KÖNYVEKRŐL

ANDERBERG, M. R.: *Cluster analysis for applications.* New York – London, 1973. Academic Press, 359 p.

Az olvasóban a könyv kiadásának dátumát olvasva felvetődhet a kérdés, hogy a SZIGMA cluster analízisnek szentelt különszámában miért ezt a már 5 éve megjelent könyvet ismertetjük.

A könyv választását a következők indokolják: A matematikai-statisztikai cluster analízis területén a 70-es évek elejéig elért eredményeket Anderberg munkája nemcsak összefoglalja, hanem nagyon alaposan és didaktikusan rendszerezi is. A könyv a cluster elemzés hazai művelőinek szinte egyöntetű véleménye szerint alapvető fontosságú.

A később megjelent cluster analízis témájú könyvek (pl. *J. A. Hartigan:* Clustering Algorithms, John Wiley and Sons, 1975) már csak egy-egy fontosabb részterülettel foglalkoznak behatóan.

Ugyanakkor a hetvenes évek elejétől igen gyors fejlődésnek induló információtudományi cluster elemzésről összefoglaló jellegű könyv még nem jelent meg, csak cikkek és szélesebb területet is felölelő tanulmányok (pl. *K. Sparck Jones:* Automatic Indexing '74. University of Cambridge, 1974).

Anderberg könyve 10 fejezetet és 8 függeléket tartalmaz.

Az első két fejezet általánosan ismerteti a cluster analízis témáját, felhasználási területeit és kapcsolatait más tudományágakkal. A cluster analízis alapfeladata az, hogy objektumok bonyolult rendszerének struktúráját feltárja, az objektumokat – előzetes tapasztalatok nélkül – kizárólag jellemzőiből adódó kapcsolataik alapján úgy csoportosítsa, hogy az egymáshoz hasonló objektumok egy csoportba (clusterbe) az egymáshoz kevéssé hasonló objektumok különböző csoportokba kerüljenek. Tágabb értelemben a cluster elemzés feladata a jellemzők csoportosítása is.

A szerző élesen meghúzza a határvonalat az osztályzási eljárások és a cluster elemzés módszerei között azzal, hogy a klasszifikálásnál a struktúra már ismert, és csak egy vagy több kategóriát kell besorolni, míg a cluster elemzés elsődleges célja egy „természetes" struktúra megalkotása.

A struktúra feltárás problémája nagy jelentőségű a biológiában, földtudományban, orvostudományban, társadalomtudományban, mérnöki tudományokban, információtudományban és végül, de nem utolsó sorban az operációkutatásban. Ebből is látszik, hogy a cluster elemzés módszereinek felhasználási területe milyen széleskörű.

A harmadik fejezet az objektumok jellemzőinek (a továbbiakban változóknak) csoportosításával foglalkozik. Az értelmezési tartomány számossága szerint megkülönböztet folytonos, diszkrét és – a diszkréten belül még – bináris változókat. Különösen jelentős a változók csoportosítása a mérési skála alapján. Jelöljön A és B két tetszőleges objektumot, x_A, ill. x_B az X változónak az A, ill. a B objektumra jellemző „értékét".

Ha X normális skálájú változó, akkor csak azt tudjuk, hogy $x_A = x_B$ vagy $x_A \neq x_B$. Ha X ordinális skálájú változó, akkor azt tudjuk, hogy $x_A = x_B$ vagy $x_A > x_B$ vagy $x_A < x_B$. Ha X intervallum skálájú változó, akkor ismerjük $x_A - x_B$ értékét. Ha X hányados skálájú változó, akkor ismerjük $x_A - x_B$ értékét. Ha X hányados skálájú változó, akkor ismerjük x_A/x_B értékét is. Ebben a fejezetben található skála konverziók részletes ismertetése is.

A negyedik fejezet a változók közötti asszociációs mértékekkel foglalkozik. Külön tárgyalja a hányados és intervallum skálájú (kvantitatív jellegű) változók, valamint a nominális és ordinális skálájú (kvalitatív jellegű) változók között használatos mértékeket. Részletesen elemzi és összehasonlítja a bináris változók asszociációs mértékeit (pl. *Dice, Tanimoto* mértéke).

Az ötödik fejezet az objektumok közötti asszociációs mértékeket tárgyalja. A kvantitatív jellegű változók esetén a metrikán alapuló mértékek fontosságát emeli ki. A kvalitatív jellegű változók esetén valószínűségi jellegű mértékek alkalmazása is elterjedt, ezért ezeket is részletesen tárgyalja. Külön foglalkozik a szerző a bináris változók esetén használatos mértékekkel. A fejezet végén értékes tanácsokat ad arra az esetre, ha az összehasonlítandó objektumok jellemzői között kvalitatív és kvantitatív jellegű változók egyaránt szerepelnek.

A következő két fejezet a cluster eljárásokkal foglalkozik. A hierarchikus cluster eljárások (hatodik fejezet) zömmel az asszociációs mérték segítségével képzett hasonlósági mátrix felhasználásán alapulnak, amely i-edik sorának j-edik eleme megadja az i-edik és a j-edik objektum, ill. változó közötti asszociációs mértéket. Az eljárások célja olyan cluster hierarchia (fa) megalkotása, amelynél két clusternek vagy nincs közös eleme, vagy az egyik tartalmazza a másikat. (Általában clusternek tekintik az egyes objektumokat és a teljes vizsgálandó objektum halmazt is.) A fejezet elsősorban az agglomeratív jellegű hierarchikus cluster eljárásokat tárgyalja (az eljárás során mindig kisebb clusterek egyesítéséből képezzük a nagyobb clustert). A fordított irányú (particionáláson alapuló) eljárások közül csak néhányat említ meg. Az agglomeratív jellegű eljárásokat aszerint csoportosítja, hogy a számítógép központi memóriájában a kiindulási adatokat vagy a hasonlósági mátrixot tárolják. Részletesen tárgyalja a széleskörűen alkalmazott módszereket (legközelebbi szomszéd, legtávolabbi szomszéd, *Ward*, *Wishart* módszerét és a középpont szerint osztályozó eljárásokat).

A hetedik fejezet az ún. nem hierarchikus cluster eljárások ismertetését tartalmazza, itt nem egy cluster hierarchia megalkotása a cél, hanem az objektumok particionálása előre megadott számú vagy az eljárás közben adódó számú csoportba (clusterbe). A fejezet elején a szerző részletesen elemzi a kialakítandó clusterek magjainak meghatározására szolgáló heurisztikus algoritmusokat. Ezt követi *Forgy*, *MacQueen*, és *Wishart* eljárásának, valamint az ISODATA módszerének ismertetése.

A nyolcadik és kilencedik fejezet azokat a technikákat és stratégiákat tartalmazza, amelyek alkalmazásával a cluster analízis hatékonysága növelhető. A hierarchikus osztályozás segédeszközeinek vázlatos ismertetése mellett részletesen tárgyalja a szekvenciális jellegű cluster eljárásokat, amelyek zömmel heurisztikus módszerek, de nagy méretű problémák megoldásához szinte nélkülözhetetlenek. Vázolja a több cluster eljárás párhuzamos alkalmazásából eredő előnyöket és esetleges hátrányokat (pl. költség) is.

Vizsgálja a cluster analízis felhasználási területeit a matematikai statisztikán belül, és a külső kritériumokat is figyelembe vevő cluster eljárások megalkotásának általános irányelveit.

A tizedik fejezet a cluster eljárások összehasonlításának módszereivel foglalkozik. Kiemeli, hogy általában nem található legjobb eljárás egy adott feladat megoldására. Kísérletet tesz a hierarchikus eljárások és a particionáláson alapuló eljárások közötti hasonlóság mérésére. Felsorolja a problémák legfontosabb jellemzőit (pl. objektumok, változók száma, típusa, változók súlyozása, clusterrel szemben támasztott követelmények) és a megoldási módszerek fő paramétereit (pl. az eredmények struktúrája, gépidő igény, memória igény, a kiindulási állapot megválasztásának hatása a clusterek struktúrájára).

Az A függelék elméleti jellegű. A nominális skálájú változók közötti asszociációs mérték megalkotásával foglalkozik. A könyv B, C, D, E, F és G függeléke FORTAN nyelven írt számítógépes programokat tartalmaz, amelyek géptől függetlenek és valóban egyszerűen felhasználhatók. A B függelékben a skála konverziók elvégzését a C és D függelékben az asszociációs mértékek meghatározását elősegítő programok szerepelnek. Az E és F függelék a leggyakrabban használt cluster eljárások programjait, a G függelék az eljárások eredményeinek interpretálását elősegítő programokat tartalmazza. A H függelék a számítógépes programok kapcsolatait mutatja be.

A könyvet több mint 150 tételes referencia lista és tárgymutató zárja.

FUTÓ PÉTER

RUBIN, J. – FRIEMAN, H. P.: (Cluster analízis és taxonomikus rendszer az adatok csoportosítására és osztályozására) *A Cluster Analisis and Taxonomy System for Grouping and Classifying Data.* IBM Contributed program library. August 1967.

Az IBM gondozásában 1967-ben megjelent könyv első részében betekintést nyújt a cluster analízis problémafelvetésébe. Ismertet egy, a súlypontok módszerén alapuló nem hierarchikus cluster modellt, illetve az adatok hasonlósági mértékének vizsgálatán alapuló osztályozási rendszer elméletét és a módszerek gyakorlati meg-

valósításának elvét. A második részben a mellékelt programcsomag használatát ismerteti. A függelékekben a gyakorlati tapasztalatokat összegezi és javaslatokat ad a különböző alkalmazási területek felhasználói részére.

Ebben az ismertetőben a cluster modell elméletének és gyakorlati megvalósításának elvét írjuk le.

Matematikai leírás

A felhasznált cluster definíció megköveteli, hogy a clusterek diszjunktak legyenek és minden elem tartozzék valamely clusterbe. Az analízis eredményétől megkívánjuk, hogy optimális legyen, vagy ha nem lehetséges az optimalizáció, akkor erről felvilágosítást adjon. Ezért célszerű egy $\mathcal{F}$ cluster függvény bevezetése, amely jellemző minden egyes felbontásra.

A cluster analízis problémája az $\mathcal{F}$ függvény maximumának/minimumának meghatározására.

A vizsgált elemek legyenek $\xi_1, \xi_2, \ldots, \xi_n$ p-dimenziós valószínűségi változók, amelyeknek létezik közös sűrűségfüggvényük. Tegyük fel, hogy előre adott a clusterek száma és a függvény olyan egyszerű szerkezetű, hogy $f(x) = k$, ha $x \in G_i$ és $f(x) = 0$, ha $x \notin G_i$ $(i = 1, 2, \ldots g)$, ahol G_i az i-edik cluster és g a clusterek száma.

A mérések eredményét mátrix alakban ábrázoljuk. X mátrix sorai $P_i = (x_{i,1}, x_{i,2}, \ldots x_{i,p})$ $(i = 1,2, \ldots, n)$ és P_i legyen a p dimenziós euklideszi-tér egy pontja. Az általánosság megszorítása nélkül feltehető, hogy az n pont tömegközéppontja az origó. Ekkor az n pont totális szórás mátrixa $T = X^T X = \sum_{i=1}^{n} P_i^T P_i$. ($T$ a transzponálás jele). Jelölje $n_1, n_2, \ldots, n_g$ az objektumok számát az egyes csoportokban úgy, hogy $n_1 + n_2 + \ldots + n_g = n$. Ezután definiáljuk a csoportok szórásmátrixát a C_k tömegközépponttal

$$W_k = \sum_{l=1}^{n_k} (P_{kl} - C_k)^T (P_{kl} - C_k)$$

A csoportok összegezésében a szórásmátrix legyen $W = \sum_{k=1}^{g} W_k$ és a csoportok közötti szórásmátrixot definiálja $B = \sum_{k=1}^{g} n_k C_k^T C_k$. Ezek után vizsgáljuk az (1) $T = W + B$ mátrix egyenletet. Mivel a feldolgozás kezdetén az adatok mértékéről és méretéről semmilyen kikötést nem tettünk, a felbontást jellemző clusterfüggvényt (továbbiakban kritériumot) úgy kell választani, hogy az invariáns legyen a pontok közötti nem szinguláris lineáris transzformációkkal szemben.

Belátható, hogy $p = 1$ esetében az (1) egyenletből kapott $\frac{T}{W} = 1 + \frac{B}{W}$ egyenletben $\frac{B}{W}$ a fenti feltételnek megfelelő kritériumot ad.

$p > 1$ esetben ha $p \leqq n - g$ akkor W pozitív definit szimmetrikus mátrix, így létezik inverze. Ekkor képezzük a Mahalanobis-távolságot a következő módon

$$md_{ij} = (P_i - P_j)\, W^{-1} (P_i - P_j)^T$$
$$(i, j = 1,2, \ldots, n)$$

Ez a távolság invariáns a pontok közötti nem szinguláris lineáris transzformációkkal szemben. Így kaptunk egy, a felbontáshoz rendelhető megfelelő kritériumot. Ebből adódik a módszer heurisztikus jellege; ha ismerjük a felbontást, megfelelő metrikát tudunk meghatározni, ha ismerjük a megfelelő metrikát, segítségével eljuthatunk az optimális felbontáshoz. Feltételezzük, hogy a számítások során eljutunk a legjobb felbontásig.

Az eljárást a következőképpen fogalmazzuk meg. Az (1) egyenletből a determinánsok szorzástételének felhasználásával képezzük a $\frac{|T|}{|W|} = |I + W^{-1} B|$ egyenletet. A bal oldal egy skalárfüggvény, ezt a $\frac{|T|}{|W|}$-t maximalizáljuk, elfogadva azt a felbontást, amelyre a $\frac{|T|}{|W|}$ a legnagyobb, majd az ehhez tartozó W-t használjuk a Mahalanobis-különbség meghatározására.

Mivel módszereink heurisztikusak, szükséges más kritériumok meghatározása is az ellenőrzés érdekében. Vezessük be a $\mathrm{Tr}A$ jelölést, amely a mátrixhoz olyan konstansot rendel, mely jellemzi a mátrix elemeinek egymáshoz való viszonyát. (Bizonyos feltételek mellett a $\mathrm{Tr}A = \Sigma a_{ii}$.) Vezessük be az (1) egyenlet alapján a Hotteling-féle trace-kritériumot. Belátható, hogy a $\mathrm{Tr}\,(W^{-1}B)$ megfelelő kritériumot ad, ahol $\mathrm{Tr}\,(W^{-1}B) = \Sigma \lambda_i$, ahol λ_i-k a $W^{-1}B$ sajátértékei, azaz a $|B - \lambda W| = 0$ egyenlet megoldásai. Ezen λ_i-k segítségével kifejezhetjük a $\frac{|T|}{|W|}$ hányadost is, melyre igaz $\frac{|T|}{|W|} = \prod_i (1 + \lambda_i)$.

A feldolgozás során a $\log(|T|/|W|)$-t fogjuk vizsgálni. A különböző feldolgozások során kitűnt, hogy nem dönthetünk egyik kritérium javára sem. A választást

az éppen adott feladat határozza meg, illetve az ismertetett programcsomag automatikusan választja ki a legmegfelelőbbet.

A cluster analízis gyakorlati megvalósítása

A clusterizáláshoz javasolt kritériumok mindegyike felhasználja az összes lehetséges g csoportra való felbontást. Ezek száma még alacsony elemszám esetén is igen nagy, ezért fontos olyan módszer kidolgozása, amely ha nem is az összes, de legalább annyi lehetőséget végigvizsgál, amelyekből már megfelelő következtetést levonhatunk. Használjuk a már sok esetben bevált, ún. hegymászó eljárást.

Mivel ez a rendszer valamilyen kezdeti felbontást feltételez az objektumok halmazán, hatékonysága nagyon függ a kiindulás jóságától.

Először legyen a kezdeti felbontás valamilyen véletlen felbontás, majd alkalmazzuk az ún. „gyorsított menet"-et. Válaszszuk ki a kezdeti felbontás valamely csoportját és minden elemét vigyük az elemhez legközelebb levő csoport súlypontjának közelébe (a mérték legyen az éppen adott felbontáshoz tartozó Mahalanobis-mérték, vagy az általános eukledeszi mérték). Minden esetben számoljuk ki az egyszeri páronkénti kritériumot ($\mathrm{Tr}\, W_k = \frac{1}{n_k}\left(\sum_{l,m=1}^{n_k}(P_{lk}-P_{mk})(P_{lk}-P_{mk})\right)$, $\mathrm{Tr}\, W = \sum_{k=1}^{g}\mathrm{Tr}\, W_k$ és ha ez kisebb, mint az előző esetben, hagyjuk az elemet az új helyén.

Másik jól használható eljárás a kezdeti felbontáshoz az ún. újrajelölési mód. Jelöljön valamilyen kiinduló felbontást Q és minden egyes objektumot abba a Q-hoz tartozó csoportba soroljunk, amelynek súlypontjához a legközelebb van (mértékül az eukledeszi mértéket használjuk). Vizsgáljuk az új felosztáshoz tartozó valamelyik kritériumot. Az eljárást addig folytatjuk, amíg vagy nem kaptunk új felbontást, vagy az új felbontáshoz gyengébb érték tartozik.

A hegymászó eljárás. Tegyük fel, hogy valamilyen módszerrel adott egy kezdeti felbontás. Vegyük az egyik objektumot és mozgassuk el csoportról csoportra. Ha egy mozgatás során a választott kritérium jobb értéket szolgáltat, akkor az objektumot az új csoportba tartozónak tekintjük és folytatjuk a mozgatását. Ha az új felbontásnál a kritérium ugyanazt az értéket szolgáltatja, meghagyjuk az eredeti felbontást. Ha a kezdeti felbontásban az adott csoportok száma kisebb, mint az a szám, amit a felhasználó megkívánt, akkor az üres halmazba való mozgatást is megvizsgáljuk. Így az egyes objektumok elvándorolnak a „jobb" csoport felé. Ezután vesszük a következő objektumot és így tovább. Az összes objektum egyszeri tologatása lesz egy hegymászó menet. Mivel véges halmaz mozgatásainak száma is véges, véges ismételt hegymászó eljárás után eljutunk a legjobb felbontásig.

Mivel a vázolt eljárás igen sok számítást igényel a cluster-analízis eredményét egyszeri hegymászó menet után kapjuk. A gyakorlati feladatoknál ez általában elegendő, de természetesen mód van a program paraméterezésének segítségével további pontos számolásokra. A hatékonyságot inkább a kezdeti felbontás jobb megadásával, illetve jobb adatelőkészítéssel lehet fokozni.

Az IBM felhasználói programkönyvtárában közreadott rendszer IBM alapú, így (ESZR gépen is) 128K, 256K, illetve 512K vagy nagyobb központi memóriájú gépen futtatható. A terjesztett mágnesszalagon az IBM OS konvekcióknak megfelelő object modul, illetve az eredeti forrásprogram is rögzítve van. A program Assembler és FORTRAN nyelven megírt modulokból áll. A program szolgálatatásai igen sokrétűek. Az adatokat a felhasználó által megadott FORTRAN formátum szerint kell megadni. A kezdeti felbontás is megadható, de kívánságra a program véletlen, vagy gyorsított, vagy újrajelöléses módon is képezhet kiinduló felbontást. Szabályozni lehet, hogy a hegymászó eljárást hányszor ismételje meg és azt, hogy a feldolgozás során a Mahalanobis ($\mathrm{Tr}\,(W^{-1}B)$), a Wilks – Lambda $\left(\log\frac{|T|}{|W|}\right)$ kritériumot, vagy ezek közelítését használja. Végül mód van arra, hogy a feldolgozást megszakítva, más időpontban a korábbi eredményeket felhasználva pontosítsuk az eredményeket.

A programrendszer a KSH Számítástechnikai Igazgatóságán instalálva van, s ott használatáról gyakorlati tapasztalatokkal is rendelkeznek.

CSICSMAN JÓZSEF