuk el. A 6.4.14. tétel szerint ez az elhagyások után keletkezett erdő a k-komponensű erdők közül minimális összhosszúságú lesz. Ezt állapítja meg a

6.5.1. állítás. A teljes single linkage eljárás során adódó erdő sorozat $O(n \cdot \log n)$ időben megadható a gráf egy adott MFF-ja segítségével. A k-komponensű erdő megkonstruáláshoz $O(n)$ lépés is elegendő.

Bizonyítás. A Kruskal-algoritmus domináns lépése az élek hosszúság szerinti rendezése. Speciálisan egy feszítőfa esetében $O(n \cdot \log n)$ az algoritmus futási ideje. A leghosszabb $k-1$ él a komponensek meghatározásához nagyságrendileg n lépés kell. A komponensek a megadott időkben előállíthatók.

Az előző állításhoz hasonlóan válaszolható meg a verifikációs kérdés is

6.5.2. állítás. Egy MFF ismeretében, tetszőleges (rögzített) k természetes számra a gráf csúcshalmazának bármelyik k-particiójáról $O(n)$ idő alatt verifikálható, hogy az a Kruskal (single linkage) algoritmus n-k. lépésében keletkezhetett-e.

Bizonyítás. Az eredeti gráf egy MFF-jának ismeretében, a módosított Kruskal-algoritmussal (3.1. pont) konstruáljuk meg a megfelelő, ún. szubdomináns (3.1. pont) ultrametrika szerinti szintgráfokat. (Mindegyik MFF az egyetlen szubdomináns ultramet-

rikát definiálja a teljes gráf élein.) Tudjuk, hogy ennek komponensei teljes részgráfokat feszítenek (5.4.4. állítás).

Határozzuk meg a MFF leghosszabb k-1 élét. Válasszuk a legkisebb hosszát szintszámnak. Ha a kérdezett k-partició minden osztályát teljesen tartalmazza az adódó szintgráf (legfeljebb) k komponensének valamelyike, akkor a verifikációs kérdésre igen a válasz, különben nem.

6.6. Kombinatorikus clusterező módszerek

A hierarchikus clusterező eljárások egy nagy családja könnyen programozható. Ezeknek az eljárásoknak a közös jellemzője, hogy minden lépésben a partició egyesítendő osztályainak meghatározása a clusterpárokra vonatkozó kritérium alapján történik. A kritérium alkalmazása a clusterek távolságára (különbözőségére) vonatkozó olyan mennyiség minimalizálását jelenti, amelyet egyszerűen lehet updatelni az egyesítő lépést követően.

A clusterezés valamelyik szintjén legyen a C_i és C_j az a két egyesítendő cluster, amelyek d_{ij} clusterek közötti távolsága a legkisebb. Legyen a C_h tetszőleges egyéb cluster ugyanezen a szinten. Jelölje C_k az egyesítés után adódó clustert és d_{hk} a C_h

távolságát a C_k clustertől.

6.6.1. definíció [58]. Kombinatorikusnak nevezzük azokat a clusterező eljárásokat, amelyekre a

$$(6.6.1) \qquad d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}|$$

összefüggés érvényes, ahol $\alpha_i, \alpha_j, \beta, \gamma$ együtthatók csak a C_i, a C_j (esetleg a C_k és a C_h) elemszámától függenek.

A clusterek távolsága a kezdeti egy-egy elemű clusterekre a pontok eredeti távolsága, a clusterezési lépések után pedig a clusterek pontpárjai közötti távolságok valamilyen egyszerű függvénye szokott lenni. A single linkage eljárás esetében

$$(6.6.2) \qquad d_{ij} = \min_{\substack{x_i \in C_i \\ x_j \in C_j}} d(x_i, x_j),$$

a complete linkage-nél

$$(6.6.3) \qquad d_{ij} = \max_{\substack{x_i \in C_i \\ x_j \in C_j}} d(x_i, x_j),$$

a group-average (csoport-átlag) linkage-re pedig

$$(6.6.4) \qquad d_{ij} = \frac{1}{|C_i| \cdot |C_j|} \sum_{\substack{x_i \in C_i \\ x_j \in C_j}} d(x_i, x_j).$$

A single, a complete, a group-average linkage, a centroid, a medián módszerek [58] kombinatorikusak, következésképpen a Krus-

kal-algoritmus segítségével $O(n^2 \log n)$ futási idejű programmal realizálhatók.

Az alábbi 6.1. táblázatból kiolvashatók a megfelelő együtthatók:

módszer	α_i	α_j	β	γ
single linkage	0.5	0.5	0	-0.5
complete linkage	0.5	0.5	0	0.5
group-average linkage	n_i/n_k	n_j/n_k	0	0
centroid	n_i/n_k	n_j/n_k	$-\alpha_i\alpha_j$	0
medián	0.5	0.5	-0.25	0

6.1. táblázat

(ahol n_i, n_j és n_k jelöli a C_i, a C_j és a C_k elemeinak számát).

A single és complete linkage eljárás során az összevonásokhoz tartozó távolságértékek monoton növő sorozatot alkotnak. Az összevonási kritérium biztosítja, hogy

$$\min (d_{hi}, d_{hj}) \geq d_{ij} \qquad (6.6.5)$$

teljesüljön. Ezért $d_{hk} \geq d_{ij}$ abban az esetben is, amikor $\gamma \geq 0$ és $\alpha_i + \alpha_j + \beta \geq 1$.

Az előbb említett sorozat monotonítása lehetővé teszi, hogy a hierarchikus módszer által az S halmaz pontjain értelmezett

súlyozott élű fa segítségével reprodukálhassuk a módszer egymás utáni lépéseit. A fát úgy definiáljuk, hogy minden lépésben az egyesítendő clusterek valamely pontpárja között élt húzunk be és az élhez a clusterek közötti távolságot rendeljük súlyként. (A single és complete linkage eljárás esetében ez a távolság megfelel egy-egy pont eredeti távolságának is, a group-average módszernél azonban általában nem található ilyen pontpár.)
Az összevonáskor fellépő távolságértékek monoton növekedése miatt pontosan a módszer egyesítő lépéseit kapjuk a megfelelő sorrendben, ha ezt a fát, mint önmaga minimális feszítőfáját a Kruskal-algoritmus segítségével építjük fel. A módszer ilyen n-pontú fagráffal való reprezentálása nem egyértelmű. Megjegyezzük, hogy a fa által a 3.1. pont példájában említett módon definiált ultrametrika és a módszerhez tartozó, a 3.1.4. tétel bizonyításában definiált ultrametrika ekvivalensek.

Ling [75] és Matula [81] kezdeményezte az ún. gráfelméleti clusterezéseket. A javasolt módszerek a gráfelméletből ismert struktúrák keresésével kívánják megoldani a clusterezés feladatát. Ezek egy része - valamilyen rögzített k mellett - **maximális k-klikkeket, k-szorosan (csúcs)összefüggő feszített részgráfokat, k-szorosan élösszefüggő komponenseket** illetve k-

kötegeket (k-clusterek, weak k-linkage clusterek, k-bond [81]) keresnek a gráfban, illetve annak speciális részgráfjaiban: a szintgráfokban. (A k-köteg olyan feszített részgráf, amiben az összes csúcsnak a részgráfra vonatkozó foka legalább k.)

A cluster struktúrát a különböző λ szinteken tekintett (különböző) részgráfok megfelelő gráfstruktúráinak csúcshalmazai adják, midőn λ a legkisebb távolságtól kezdve a legnagyobbig nő.

A maximális k-klikk és k-szorosan összefüggő komponenst adó eljárások nem feltétlenül diszjunkt felbontását adják a gráf csúcshalmazának, míg a másik két struktúra hierarchikus agglomeratív clusterezési eljáráshoz vezet.

Egy másik megközelítés során az adott n-pontú gráfra minden lehetséges k érték (k=n,n-1,...,1) mellett meghatározzuk a maximális k-kötegeket illetve k-szorosan élösszefüggő komponenseket. A kapott csúcshalmaz sorozatokat kiegészítve az alaphalmazzal egy-egy hierarchikus struktúrát kapunk.

7. Statisztikai hipotézisvizsgálatok

Végül szeretnék néhány megjegyzést fűzni a cluster analízishez, alkalmazói szempontból.

Az utóbbi évtizedekben számos olyan alkalmazás született, ami rávilágít a cluster analízisben rejlő lehetőségekre. Az alkalmazások során született eredményeknek a megértése, verifikálása, jelentőségük felismerése elsősorban a problémát felvető szakág szemszögéből történt meg. A matematikai szempontok alapján történő kiértékelés általában elmarad.

Ebben a fejezetben ismertetjük azokat az erőfeszítéseket, amelyeket az állítások statisztikai verifikálhatósága érdekében tettek. Egyelőre nem rendelkezünk olyan apparátussal, amit közvetlenül használhatnánk, legfeljebb nagy minta esetén közelítő értelemben.

Természetesen ezek a - talán szkeptikusnak tűnő - megjegyzések nem azt a célt szolgálják, hogy megkérdőjelezzük a clusterezés létjogosultságát. Semmi esetre sem szeretnénk a szakági keretek között tudományosan megmagyarázható, újat hozó eredményeket kétségbe vonni, hiszen a szakág szempontjából lényegtelen az alkalmazott módszer matematikai apparátusa (eltekintve attól,

hogy esetleg reális időben nem hajtható végre).

7.1. A clusterezések statisztikai kiértékelése

Az eredmények objektív értékeléséhez a matematikai statisztika eszközeivel szeretnénk módszert találni.

Dönteni kívánunk két hipotézis között. Az egyik szerint a mintaelemek eloszlása az eloszlásfüggvények valamilyen paraméterekkel, vagy egyéb módon meghatározott halmazának egy eleme. Az ellenhipotézis (alternatív hipotézis) egy másik halmazt jelöl meg.

A hipotézisek vizsgálatára a statisztikai próbáknak nevezett eljárások szolgálnak. Meghatározzuk a megfigyelések valamilyen statisztikájának az eloszlását mind a null-, mind az ellenhipotézis esetén. A statisztika számolt értéke alapján döntünk arról, hogy melyik hipotézist fogadjuk el: és pedig aszerint, hogy ez az érték az elfogadási vagy a kritikus tartományba esik.
A döntés során elkövethető hibák valószínűségét a statiszika és a fenti két tartomány megválasztása határozzák meg. Ha semmiféle előzetes információnk sincs a mintaelemek közös eloszlásáról, akkor első lépésben a (jelenség sajátosságaiból adódó) megfelelő egyszeresen összefüggő tartományban egyenletes eloszlást vá-

laszthatjuk nullhipotézisnek. Az ellenhipotézisbe az összes ettől eltérő eloszlás tartozik. Ha a nullhipotézist fogadjuk el, akkor ez azt támasztja alá, hogy a jelenség kimenetelei között nem figyelhető meg szisztematikus csoportosulás.

Ha az előző vizsgálatban az ellenhipotézist részesítettük előnyben, akkor az eloszlásosztályok alkalmas választásával a csoportosulások jellegét tovább tesztelhetjük. A jelenséget elemző szakemberek leginkább arra kíváncsiak, hogy a megfigyelések eloszlása milyen és hány "egyszerűbb" eloszlás keverékeloszlásaként adódik.

A különböző hipotézisek között esetleg más és más statisztika segítségével lehet hatékonyan választani.

Sajnos, kevés olyan eredmény ismert, ami a clusterező eljárások során használt kritériumok alapján számolt statisztikák eloszlását adná meg. Az ilyen eredményekre nagy szükség lenne, hiszen ezek segítségével választhatnánk ki a jelenséget leíró eloszlásosztályt jól "felismerő" clusterező eljárásokat.

A hierarchikus módszerekkel kapcsolatban megemlítjük, hogy a complete linkage és az average linkage eljárásokra vonatkozóan [42]-ben található idevágó eredmény egydimenziós eloszlások

esetén.

Nem találtunk hasonló eredményt a single linkage eljárással kapcsolatban $m)2$ dimenzióban, de érdemes megemlíteni Steele [103] eredményét a minimális feszítőfa hosszával kapcsolatban.

7.1.1. tétel. Tegyük fel, hogy az X_i, $1 \leq i \leq n$ független, azonos $F(x)$ eloszlású m-dimenziós ($m \geq 2$) valószínűségi változók. Tegyük fel, hogy az eloszlás korlátos tartójú. Ekkor 1 valószínűséggel

$$\lim_{n \to \infty} n^{-(m-1)/m} M(X_1, X_2, \ldots, X_n) = c_m \int_{R^m} p(x)^{(m-1)/m} dx,$$

ahol az $M(X_1, X_2, \ldots, X_n)$ az X_i, $1 \leq i \leq n$ pontokhoz tartozó euklideszi minimális feszítőfában az élek hosszának összegét, az $p(x)$ pedig az eloszlás abszolút folytonos részének a sűrűségfüggvényét jelöli. A c_m csak az m dimenziótól függő konstans.

A konvergencia sebességéről és a c_m számról keveset tudunk, bár az utóbbi becsülhető (pl. Monte-Carlo módszerrel, az egyenletes eloszlás esetében a $[0,1]^m$ egységkockára alkalmazva). Sajnos, az $M(X_1, X_2, \ldots, X_n)$ mennyiségre centrális határeloszlás jellegű tétel és megfelelő próba sem ismeretes.

A nem hierarchikus módszerekkel kapcsolatban több eredmény található az irodalomban. A 7.2. pontban vázoljuk a szorosan kapcsolódó kvantizálási eredményeket, végül a 7.3. pontban a legkisebb négyzetes kritériumra nézve optimális kvantizálások/clusterezések tulajdonságait vizsgáljuk.

7.2. Clusterezés és kvantizálás

A clusterezés szorosan kapcsolódik az ún. kvantizálási témakörhöz. Feltételezzük, hogy minden ξ megfigyelés a többitől függetlenül figyelhető meg és az m-dimenziós **ismert** $F(x)$ eloszlás szerint oszlik el. gy gondoljuk, hogy az F eloszlás k, homogénnek tekintett eloszlás keveréke. Ezek után a megfigyelési értékek lehetséges halmazának - általában az R^m-nek - egy particióját keressük. A két problémakört tehát aszerint különböztetjük meg, hogy az eloszlás ismeretében, illetve ennek hiányában - valamilyen minta alapján - kell az optimális veszteséget és a struktúrát meghatározni. Az előbbi esetben kvantizálási problémáról beszélünk, míg az utóbbinál használjuk a clusterezés megjelölést.

A kvantizálás során az a feladatunk, hogy megadjunk egy valamilyen értelemben legjobb vagy elég jó ún. kvantizáló függvényt, ami a partició különböző osztályain az osztályra jellemző különböző (m-dimenziós) konstans értéket vesz fel. Ezeket az értékeket nevezzük reprezentáns pontoknak. Az alábbiakban ξ-nek a megfelelő reprezentáns ponttól mért euklideszi távolsága r. hatványával mérjük a kvantizálás hibáját, és L_r-normában, azaz az átlagos értelemben legjobb kvantizáló függvényt keressük.

Példa: Egy mérőszámokkal jellemezhető véletlen eseményről szeretnénk információkat továbbítani, de minden eseményről csak egy adott számú, mondjuk k-féle jel valamelyike lehet az üzenet. A fogadó szeretné minél kisebb hibával azonosítani a valódi információt: keressünk egy olyan k-értékű kvantizáló függvényt, amelyikre legkisebb az eredeti ξ információ (input) és a kvantizáló függvény (output) L_r-normában vett eltérése.

Az algoritmikus szempontból a 2.2. pontban bevezetett k-közép clusterező és Lloyd kvantizáló eljárása lényegében nem különböznek egymástól.

Nagy clusterezendő minta esetén azt várhatjuk, hogy az optimális clusterezések struktúrájáról, a veszteség nagyságáról a kvantizá-

lás eszközeivel következtetéseket vonhatunk le.

A (2.2.1)-ben definiált $W_n(S)$ statisztikát használjuk a hipotézisvizsgálat során. A (7.3.5) kifejezésben definiált $W(k;P,r)$ mennyiség (az optimális kvantizálási hiba) adja a P mértékre vonatkozó elméleti értéket. A 7.3. pontban két centrális határeloszlás típusú eredményt idézünk, amelyek nagy megfigyelésszám esetén valamilyen formában esetleg alkalmazhatók a döntések során.

7.3. A legkisebb négyzetes kritériumra nézve optimális kvantizálások és clusterezések tulajdonságai

Optimális struktúrák

Speciálisan az egyenletes eloszlás esetén pontosan ismert az optimális kvantizáló struktúra az egyenesen, a síkon (Fejes Tóth [28]), illetve a térben a speciális, ún. rácspont kvantizálókra (Barnes és Sloane [5]). Azonban olyan egyszerű esetekben sem ismerjük az optimális struktúrát, mint amikor a kétdimenziós normális eloszlást kell kvantizálni (Gray és Karnin [36]).

Általában nem lehetséges általános eloszlás esetében visszavezetni az optimális kvantizálási struktúra keresését az egyenletes eloszlás esetére (Heppes és Szüsz [43]).

Optimális struktúrák unicitását biztosító feltételek és a struktúrák keresésének algoritmikus kérdései

Általános esetben egy eloszlásnak több lokális optimumot adó kvantizálása is lehetséges és a lokális optimumok közül több is a globális optimummal egyenlő veszteséget adhat. Azonban a 7.3.2. tétel szerint az egydimenziós esetben, az eloszlásra vonatkozó elég általános, az optimumhely **unicitását** biztosító feltételek mellett a Lloyd-algoritmussal tetszőleges pontossággal, a pontosság függvényében viszonylag gyorsan megkaphatjuk az optimális kvantizálás struktúráját meghatározó elválasztó pontokat.

Lloyd [77] 1982-ben publikálta teljes terjedelemben 1957-ben elért, az optimális kvantizálásra vonatkozó eredményeit. Nem talált elégséges feltételt, ami a minimumhely unicitását biztosította volna, azonban megadott egy a későbbiekben jól használhatónak bizonyult algoritmust (2.2. pont, Lloyd-algoritmus).

Fleischer [31] 1964-ben talált olyan elég általános feltétele-

ket, amelyek biztosították egyetlen lokális minimumhely létezését. Ez egyúttal globális optimumhely is és Lloyd módszere ehhez konvergál (vagy véges sok lépésben eljut).

Truskin [106] 1982-ben általánosította a feltételeket egy általános hibafüggvényosztály elemeire.

Jelölje P illetve F a szóbanforgó egydimenziós valószínűségi változó által indukált mértéket illetve ismert eloszlásfüggvényét.

7.3.1. tétel (Truskin [106], 4. tétel). Legyen $I=(v,w)$, $(-\infty \leq v < w \leq \infty)$ olyan nyílt intervallum, hogy az $F(x)$ eloszlásfüggvény $p(x)$ sűrűségfüggvénye pozitív az intervallum belső pontjaiban, míg $p(x)=0$ különben. Tegyük fel ezen kívül, hogy a $p(x)$ logkonkáv az I intervallumban és véges második momentuma van (azaz $\int_v^w x^2 p(x)\,dx < \infty$). Ekkor az L legkisebb négyzetes hibafüggvénynek egyetlen stacionárius pontja, így egyetlen lokális minimumhelye van.

Kieffer azt vette észre, hogy ha egy kicsit finomítja a feltételeket, akkor a minimumérték viszonylag kevés számolással is jól közelíthető. (Most tulajdonképpen nemcsak azt tételezzük fel, hogy a p(x) ismert, hanem azt is, hogy a Lloyd-algoritmus során

számolandó $q_i^{(n)}$ kvantizáló értékeket egy orákulum ([79]) tetszőleges általunk kívánt pontossággal megadja, így az integrálásból származó bonyodalmaktól eltekinthetünk.) A minimális veszteség n tizedesjegy pontosságig való meghatározásához az iterációs lépések száma legfeljebb lineárisan nő az n függvényében. Jelölje (x_i^*) illetve (q_i^*) az optimális elválasztó illetve kvantizáló sorozatot, ekkor

<u>7.3.2. tétel (Kieffer [51])</u>. Ha a p(x) sűrűségfüggvény kielégíti a 7.3.1. tétel feltételeit, valamint

vagy a/ $v \neq -\infty$ vagy $w \neq +\infty$

vagy b/ $(v,w) = (-\infty,+\infty)$, de log p(x) nem szakaszosan lineáris,

akkor létezik olyan λ: $0 < \lambda < 1$, hogy elég nagy n esetén

$$\max_{i=1,2,\ldots,k-1} \left| x_i^{(n)} - x_i^* \right| < \lambda^n, \tag{7.3.1}$$

$$\int_{-\infty}^{+\infty} \left(Q^{(n)}(t) - Q^*(t)\right)^2 dF(t) < \lambda^n, \tag{7.3.2}$$

és

$$L\left((x_i^{(n)}),(q_i^{(n)})\right) - L\left((x_i^*),(q_i^*)\right) < \lambda^n, \tag{7.3.3}$$

ahol $Q^{(n)}$ a Lloyd-algoritmus n. lépésében adódó, a Q^* pedig az optimális kvantizáló függvényt jelöli.

A 7.3.1. illetve a 7.3.2. tétel abszolút folytonos eloszlásokról szól, és a feltételeket nyilván nem teljesítik azok az eloszlások, amelyek sűrűségfüggvénye nem unimodális, hiszen a logkonkáv tulajdonság maga után vonja az unimodalitást.

Magasabb dimenzióban a fenti tételekben szereplő elégséges feltételekhez hasonló feltételek nem ismeretesek. Egyszerűbb esetekben analitikusan is meg lehet határozni a stacionárius pontokat. Ott, ahol ez nem lehetséges a Lloyd-algoritmus most is adhat némi támpontot.

Minden forgásszimmetrikus eloszlásra, így a standard normálisra is igaz, hogy a legkisebb négyzetes kvantizálási hiba invariáns a kvantizáló ponthalmaz origó körüli elforgatására.
Legyen $G:R^m \rightarrow R^m$ egy olyan leképezés, amelyik az eredeti koordinátákat permutálja és esetleg tükrözi. Ilyen módon egy távolságtartó leképezést definiáltunk. Ennek következtében a legkisebb távolság particiók (LTP) legkisebb távolság particiókba mennek át a G alkalmazása során. Mivel a G Jacobi-determinánsa 1, így a LTP centroidjaként adódó kvantizáló vektorokat a képtérbeli megfelelő kvantizáló vektorokba viszi a G transzformáció, és a legkisebb négyzetes kvantizálási hiba változatlan marad.

A fentiek szerint forgásszimmetrikus eloszlások esetén nem érdemes különbözőnek tekinteni a lokális optimumot adó konfigurációk közül azokat, amelyek egymásból elforgatással vagy a koordináták permutálásával/tükrözésével (vagy az előbbiek kombinációjával) keletkeznek.

Amíg az egydimenziós normális eloszlásra a 7.3.1. illetve a 7.3.2. tétel feltételei teljesülnek, és ezek a lokális optimumhely unicitását biztosítják, addig a legegyszerűbbnek tűnő esetben, a kétdimenziós standard normális eloszlásra nem ismeretes a lokális optimum unicitását biztosító elégséges feltétel.

A kvantizálási hiba nagyságrendjének aszimptotikus vizsgálata

Vizsgáljuk meg, hogy aszimptotikusan hogyan függ a kvantizálás hibájának a nagyságrendje a kvantizáló értékek számától akkor, amikor a veszteséget az input és output közötti eltérések r. hatványával mérjük.

Cox [19] 1957-ben veti fel a kérdést a clusterezés terminológiájával megfogalmazva (az r=2 esetre). Fisher és Van Ness [30] közvetve próbált a clusterezési veszteség nagyságrendjének a megtippelésével a clusterek választandó számára ötletet adni.

Bolsev [10] matematikai statisztikai heurisztikával állapította meg a nagyságrendet. Lengyel és Ruda [61], [62] lényegében megoldották Cox problémáját.

Jelölje p a Lebesgue-mértékre nézve abszolút folytonos F eloszlás sűrűségfüggvényét illetve P a megfelelő mértéket. Legyen $|Q|=k$ és

$$W(Q;P,r) = \int_{R^m} \min_{q_i \in Q} \|x-q_i\|^r \, dF(x)$$

a k-szintű kvantizálás során keletkező - az input-output eltérések r. hatványával mért - átlagos hiba. Jelölje $V(k;P,r)$ az előző mennyiség Q szerint vett $W(k;P,r)$ minimumának a k-val való következő súlyozott függvényét

$$(7.3.4) \qquad V(k;P,r) = k^{r/m} W(k;P,r),$$

ahol

$$(7.3.5) \qquad W(k;P,r) = \inf_{|Q|=k} W(Q;P,r).$$

7.3.3. tétel [61]. Tekintsünk egy m-dimenziós valószínűségi vektorváltozót, aminek jelölje $F(z)$ az ismeretlen eloszlásfüggvényét illetve P a megfelelő valószínűségi mértéket. Tegyük fel, hogy

(i) $\exists$ olyan $C \subset R^m$ korlátos és zárt (azaz kompakt), valamint konvex halmaz, hogy $P(C)=1$ úgy, hogy

(ii) $P(U) > 0$, $\forall$ nyílt $U \subset C$ esetén;

(iii) az $F(z)$ abszolút folytonos

(iv) létezik olyan pozitív, véges K_1 és K_2, hogy $K_1 \leq p(z) \leq K_2$ $\forall z \in C$ pontra, ahol $p(z)$ jelöli az $F(z)$ sűrűségfüggvényét.

Ekkor létezik olyan (az F eloszlástól függő) pozitív, véges K_3 és K_4, hogy $\forall k \geq 1$-re:

$$K_3 \leq V(k;P,2) \leq K_4.$$

Megjegyezzük, hogy a (i)-(iii) feltételek azonosak a [80] cikkben találhatókkal, és a (iii)-(iv) maga után vonja a (ii) feltétel teljesülését.

A bizonyítás előtt néhány megjegyzést teszünk a tétellel kapcsolatban.

Kvantizálási kérdésekkel a 40-es évek végén kezdtek el foglalkozni. Később, Zador [112] kvantizálással foglalkozó disszertációja a címben említett legfontosabb kérdésekre választ adott. A nagyságrendi kérdést teljesebben válaszolja meg a

7.3.4. tétel (Bucklew, Wise [12]). Tegyük fel, hogy az abszolút folytonos F eloszlásnak véges $r+\varepsilon$. momentuma létezik valamilyen $\varepsilon > 0$-ra. Ekkor

$$\lim_{k\to\infty} V(k;P,r) = J_{m,r}\,\|p\|_{m/(m+r)}, \tag{7.3.6}$$

ahol a $J_{m,r}$ független a p sűrűségfüggvénytől ($\|p\|_s$ a p függvény L_s-beli függvénynormáját jelöli).

Az egydimenziós egyenletes eloszlás esetén könnyen látható, hogy $J_{1,2} = 1/12 \approx 0.0833$ és az optimális kvantizálást a k egyenlő szakaszra bontás adja. Általában is $J_{m,2} \leq m/12$ (Lengyel [61]).

Az m=2 esetben a szabályos hatszögekre és a határ mentén keletkező maradékokra vágás adja az optimális kvantizálást (Fejes Tóth [27], [28], Newman [84]) és $J_{2,2}/2 = (5/36\sqrt{3}) \approx 0.0802$.

Zador [112] alsó és felső korlátot adott meg a $J_{m,r}$ mennyiségre. Megjegyezzük, hogy azt sejtik, hogy a legjobb háló kvantizáló a legsűrűbb háló kitöltés duálisa ([16]).

A 7.3.3. tétel bizonyítása [61]. A felső korlát bizonyítása elég egyszerű: három lépésben történik. Az alsó korlát egy diszkrét geometriai eredmény felhasználásával adódik.

Először a felső korlátot bizonyítjuk, méghozzá a racionális élű téglatestekben egyenletes eloszlásokra. Legyen $C = \sum_{i=1}^{m} [a_i, b_i]$ egy olyan m-dimenziós téglatest, amelynek összes élhossza racionális szám. Legyen $b_i - a_i = p_i/q_i$ és $k = r_1 r_2 \dots r_m$, ahol p_i,

q_i, r_i pozitív egészek $\forall\, i=1,2,\ldots,m$-re. A téglatestet az "első" lapjával párhuzamosan r_1, a "másodikkal" párhuzamosan r_2, ..., az m. lapjával párhuzamosan r_m egyforma vastag téglatestre vágjuk fel. A C-n egyenletes eloszlású valószínűségi vektorváltozó komponensei függetlenek. Az így adódó k-particióra nyilván

$$(7.3.7)\quad W(k;P,2) \leqslant \frac{1}{12}\sum_{i=1}^{m}\left(\frac{b_i-a_i}{r_i}\right)^2 .$$

A k alkalmas megválasztásaival elérhető, hogy a darabolás után keletkező k téglatest egyforma kocka legyen. Legyen ugyanis $c=t.\prod_{i=1}^{m} q_i$, valamint $k=c^m.\prod_{i=1}^{m}(p_i/q_i)$ és $r_i=cp_i/q_i$ (t tetszőleges pozitív szám). Ekkor a keletkező kockák élhossza $1/c$ lesz. Az ilyen k értékekre könnyű látni a $V(k;P,2)$ korlátosságát, midőn $k\to\infty$, hiszen a (7.3.7) jobb oldala legfeljebb

$$\frac{m}{12}\,\lambda(C)^{2/m}k^{-2/m}$$ (ahol $\lambda(C)$ jelöli a C halmaz

Lebesgue mértékét). A $W(k;P,2)$ sorozat monoton csökkenő k-ban, és fenti megválasztása (t=1,2,...) már biztosítja a V-sorozat tetszőleges k-ra vonatkozó korlátosságát is.

A következő lépésben tetszőleges egyenletes eloszlásra bizonyítjuk be a korlátosságot. A módszer lényege az, hogy az eloszlás C' tartóját belefoglaljuk egy az előző lépésben használt C tégla-

testbe ((i) feltétel). Ezek után közvetlenül adódik a bizonyítás, ugyanis tetszőleges k-elemű Q halmazra

$$W(k;P_{C'},2) \leq W(Q;P_{C'},2) \leq K\, W(Q;\Lambda_C,2),$$

ahol $P_{C'}$ illetve Λ_C jelöli a C' illetve C halmazon egyenletes eloszláshoz tartozó mértéket, a K konstans pedig a C és C' Lebesgue mértékének arányával egyenlő.

Végül a feltételeknek eleget tevő C' tartójú általános eloszlás esetét visszavezetjük az előzőre. A nem centrális nyomatéknak a centrális nyomatéktól való eltéréséről szóló Steiner-tétel alkalmazásával a (iii), (iv) feltétel adja a bizonyítandó állítást. Jelölje ugyanis Q' a Q által meghatározott LTP-ban, a $P_{C'}$ mérték szerint vett feltételes várhatóérték-vektorokból képzett halmazt. Ekkor $W(Q;P,2) \leq K_2\, W(Q';P_{C'},2)$.

Az alsó korlát bizonyításához használjuk a (iv) feltétel $p(z) \geq K_1 > 0$ felét. Az első lépés az egyenletes eloszlásra vonatkozik, míg a második lépés már az általános eloszlásra.

Tekintsük az m-dimenziós euklideszi térben az $1/l$ élhosszú kocka rácsot (az l tetszőleges pozitív egész). Válasszunk ki n különböző rácspontot. Jelölje d_{ij} az i. és j. pont közötti euklideszi távolságot és legyen

$$G(n,m,\frac{1}{l}) = \min \sum_{i,j=1}^{n} d_{ij}, \quad \text{illetve}$$

$$G_1(n,m,\frac{1}{l}) = \frac{1}{2n} \min \sum_{i,j=1}^{n} d_{ij}^2,$$

ahol a minimumot az összes lehetséges n rácspontból álló kiválasztásra értjük.

Ruda ([97], 24. tétel) következő eredményére lesz szükségünk:

$$G(n,m,1) \geq \frac{m}{4(m+1)} n^{2+\frac{1}{m}},$$

ha n elég nagy. Innen a számtani és kvadratikus közép közötti egyenlőtlenségből kapjuk, hogy

$$G_1(n,m,\frac{1}{l}) \geq c_1 \cdot n^{1+\frac{2}{m}},$$

ha n elég nagy (itt a c_1 csak az m-től és l-től függő pozitív számot jelöl).

A C-n egyenletes eloszlás esetében a $W(Q;P,2)$ integrálösszeg tagjait az l növelésével egyre sűrűbb rács pontjain vett megfelelő összeggel közelítjük. A Q által definiált legkisebb távolság partícióban (LTP) a q_i ponthoz tartozó poliéderbe - a C-be eső $n(l)$ pont közül - $n_i(l)$ pont kerüljön. Ekkor a pontok alkalmas

indexelésével a (5.2.2) és (5.2.3) kifejezések alapján

$$W(Q;P,2) = \lim_{l\to\infty} \frac{1}{n(l)} \sum_{i=1}^{k} \frac{1}{2n_i(l)} \sum_{r,s=1}^{n_i(l)} d_{rs}^2 .$$

A G_1-re kapott alsó korlátot alkalmazva kapjuk a V-re vonatkozó alsó korlátot.

Általános eloszlásra a (iv) feltétel felhasználásával - a felső korlát bizonyításához hasonló módon - adódik a bizonyítás.

Clusterezési kérdések

Amennyiben nem ismerjük a P mértéket, amely szerint az n független mintaelemet megfigyeltük, akkor a megfigyelések számának növelésével azt reméljük, hogy az optimális kvantizálást optimális clusterezések sorozatával közelíthetjük. Ezt az elvárást a következő eredmények támasztják alá.

MacQueen [80] a k-közép módszer adaptív egyszerűsítéséről az r=2 esetben bebizonyította, hogy az egymás utáni lépések során keletkező clusterezési hibák 1 valószínűséggel konvergálnak.

Hartigan [41] az m=1, k=r=2 esetben igazolta, hogy az n mintaelem alapján meghatározott optimális clusterezésben - a Lloyd-módszer

mintájára R^1-ben definiálható - elválasztó pontok és a hibák sorozata mértékben konvergál az optimális kvantizálás "választópontjához", illetve hibájához. A [41] cikkben - két egyszerű hipotézis közötti döntés céljára szolgáló - valószínűséghányados statisztikára centrális határeloszlás tételt is bizonyított.

Pollard [85] az általános esetre bizonyította, hogy a k-közép eljárás során adódó cluster közepek 1 valószínűséggel konvergálnak (néhány mellékfeltétel teljesülése esetén) az optimális kvantizálási értékekhez. Pollard [86] centrális határeloszlás jellegű tételt is megadott; mindkét tétel a kvantizálási optimumhely unicitására vonatkozó kikötést tartalmaz.

Köszönetnyilvánítás

Ezúton szeretném köszönetemet nyilvánítani mindazoknak, akik a dolgozat megírásában közvetlenül vagy közvetve segítségemre voltak.

Elsősorban aspiránsvezetőmnek, Dr. Babai Lászlónak tartozom köszönettel, aki az értekezés megírását megelőzően is sok útmutatással és ötlettel segített. Beszélgetéseink során nagyon sokat tanultam tőle. A disszertáció írása során a kutatás irányait tágító kérdésekkel inspirált és értékes tanácsokkal látott el.

Köszönetemet fejezem ki társszerzőimnek, Dr. Kolosi Tamásnak és Dr. Ruda Mihálynak, mert egy érdekes alkalmazás kapcsán társadalomtudományi problémák kutatásába bevontak illetve előttem kevésbé ismert diszkrét geometriai eredményekkel megismertettek.

A kvantizálással foglalkozó irodalom megismerésében T. Cover, H. Solomon, M. Steele professzorok nyújtottak segítséget.

Köszönettel tartozom azoknak, akik az értekezés verzióit illetve részeit olvasva hasznos tanácsokkal láttak el.

Köszönet illeti munkahelyi vezetőimet, Dr. Békéssy Andrást, Dr. Demetrovics Jánost és az általuk vezetett kollektívát azért a segítőkészségért és támogatásért, ami a kutató munkát és a dolgo-

zat megírását lehetővé tette.

Végül nem feledkezhetem meg arról a segítségről sem, amelyhez a TEXTER szövegszerkesztő alkalmazása révén jutottam. Ez az eszköz tette lehetővé számomra, hogy témavezetőmmel - a nem ritkán földrésznyi távolság ellenére is - kommunikálni tudjak.

Irodalomjegyzék

1 A. V. Aho, J. E. Hopcroft, J. D. Ullman, The Design and Analysis of Computer Algorithms, Addison-Wesley, Reading, MA, 1974.

2 M. R. Anderberg, Cluster Analysis for Applications, Academic Press, New York, 1973.

3 Arató M., Fejezetek a matematikai statisztikából számítógépes alkalmazásokkal, II. Többdimenziós analízis, SZAMKI Közlemények 22, Budapest, 1979

4 Babai L., személyes közlés, 1982

5 E. S. Barnes, N. J A. Sloane, The optimal lattice quantizer in three dimensions, SIAM J. Alg. Disc. Meth., 4(1983), 30-41.

6 J. Barthelemy, An asymptotic equivalent for the number of total preorders on a finite set, Discrete Mathematics, 29 (1980), 311-313.

7 A. Benczúr, A. Krámli, J. Pergel, On the Bayesian approach

to optimal performence of page storage hierarchies. Acta Cybernetica, Tom. 3. Fasc. 2. (1977), Szeged, 79-89.

8 E. Bender, Asymptotic enumeration, SIAM Review, 16 (1974), 485-515.

9 J. L. Bentley, M. I. Shamos, Divide and conquer for linear expected time, Information Processing Letters, 7(1978), 87-91.

10 L.N. Bolshev, Cluster analysis, Bull. ISI, 23(1969), 411-425.

11 K. Q. Brown, Voronoi diagrams from convex hulls, Information Processing Letters, 9(1979), 223-228.

12 J. Bucklew, G. Wise, Multidimensional asymptotic quantization theory with r^{th} power distortion measure, IEEE Trans. Inform. Theory, IT-28(1982), 239-247.

13 P. Buneman, A note on the metric properties of trees, Journal of the Combinatorial Theory (B), 17(1974), 48-50.

14 D. Cheriton, R. E. Tarjan, Finding minimum spanning trees, SIAM J. Comput., 5(1976), 724-742.

15 L. Comtet, Advanced Combinarorics, Reidel, Dordrecht-Boston, 1974

16 J. H. Conway, N. J. A. Sloane, Voronoi regions of lattices, second moments of polytopes, and quantization, IEEE Trans. Inform. Theory, IT-28(1982), 211-226.

17 R. M. Cormack, A review of classification, Journal of the Royal Statistical Society, Series A, 134 (1971) 321-367.

18 T. M. Cover, Geometrical and statistical properties of systems of linear inequalitites with applications in pattern recognition, IEEE Trans. Electronic Computers, EC-14(1965), 326-334.

19 D. R. Cox, Note on grouping, J. Amer. Statist. Ass., 52(1957) 543-547.

20 E. W. Dijkstra, A note on two problems in connections with graphs, Numer. Math., 1(1959), 269-271.

21 D. Dobkin, R. J. Lipton, Multidimensional searching problems, SIAM J. Comput., 5(1976), 181-186.

22 P. Doubilet, G.-C. Rota, R. Stanley, On the foundations of combinatorial theory, VI: The idea of generating function, Proc. 6th Berkeley Symp. on Math. Stat. and Prob., vol II, Univ. Calif. Press, Berkeley and Los Angeles, 1970, 267-318.

23 J. C. Dunn, Well-separated clusters and optimal fuzzy partitions, Journal of Cybernetics, 4(1974), 95-104.

24 P. Erdős, J. Spencer, Probabilistic Methods in Combinatorics, Academic Press, New York, Akadémiai Kiadó, Budapest, 1974.

25 B. S. Everitt, Cluster Analysis, Heinemann Educational Books, London, 1974.

26 B. S. Everitt, Unresolved problems in cluster analysis, Biometrics, 35(1979), 169-181.

27 L. Fejes Tóth, Lagerungen in der Ebene auf der Kugel und im Raum, Springer-Verlag, 1953.

28 L. Fejes Tóth, Sur la representation d'une population infinie par un nombre fini d'elements, Acta Math. Acad. Scient. Hung., 10(1959), 299-304.

29 W. D. Fisher, On grouping for maximum homogeneity, J. Amer. Statist. Ass., 53(1968), 789-798.

30 L. Fisher, J. W. Van Ness, Admissible clustering procedures, Biometrika, 58(1971), 91-104.

31 P. E. Fleischer, Sufficient conditions for achieving minimum distortion in quantizer, IEEE Int. Convention Record, 12(1964), 104-111.

32 M. Garey, D. Johnson, Computers and Intractability, A Guide to the Theory of NP-Completeness, Freeman, San Francisco, 1978

33 R. E. Gomory, T. C. Hu, Multi-terminal network flows, SIAM J. Appl. Math., 9(1961), 551-570.

34 J. C. Gower, G. J. S. Ross, Minimum spanning trees and single linkage cluster analysis, Appl. Stat., 18(1969), 54-64.

35 R. L. Graham, Problem 85-5 EATCS, Euclidean minimum spanning trees, Journal of Algorithm, 6(1985), 285-286.

36 R. M. Gray, E. D. Karnin, Multiple local optima in vector quantizers, IEEE Trans. Inform. Theory, IT-28(1982), 256-261.

37 M. Grötschel, L. Lovász, A. Schrijver, The ellipsoid method and its consequences in combinatorial optimization, Combinatorica, 1(1981), 169-197.

38 A. Hardy, J. R. Rasson, L. Szary, P. Schindler, An optimal clustering method based on the Rasson's criterion and resulting from a new approach, in Pattern Recognition Theory and Applications, (J. Kittler, K. S. Fu, L. F. Pau Eds.), 63-71, Reidel, Dordrecht, 1982.

39 L. H. Harper, Stirling behavior is asymptotically normal, Ann. Math. Statist., 38(1967), 410-414.

40 J. A. Hartigan, Clustering Algorithms, Wiley, New York, 1975.

41 J. A. Hartigan, Asymptotic distributions for clustering criteria, Annals of Statistics, 6(1978), 117-131.

42 J. A. Hartigan, Distribution problems in clusterting, in Classification and Clustering, ed. J. Van Ryzin, Academic

Press, New York, 1977, 45-71.

43 A. Heppes, P. Szüsz, Bemerkung zu einer Arbeit von L. Fejes Tóth, El. Math., 15(1960), 134-136.

44 D. Hochbaum, D. Shmoys, Powers of graphs: A powerful approximation technique for bottleneck problems, Proc. 16th Annual ACM Symposium on Theory of Computing, Washington, 1984, 324-333.

45 L. C. Hsu, Note on an asymptotic expansion of the n-th difference of zero, Ann. Math. Statist., 19(1948), 273-277.

46 N. Jardine, R. Sibson, Mathematical Taxonomy, Wiley, London, 1971.

47 R. A. Jarvis, On the identification of the convex hull of a finite set of points in the plane, Information Processing Letters, 2(1973), 18-21.

48 M. E. Jansen, J. G. Bethlehem, An application of cluster analysis to geographical classification, Centraal Bureau voor de Statistiek, Technical Report, 9613-79-M1

49 R. E. Jensen, A dynamic programming algorithm for cluster analysis, Operational Research, 17(1968), 1034-1056.

50 S. C. Johnson, Hierarchical clustering schemes, Psychometrika, 32(1967), 241-254.

51 J. C. Kieffer, Exponential rate of convergence for Lloyd's method I, IEEE Trans. Inform. Theory, IT-28(1982), 205-210.

52 D. G. Kirkpatrick, R. Seidel, The ultimate planar convex hull algorithm?, Proceedings of the 20th Allerton Conference on Communication, Control and Computing, Monticello, Illinois (*1982*), 35-42.

53 D. E. Knuth, The Art of Computer Programming, Vol 3, Sorting and Searching, Addison-Wesley, Reading, Mass.

54 Kolosi T., Lengyel T., Az egyenlőtlenség-tudat vizsgálata cluster elemzéssel, Szociológia, 1979, 35-48.

55 J. Komlós, Linear verification for spanning trees, Combinatorica, 5(1985), 57-65.

56 M. Krivánek, J. Morávek, Clustering to minimize the sum of

volumes of convex hulls of clusters is NP-complete, Proc. of the FCT'85, 234-241.

57 J. B. Kruskal, Jr., On the shortest spanning subtree of a graph and the travelling salesman problem, Problem. Amer. Math. Soc., 7(1956) 48-50.

58 G. N. Lance, W. T. Williams, A general theory of classificatory sorting strategies, I. Hierarchical systems, Computer J., 9(1967), 373-380.

59 G. N. Lance, W. T. Williams, A general theory of classificatory sorting strategies, II. Clustering systems, Computer J., 10(1967), 271-277.

60 E. Lawler, Kombinatorikus optimalizálás: hálózatok és matroidok, Műszaki Könyvkiadó, Budapest, 1982

61 Lengyel T., Legkisebb négyzetes eljárások a cluster analízisben, egyetemi doktori értekezés, Eötvös Loránd Tudományegyetem, Budapest, 1979.

62 T. Lengyel, M. Ruda, On the asymptotic behaviour of the loss function of the minimum within-class variance

clustering procedure, 11th European Meeting of Statisticians, Oslo, 1978, 184-185.

63 T. Lengyel, A note on the number of clusterings, 12th European Meeting of Statisticians, Varna, 1979, 148.

64 Lengyel T., A kanonikus korrelációanalízis és néhány kapcsolódó probléma, Alkalmazott Matematikai Lapok, 5(1979), 385-393.

65 T. Lengyel, On the number of agglomerative clustering hierarchies, COMPSTAT'82, vol. 2, (H. Cassinus, P. Ettinger, J. R. Mathieu, Eds.), 177-178, Physica-Verlag, Wien, 1982

66 T. Lengyel, An application of canonical correlation analysis to predict coronaria thrombosis, MTA SzTAKI Közlemények, 28/1982, 35-43.

67 Lengyel T., Kanonikus korrelációanalízis, a "Többváltozós statisztikai módszerek" jegyzet 8. fejezete, Bolyai János Matematikai Társulat, szerk.: Rejtő L., Budapest, 1983/84, 251-271.

68 T. Lengyel, On a recurrence involving Stirling numbers,

European Journal of Combinatorics 5(1984), 313-321.

69 T. Lengyel, On the number of dendrograms and other tree structures, in Report on the Joint Stanford-Hungarian Workshop in Multidimensional Analysis, Information Theory and Asymptotic Methods, (I. Olkin, Ed.), Stanford University, 1984, 33-36.

70 Lengyel T., Geometriai problémák algoritmikus kérdéseiről, MTA SzTAKI Working paper, IV/70, 30.o., Budapest, 1985.

71 Lengyel T., A clusterezés és a kvantizálás közös kérdéseiről, MTA SzTAKI Working paper, IV/71, 28.o., Budapest, 1985.

72 Lengyel T., Kanonikus korrelációanalízis, megjelenik a Többváltozós statisztikai módszerek c. könyvben, Műszaki Könyvkiadó, Budapest, 1986, szerk. Székely J. G., Móri T.

73 T. Lengyel, On the computational complexity of some geometrical and clustering problems, megjelenik a Proc. of DIANA II on Recent Results of Discriminant Analysis, Cluster Analysis, Factor Analysis, Statistical Problems of Classification and Other Methods in Multivariate Data

Analysis kötetben, 1986

74 T. Lengyel, A convergence criterion for recurrent sequences with application to the partition lattice, előkészületben

75 R. F. Ling, On theory and construction of k-clusters, Computer J., 15(1972), 326-332.

76 R. F. Ling, A probability theory of cluster analysis, J. Amer. Statist. Ass., 68(1973), 156-169.

77 S. P. Lloyd, Least squares quantization in PCM, IEEE Trans. Inform. Theory, IT-28(1982), 129-137.

78 L. Lovász, Combinatorial Problems and Exercises, Akadémiai Kiadó, Budapest, 1979

79 Lovász L., Gács P., Algoritmusok, Műszaki Könyvkiadó, Budapest, 1978

80 J. MacQueen, Some methods for classification and analysis of multivariate observations, Proc. 5th Berkeley Symp. Math. Statist. Prob. 1, 281-297.

81 D. W. Matula, k-components, clusters, and slicing in graphs, SIAM J. Appl. Math., 22(1972), 459-480.

82 D. W. Matula, Graph theoretic techniques for cluster analysis algorithms, in Classification and Clustering, ed. J. Van Ryzin, Academic Press, New York, 1977, 95-129.

83 F. Murtagh, Counting denrograms, a survey, Discrete Applied Mathematics 7(1984), 191-199.

84 D. J. Newman, The hexagon theorem, IEEE Trans. Inform. Theory, IT-28(1982), 137-139.

85 D. Pollard, Strong consistency of k-means clustering, Annals of Statistics, 9(1981), 135-140.

86 D. Pollard, A central limit theorem for k-means clustering, Annals of Probability, 10(1982), 919-926.

87 F. P. Preparata, Editor, Steps into Computational Geometry, Rep. R-760, Coordinated Science Laboratory, Applied Computation Theory Group, Univesity of Illinois, Urbana 1977.

88 F. P. Preparata, S. J. Hong, Convex hulls of finite sets

of points in two and three dimensions, CACM 20 (2) (1977) 87-93.

89 R. C. Prim, Shortest connection networks and some generalizations, Bell System Tech. J., 36(1957), 1389-1401.

90 P. Racskó, Imitácionnaja model' roszta gyereva. Posztrojé-nyije modeli, Zsurnal Obsej Biologii, Tom XXXIX (1978), Moszkva, 563-571.

91 H. Raynaud, Sur l'enveloppe convexe des nuages de points aleatoires dans R^n. I, J. Appl. Prob. 7 (1970), 35-48.

92 A. Rényi, Some remarks on the theory of trees, Mat. Kut. Int. Közl., 4(1959), 73-85.

93 A. Rényi, R. Sulanke, Über die konvexe Hülle von n rotationsymmetrisch verteilten Punkten I, Z. Wahrscheinlichkeitsth., 2(1963), 75-84.

94 A. Rényi, R. Sulanke, Über die konvexe Hülle von n rotationsymmetrisch verteilten Punkten II, Z. Wahrscheinlichkeitsth., 3(1964), 138-147.

95 C. Rogers, The probability that two samples in the plane will have disjoint convex hulls, J. Appl. Prob., 15(1978) 790-802.

96 G.-C. Rota, On the foundations of combinatorial theory, I. Theory of Möbius functions, Z. Warscheinlichkeitstheorie und Verw. Gebiete, 2(1964), 340-368.

97 Ruda M., Alakzatok tömörségével kapcsolatos vizsgálatok, MTA III. Osztály Közleményei, 23(1974), 203-237.

98 M. Schader, Hierarchical analysis: classification with ordinal object dissimilarities", Metrika, 27(1980), 127-132.

99 D. Schröder, Vier combinatorische Probleme, Z. für M. Phys, 15(1870), 361-76.

100 M. I. Shamos, Geometric complexity, Proc. 7th Annual ACM Symposium on Theory of Computing, 1975, 224-233.

101 M. I. Shamos, D. J. Hoey, Closest-point problems, Proc. 16th Annual IEEE Symposium on Foundations of Computer Science, 1975, 151-162.

102 P. H. A. Sneath, R. R. Sokal, Numerical Taxonomy, Freeman, San Francisco, 1973.

103 J. M. Steele, Growth rates of minimal spanning trees of multivariate samples, Technical Report No. 162, 1980, Stanford University

104 H. Steinhaus, Mathematical Snapshots, Oxford University Press, 1969, (Matematikai kaleidoszkóp, Műszaki Könyvkiadó, Budapest, 1985)

105 G. F. Toussaint, Pattern recognition and geometrical complexity, in Proc. 5th Int. Conf. on Pattern Recognition, Miami Beach, 1980, 1324-1346.

106 A. V. Trushkin, Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions, IEEE Trans. Inform. Theory, IT-28(1982),187-198.

107 Tusnády G., Mátrixok szinguláris felbontása, Matematikai Lapok, 5(1979), 375-384.

108 Tusnády G., Keverékek felbontása, Mat. Lapok, 20(1978-1982), 59-67.

109 A. C. Yao, On constructing minimum spanning trees in k-dimensional spaces and related problems, SIAM J. Comput., 11(1982), 721-736.

110 J. W. Van Ness, Admissible clustering procedures, Biometrika, 60(1973), 422-424.

111 R. O. Winder, Partitions of N-space by hyperplanes, J. SIAM Appl. Math., 14(1966), 811-818.

112 P. Zador, Development and evaluation of procedures for quantizing multivariate distributions, Ph. D. disszertáció, Stanford University, 1963.

1986-BAN EDDIG MEGJELENTEK:

179/1986 Terlaky Tamás: Egy véges criss-cross módszer és alkalmazásai

180/1986 K.N. Čimev: Separable sets of arguments of functions

181/1986 Renner Gábor: Kör approximációja a számitógépes geometriai tervezésben

182/1986 Proceedings of the Joint Bulgarian-Hungarian Workshop on "Mathematical Cybernetics and Data Processing" Scientific Station of Sofia University, Giulecica /Bulgaria/, May 6-10, 1985 /Editors: J. Denev, B. Uhrin/ Vol I

183/1986 Proceedings of the Joint Bulgarian-Hungarian Workshop on "Mathematical Cybernetics and Data Processing" Scientific Station of Sofia University, Giulečica /Bulgaria/, May 6-10, 1985 /Editors: J. Denev, B. Uhrin/ Vol II

184/1986 HO THUAN: Contribution to the theory of relational databases

185/1986 Proceedings of the 4th International Meeting of Young Computer Scientists IMICS'86 /Smolenice, 1986/ /Editors: J. Demetrovics, J. Kelemen/

186/1986 PUBLIKÁCIÓK - PUBLICATIONS 1985 Szerkesztette: Petróczy Judit

187/1986 Proceedings of the Winter School on Conceptual modelling /Visegrád, 27-30 January, 1986/ /Editors: E. Knuth, A. Márkus/

1985-BEN MEGJELENTEK:

166/1985 Radó Péter: Információs rendszerek számitógépes tervezése

167/1985 Studies in Applied Stochastic Programming I. Szerkesztette: Prékopa András /utánnyomás/

168/1985 Böszörményi László - Kovács László - Martos Balázs - Szabó Miklós: LILIPUTH

169/1985 Horváth Mátyás: Alkatrészgyártási folyamatok automatizált tervezése

170/1985 Márkus Gábor: Algoritmus mátrix alapu logaritmus kiszámitására kriptográfiai alkalmazásokkal

171/1985 Tamás Várady: Integration of free-form surfaces into a volumetric modeller

172/1985 Reviczky János: A számitógépes grafika terület-kitöltő algoritmusai

173/1985 Kacsukné Bruckner Livia: Mozgáspálya generálás bonyolult geometriáju felületek 2 1/2D-s NC megmunkálásához

174/1985 Bolla Marianna: Mátrixok spektrálfelbontásának és szinguláris felbontásának módszerei

175/1985 Hannák László, Radó Péter: Adatmodellek, adatbázis-filozófiák

176/1985 Számitógépes képfeldolgozási és alakfelismerési kutatók találkozója.
Szerkesztette: Csetverikov Dmitirj,
Főglein János és Solt Péter

177/1985 Gyárfás András: Problems from the world surrounding perfect graphs

178/1985 PUBLIKÁCIÓK'84
Szerkesztette: Petróczy Judit