S. Benedikt Vera — Vári Anna

Egyes clusteranalízis-eljárások és gazdasági alkalmazásuk

Bonyolult rendszerekkel kapcsolatos modellezési feladatok megoldásában matematikai, számítástechnikai és humán oldalról hasonló problémák (a feladat egyszerűsítésére, az implicit összefüggések feltárására stb.) merülnek fel. A cikk első része e problémákat általában tárgyalja, majd ismerteti az ezek megoldását segítő statisztikai osztályozó eljárásokat. Végezetül bemutatjuk az eljárások alkalmazási lehetőségeit strukturális döntések szimulációs modellezésében.

A modellezési feladatok által felvetett problémák

Komplex rendszerek *modelljének megalkotásához* gyakran nélkülözhetetlen a rendszer leíró változói közötti kapcsolatok feltárása, a változók struktúrába rendezése, számuk redukálása. Ugyancsak felvetődhet az igény a rendszer objektumainak (alrendszereinek) leíró változóik alapján történő tipologizálására és összevonására is.

A változók és objektumok strukturálása és összevonása nem az egyetlen módja a rendszerelemzés megkönnyítésének. Másik módja a kvantitatív skálákon mért változóknak (illetve változócsoportoknak) kvalitatív változókká történő transzformációja. A kvalitatívvá transzformált változók és a rendszer egyéb – kvantitatív vagy kvalitatív – változói között általában jóval egyszerűbben kezelhető statisztikai, illetve logikai összefüggések állapíthatók meg, mint ez utóbbiak és az eredeti kvantitatív változók (változócsoportok) között.

A modell-változók vagy objektumok összevonása, valamint a kvalitatív skálák alkalmazása a rendszerelemzés matematikai és számítástechnikai problémáinak enyhítésén túl a modellek alkalmazásával kapcsolatos *emberi döntések* hatékonyságát is növeli. Döntéselméleti kísérletek bizonyították, hogy a döntéshozóknak egy-egy szituáció felismerésében és megítélésében nyújtott teljesítménye a szituációt leíró változók számának, pontosabban a változók lehetséges értékkombinációi számának növekedésével rohamosan csökken (*Miller*, 1967; *Edwards, Philips*, 1966; *Tversky, Kahneman*, 1972).

A következőkben néhány, a modell-alkotás és a modell-használat során felmerülő tipikus döntési feladaton mutatjuk be a fenti elvek érvényesítési lehetőségeit.

Kvantitatív skáláknak leképezésekor kvalitatív skálára az osztályok határainak kijelölésével tulajdonképpen meghatározzuk, hogy a kvantitatív változók értékeinek különbsége mely tartományok között lényegi, és mely tartományo-

kon belül hanyagolható el az adott célú elemzés szempontjából. Ez igen hasznos lehet pl. a *modellek érvényességének vizsgálatánál*, amikor a modell által előállított egyes adatokat a valóságos rendszer tényleges adataival kell összehasonlítanunk. A modell érvényességének kritériuma az, hogy a modellezett rendszer és a modell megfelelő adatai adott pontossági korláton belül megegyezzenek. A kívánt pontosság megadása körül felvetődő problémákat a vizsgált változók kvalitatív leképzésével, azaz a lényeges és kevésbé lényeges különbségek előzetes szétválasztásával jórészt kiküszöbölhetjük.

A kvalitatív skálák alkalmazásának említett előnye fontos lehet azokban az esetekben is, amikor nem egyetlen jól definiált feltételrendszer és célfüggvény alapján keressük a legjobb döntési alternatívát, hanem több *döntési politikát* akarunk összevetni az ezek eredményességét kifejező ún. cél-változókra gyakorolt hatásuk alapján. (Ez a helyzet általában a heurisztikus eljárásoknál.) A döntéshozó a több cél-változónak egyetlen kvalitatív preferenciaskálára való leképezésével értékítéleteit, a cél-változóknak az értékelés szempontjából ekvivalens tartományait határozza meg. Ez nagymértékben megkönnyíti a döntési politikák értékelését és összehasonlítását.

Másrészről a döntési politikákat reprezentáló kvantitatív változókat – az ún. tényezőket – is átalakíthatjuk kvalitatív változókká. A tényezők vagy a cél-változók számának csökkentése, illetve ezek kvalitatív transzformációja megkönnyíti annak megállapítását, hogy a döntési politika tényezői milyen hatást gyakorolnak a modell működésére *(modell-érzékenységvizsgálat)*, s ezáltal elősegíti a legeredményesebb politikai megtalálását.

A leíró változók vektorai és az osztályok közötti kapcsolat (osztályhatárok vagy osztályba tartozási függvény) explicit megadása ugyancsak nehéz döntési problémát vethet fel, különösen sokváltozós terek esetében. Sok esetben alkalmazható az a módszer, hogy a döntéshozók implicit módon, azaz összetartozó kvantitatív-kvalitatív értékeket (ún. tanítási mintákat) adnak meg és ebből számítógépi úton megfelelő statisztikai eljárások állítják elő a minta alapján legvalószínűbb összefüggéseket. Ezen eljárásoknak természetesen alkalmasaknak kell lenniök a döntéshozó inkonzisztens döntéseinek kiszűrésére is. Így lehetővé válik, hogy az ember is fokozatosan tanuljon a gépi eljárások által szolgáltatott eredményekből.

A döntések megfelelő előkészítése, az egyidejűleg figyelembe veendő változók számának és típusának helyes meghatározása, az emberi tanulás biztosítása különösen a folyamatos emberi beavatkozást igénylő, interaktív eljárások (pl. számítógépes szimuláció) alkalmazása esetén lényeges.

Statisztikai osztályozó eljárások

A fentiekben felsorolt problémák megoldására többek között a statisztikai osztályozó (alakfelismerő) eljárások alkalmasak.

Ezek egyik csoportját a *tanító nélkül osztályozó* (cluster analízis) eljárások képezik, amelyek az objektumok tulajdonságait leíró változók értékei alapján az objektumok osztályozására vonatkozó hipotéziseket generálnak. A másik csoportba a *tanítóval működő osztályozó* eljárások tartoznak, amelyek az elemeikkel megadott osztályokból (tanítási mintákból) indulnak ki. Az eljárások feladata, hogy a leíró változók és az osztályokat reprezentáló kvalitatív változók között összefüggéseket határozzanak meg, s egyben lehetővé tegyék

újabb objektumok osztályba sorolását. Mindkét eljárástípusnál az az objektumok összevonásának, az objektumok és osztályok egymáshozrendelésének kritériuma valamely geometriailag vagy statisztikusan értelmezhető távolság (pl. euklideszi távolság, osztályon belüli négyzetes eltérés stb.) minimalizálása.

A clusteranalízis-eljárásokat az általuk igényelt kiindulási információ mennyisége és jellege alapján két nagy csoportba sorolhatjuk.

A hierarchikus eljárások az objektumok eloszlására, illetve az osztályok számára vonatkozóan semmiféle előzetes információt nem igényelnek. Az objektumok összevonásával a közöttük levő összefüggések hierarchiáját határozzuk meg. *A nem hierarchikus eljárások* adott (rögzített) vagy az algoritmusok által iteratív módon változtatható számú osztályt képeznek az objektumokból. Az osztályok számára vonatkozó információt tehát előre meg kell adni számukra.

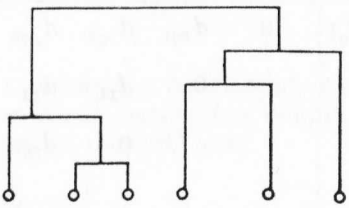
A modellezés során – mint azt példánkból is látni fogjuk – célszerű a különböző mennyiségű kiindulási információt igénylő hierarchikus és nem hierarchikus cluster analízist, valamint a tanítóval működő eljárásokat kombináltan, egymásra építve alkalmazni.

A következőkben részletesen ismertetjük a vizsgálatainkban alkalmazott cluste ranalízis eljárásokat. Ezek programját FORTRAN nyelven a CDC 3300 számítógépre készítettük el, s a közeljövőben elkészül egy, az R20 számítógépen futtatható változatuk is. A vizsgálatainkban ugyancsak felhasznált tanítóval működő – ún. potenciálfüggvényes – osztályozó eljárásra vonatkozóan l. pl. *Andrews*, 1972.

A Wishart-féle hierarchikus clusteranalízis-eljárás

A hierarchikus cluster analízis az egyes objektumokat rangsorolja, ún. hierarchiát állapít meg közöttük. Az egyes objektumok közötti hierarchikus összefüggést fával vagy más néven dendogrammal szokták ábrázolni (1. sz. ábra). Az ábrán látható fán a körök egy-egy objektumot jelentenek, az összekötő vonalak az objektumok közötti kapcsolatokat. A dendogram függőleges tengelye az egyesítési szintek megadására szolgál. A hierarchikus clusteranalízis fogalmát az alábbiakban határozzuk meg.

Legyen $E = \{x_1, x_2, \ldots, x_n\}$ az objektumok halmaza, amelyet kiindulásként tekintsünk $P_1, \ldots P_n$ egyelemű clusterek halmazának. Válasszuk ki közülük azt a P_p és P_q osztályt, amelyek valamilyen értelemben a legköze-



1. ábra

lebb vannak egymáshoz, s ezeket vonjuk össze egy osztályba. Az így kapott osztályhalmaznak már csak $n-1$ eleme lesz:

$$P_1, P_2 \ldots, (P_p, P_q), \ldots, P_n$$

Ismételjük meg ezt az eljárást. Így az osztályhalmazoknak egy olyan sorozatát kapjuk, amelynek a továbbiakban $n-2$ eleme, majd $n-3$ stb. eleme lesz. Végül egyetlen osztályt kapunk, amely az eredeti n osztályt tartalmazza.

Kérdés, hogy mely osztályokat nevezzük legközelebbieknek. Ehhez definiálni kell két objektum távolságát és ennek alapján meg kell határozni, hogy mit értünk két osztály távolságán. Érthetjük ezalatt pl. egymáshoz legközelebbi, illetve egymástól legtávolabbi elemeik, vagy középpontjaik (centroidjaik) távolságát stb., a különböző hierarchikus módszerek lényegében ebben különböznek egymástól.

Legyen az osztályok közötti távolságfüggvény

$$d: E^* \times E^* \to R^-,$$

ahol E^* az E részhalmazainak halmaza, és R^+ a nem negatív valós számok halmaza.

Jelöljük a P_i és P_j osztályok távolságát, azaz $d(P_i, P_j)$-t d_{ij}-vel. A kezdeti n számú egyelemű osztályra így egy $n \times n$-es D_0 távolság-mátrixot nyerünk:

$$D_0 = \begin{array}{c|cccc} & P_1 & P_2 & P_3 & \ldots P_n \\ \hline P_1 & 0 & d_{12} & d_{13} & d_{1n} \\ P_2 & & 0 & d_{23} & d_{2n} \\ \vdots & & & 0 & d_{3n} \\ P_n & & & & 0. \end{array}$$

Tegyük fel, hogy P_p és P_q vannak egymáshoz a legközelebb. Akkor a P_p és P_q összevonása után egy új $(n-1)\times(n-1)$ dimenziós távolság-mátrix elemeit kell meghatározni.

$$D_1 = \begin{array}{c|ccccc} & (P_p, P_q) & P_1 & P_2 & P_3 & \ldots P_n \\ \hline (P_p, P_q) & 0 & d_{pq1} & d_{pq2} & d_{pq3} & \ldots d_{pqn} \\ P_1 & & 0 & d_{12} & d_{13} & \ldots d_{1n} \\ P_2 & & & 0 & d_{23} & \ldots d_{2n} \\ \vdots & & & & & \vdots \\ P_n & & & & & 0. \end{array}$$

A D_1 mátrixnak $n-2$ sora azonos a D_0 mátrix megfelelő sorával csak egy sorát kell újra kiszámítani. Ha azonban meg tudnánk adni a D_i, $i = 1 \ldots n-1$ számolására egy olyan transzformációs formulát, amely nem az eredeti objektumok, hanem csak az előző mátrix adatait használja, akkor ez az eljárás legegyszerűsödne.

Wishart adott meg egy olyan rekurzív formulát, amelynek segítségével hat különböző hierarchikus módzsert lehet megoldani (*Wishart*, 1969). A módszerek egymástól a d függvény definíciójában különböznek. Az objektumok között értelmezett távolság valamennyi módszernél az euklideszi távolság. Ha a P_p osztályt egyesítjük a P_q osztállyal, akkor az így kapott új P_r osztály távolságát a többi P_t ($t = 1, \ldots n$; $t \neq p$, $t \neq 0$) osztálytól is ki kell számítani. Így a távolság mátrix is megváltozik. Az új távolság-mátrixot a következő transzformációs formula segítségével számítjuk:

$$d_{ir}^2 = \alpha_p d_{ip}^2 + \alpha_q d_{iq}^2 + \beta d_{pq}^2 + \gamma \mid d_{ip}^2 - d_{iq}^2 \mid, \tag{1}$$

ahol α_p, α_q β és γ paraméterek, és $P_r = P_p \cup P_q$.

Jelöljük k_i-vel az i-edik osztály elemeinek számát. Az α_p, α_q, β és γ paraméterek különböző megválasztásával a következő hat módszert kapjuk:

A) *Legtávolabbi szomszéd módszer*

Két osztály közötti távolságot akkor tekintjük minimálisnak, ha az összevonásukkal nyert osztály legtávolabbi objektumai közötti távolság minimális. Ebben az esetben a paraméterek:

$$\alpha_p = \alpha_q = \frac{1}{2}; \quad \beta = 0; \quad \gamma = \frac{1}{2}.$$

B) *Legközelebbi szomszéd módszer*

A két osztály közötti távolság akkor minimális, ha az összevonással kapott osztály legközelebbi objektumai közötti távolság minimális. A paraméterek:

$$\alpha_p = \alpha_q = \frac{1}{2}; \quad \beta = 0; \quad \gamma = -\frac{1}{2}.$$

C) *Centroid módszer*

Két osztály közötti távolságot a centroidjaik közötti távolsággal definiáljuk. Olyan osztályokat von tehát össze a módszer, amelyek centroidja közötti távolság minimális.

A transzformációs formula paraméterei:

$$\alpha_p = \frac{k_p}{k_r}; \quad \alpha_p = \frac{k_p}{k_r}; \quad \beta = \alpha_p \alpha_q; \quad \gamma = 0.$$

D) *Medián módszer*

Az osztályok közötti távolságot a mediánjaik közötti távolsággal definiáljuk. Az osztályozásnál azokat az osztályokat vonjuk össze, amelyek mediánjai közötti távolság minimális. A paraméterek:

$$\alpha_p = \alpha_q = \frac{1}{2}; \quad \beta = -\frac{1}{4}; \quad \gamma = 0.$$

E) *Csoportátlag módszer*

A módszer a két osztály közötti távolságot a két osztály elemei közötti átlagos távolsággal definiálja és az osztályozásnál azokat az osztályokat vonja össze, ahol az átlagos távolság minimális.

A paraméterek:

$$\alpha_p = \frac{k_p}{k_r}; \quad \alpha_q = \frac{k_q}{k_r}; \quad \beta = \gamma = 0.$$

F) *Ward módszere*

Az osztályozásnál olyan új osztályok létrehozására törekszünk, amelyekben négyzetes hibának az összevonás által eredményezett növekedése minimális.

A paraméterek:

$$\alpha_p = \frac{k_t + k_p}{k_t + k_r}; \quad \alpha_q = \frac{k_t + k_q}{k_t + k_r}; \quad \beta = \frac{-k_t}{k_t + k_r}; \quad \gamma = 0.$$

A Diday-féle hierarchikus clusteranalízis-eljárás

A nem hierarchikus eljárások lényege, hogy az objektumok egy kezdeti osztályozásából, vagy az osztályok valamely feltételezett jellemzőiből kiindulva iteratív módon változtatják az objektumok besorolását, mindaddig, amíg valamilyen szempont szerint jobb osztályozást nyernek.

Az eljárásoknak az objektumokat előre rögzített, vagy az eljárás során iteratív módon meghatározott K számú osztályba kell sorolniok. Ettől függően megkülönböztetünk rögzített számú osztályt feltételező algoritmusokat (pl. *Forgy*, 1966; *McQueen*, 1967 stb.), és e változó osztályszámú algoritmusokat (pl. *Ball–Hall*, 1965). Az előbbi csoportba tartozik az alábbiakban ismertetendő Diday-féle eljárás is (*Diday, Govaert*, 1974).

Legyen:

E az R^q (q dimenziós euklideszi tér) részhalmaza, az objektumokat reprezentáló vektorok halmaza,

E^* az E részhalmazainak halmaza,

P_K^* az E K-osztályú particióinak halmaza,

$$P \in P_K^* \Longleftrightarrow P = (P_1, P_2, \ldots, P_K), \qquad P_i \in E^*,$$
$$P_i \cap P_j = \Phi \quad \text{ha} \quad i \neq j, \qquad UP_i = E.$$

L^* az ún. „magok" tere. A magok az E részhalmazaihoz rendelt jellemzők,

L_K^* az L^*-ból képzett K-elemű sorozatok halmaza,

$$L \in L_K^* \Longleftrightarrow L = (\lambda_1, \lambda_2, \ldots, \lambda_K), \lambda_i \in L^*,$$

D az objektumok és magok közötti értelmezett távolságfüggvény,

$$D : E \times L^* \to R^+,$$

ahol R^+ a nem negatív valós számok halmaza.

R a magok és az objektumokból képzett halmazok között értelmezett távolságfüggvény

$$R : L^* \times E^* \to R^+$$

W az R által szolgáltatott távolságoknak egy-egy K-partícióra történő összegezése, egyúttal a minimalizálandó célfüggvény,

$$W : L_K^* \times P_K^* \to R^+,$$

$$W(L, P) = \sum_{i=1}^{K} R(\lambda_i, P_i).$$

Definiáljuk továbbá az alábbi leképezéseket:

f: $L_K^* \to P_K^*$,

olyan leképezés, amelynek eredménye olyan új P partíció:

$$L = (\lambda_1, \ldots, \lambda_K); f(L) = P = (P_1, \ldots, P_K),$$

amelyre:

$$P_i = \{x \in E \mid D(x, \lambda_i) \leq D(x, \{\lambda_j), \forall_j \neq i\}.$$

g: $P_K^* \to L_K^*$

olyan leképezés, amelynek eredménye olyan új K-elemű L mag-halmaz:

$$P = (P_1, \ldots, P_K); g(P) = L = (\lambda_1, \ldots, \lambda_K),$$

amelyre:

$$R(\lambda_i, P_i) = \min_{\lambda \in L^*} R(\lambda, P_i).$$

Legyen adott P^0 kezdeti K-partició, vagy L^0 kezdeti K-elemű mag-halmaz.
Az f és g leképezések végrehajtásával rekurzív módon előállíthatók az alábbi sorozatok:

$$\begin{aligned} L^n &= g(P^n), \\ P^{n+1} &= f(L^n), \\ w_n &= W(L^n, P^n). \end{aligned}$$

Bizonyítható, hogy amennyiben

$$W[L, f(L)] \leq W(L, P), \qquad L \in L_K^*, P \in P_K^*, \tag{2}$$

akkor a w_n sorozat monoton csökken és véges számú lépés után *lokális* minimumot ér el.
A (2) feltétel teljesüléséhez elégséges, hogy

$$R(\lambda_i, P_i) = \sum_{x \in p_i} D(x, \lambda_i) \tag{3}$$

A Diday-féle algoritmus P^0-ból vagy L^0-ból kiindulva az f és g leképezések sorozatát hajtja végre, mindaddig, amíg a W értéke tovább már nem csökken. Tetszőleges D távolságfüggvény alkalmazható, R-nek pedig a (3) feltételt kell teljesítenie, így az algoritmus

$$W(L, P) = \sum_{i=1}^{K} \sum_{x \in P_i} D(x, \lambda_i)$$

típusú célfüggvény minimalizálására alkalmas.

A következőkben a Diday-féle algoritmus három speciális esetét vizsgáljuk meg:

A) Legyenek magok az osztályok középérték-vektorai:

$$\lambda_i = \mu_i = \frac{1}{N_i} \sum_{x \in P_i} x,$$

és alkalmazzuk távolságfüggvényként az euklideszi távolságot:

$$D(, \lambda_i) = (x - \mu_i)^T (x - \mu_i).$$

Ennek megfelelően a célfüggvény az osztályon belüli négyzetes eltérések összege, azaz:

$$W(L, P) = \sum_{i=1}^{K} \sum_{x \in P_i} (x - \mu_i)^T (x - \mu_i).$$

Megjegyezzük, hogy a Diday-féle algoritmus ezen speciális esete megegyezik a Forgy által kidolgozott eljárással (*Forgy*, 1966).

B) Válasszuk magnak egy-egy osztály középérték-vektorát és kovariancia mátrixát:

$$\lambda_i = (\mu_i, V_i)$$

$$V_i = \frac{1}{N_i} \sum_{x \in Pi} (x - \mu_i) (x - \mu_i)^T.$$

Alkalmazzuk távolságfüggvényként a *Mahalanobis*-távolságot:

$$D(x, \lambda_i) = (\det V_i)^{1/q} (x - \mu_i)^T V_i^{-1} (x - \mu_i).$$

Ekkor:

$$W(L, P) = \sum_{i=1}^{K} (\det V_i)^{1/q} \sum_{x \in P_i} (x - \mu_i)^T V_i^{-1} (x - \mu_i),$$

azaz a célfüggvény az osztályok inercia-főtengelyeire transzformált négyzetes eltérések összege.

C) Tegyük fel, hogy az objektumok K számú ismert típusú eloszlás keverékéből származnak, azaz:

$$F(x) = \sum_{i=1}^{K} p_i \varphi(\lambda_i, x),$$

ahol $F(x)$ a keverékeloszlás sűrűségfüggvénye, $\varphi(\lambda_i, x)$ az i-edik komponens sűrűségfüggvénye, p_i az i-edik komponens a priori valószínűsége.

Válasszuk magnak az osztályok sűrűségfüggvényeinek λ_i paramétereit, a távolságfüggvényt pedig definiáljuk egy-egy objektumnak egy-egy osztályba való tartozása a posteriori valószínűségének függvényeként – (az osztályok a priori valószínűségeit egyenlőnek tekintve) – az alábbi módon:

$$D(x, \lambda_i) = \log [C/\varphi(\lambda_i, x)],$$

ahol C egy 1-nél nagyobb konstans.
Ekkor:

$$W(L, P) = C' - \log \prod_{i=1}^{K} \prod_{x \in P_i} \varphi(\lambda_i, x).$$

Az így használt Diday-féle algoritmus tehát az objektumoknak a hozzájuk rendelt osztályokba tartozásának – a teljes objektumhalmazra számított – együttes valószínűségét igyekszik maximalizálni.

Speciálisan Gauss-eloszlást feltételezve, $\lambda_i = (\mu_i, V_i)$-t választva, ahol μ_i az i-edik osztály középértékvektora, V_i az i-edik osztály kovariancia mátrixa:

$$D(x, \lambda_i) = C + \frac{1}{2} [\log(\det V_i) + (x - \mu_i)^T V_i^{-1} (x - \mu_i)]$$

$$W(L, P) = C' + \frac{1}{2} \left[\sum_{i=1}^{K} N_i \log(\det V_i) + \sum_{i=1}^{K} \sum_{x \in P_i} (x - \mu_i) V_i^{-1} (x - \mu_i) \right].$$

Alkalmazási példák

1. Vizsgálat

A vizsgálat célja az volt, hogy egy ágazati modell, a magyar szénhidrogénipar strukturális döntéseinek vizsgálatára készült szimulációs modell (*Vári, Kelemen*, 1974) eredményeit elemezzük és meghatározunk egy kellőképpen eredményes döntési politikát.

A modell a szénhidrogénipari vertikum termelési, tárolási és értékesítési folyamatait írta le. A strukturális döntési politikák a rendszer termelő-, tároló- és szállítókapacitásait érintő döntésekből (pl. beruházások, átcsoportosítások stb.) tevődtek össze, az eredményességet pedig – többek között – a hazai igények kielégítettségi fokával mértük. Így a vizsgált tényezők az említett kapacitások idősorai, a cél-változók pedig a rendszer által kibocsátott legfontosabb termékek kínálatai és keresleti idősorai voltak.

Mivel az egyes termékek termelőkapacitásai nem változtathatók meg egymástól függetlenül (ez a szénhidrogénipari technológia sajátos ága), a szállító- és tárolóeszközök pedig a különböző termékfajták között bizonyos mértékig átcsoportosíthatók, így szükséges volt a termékek adatainak együttes vizsgálata. A 10 legfontosabb terméknek 13 évre, azaz 52 negyedévre vonatkozó kínálati és keresleti adatainak együttes áttekintése, illetőleg ennyi adatra nézve az érzékenységvizsgálat elvégzése nehéz feladatot adott volna, ezért előzleg cluster analízisnek vetettük alá az eredményeket.

A szénhidrogénipari termékek nagy részénél (üzemanyagok, fűtőanyagok) a keresleti idősorok szezonális ingadozásokat tartalmaznak, így a kereslet-kínálat viszony évközi alakulása is jelentős eltéréseket mutathat a különböző anyagoknál és időszakokban. Az osztályozandó vektorokat ezért egy-egy termék egy-egy évre vonatkozó negyedéves bontású kínálat-kereslet hányadosaiból alakítottuk ki. Az így nyert 130 db négyegyelemű vektorból először hierarchikus osztályozással egy megfelelő induló osztályozást képeztünk ki. A hierarchikus elemzést a *Wishart*-féle eljárással (a *Ward*-módszerrel) elvégezve, azon vektorokból alakítottunk ki osztályokat, amelyek viszonylag korán kapcsolódtak össze és összevonásuk más osztályokkal a négyzetes eltérések összegét viszonylag nagymértékben megnövelte volna.

Az így nyert osztályozásból kiindulva a Diday-féle nem hierarchikus algoritmussal néhány iteráció után olyan új osztályozást nyertünk, amelyhez tartozó összes négyzetes eltérés az indulási értéknek kb. 2/3 része volt. Mind a hierarchikus, mind a nem hierarchikus osztályozásnál euklideszi távolsági mértéket alkalmaztunk.)

A kialakult 17 osztály mindegyike az alábbi 4 főbb viselkedéstípus valamelyikét reprezentálta:

a) a kínálat jól követi a keresletet,
b) hiányok és többletek váltják egymást,
c) állandó hiány,
d) állandó túlkínálat.

A modell összefüggéseinek ismeretében általánosságban megállapítható, hogy a b) típusú osztályoknál a tárolási és szállítási kapacitások növelése, a c) típusúaknál az adott termék (esetleg ennek nyersanyagai) termelési kapacitásainak növelése, a d) típusúaknál az exportlehetőségek bővítése eredményezheti a kereslet-kínálati viszonyok javulását. A felsorolt típusokon belüli osztályok a hiányok és többletek mértékében, vagy fázisviszonyaiban különböztek egymástól. Ezek további vizsgálata alapján feltárhatók a termelő-, tároló- és szállítókapacitások időszakos átcsoportosítási lehetőségei is.

A modell összefüggéseinek ismeretén túl érzékenységi vizsgálatra is szükség volt ahhoz, hogy a strukturális változtatások arányait és mértékét megfelelően határozhassuk meg. Az érzékenységi vizsgálat során azt kellett feltárni, hogy az egyes kínálat-keresleti vektorok mozgása milyen törvényszerűségeket követ a strukturális döntések függvényében. Első lépésben a vektoroknak az osztályok közötti mozgását vizsgáltuk meg, ezután került sor – a kritikus termékeknél és időszakokban – a finomabb kvantitatív összefüggések feltárására.

Példánkban tehát a clusteranalízis az eredmények áttekintését, a lényegi összefüggések megragadását, s az érzékenységi vizsgálat hatékony elvégzését segítette elő.

2. *Vizsgálat*

Célunk az volt, hogy egy beruházási szimulációs játék (*Rabár, Kelemen*, 1975) modelljét, a döntési lehetőségeket didaktikai szempontból helyesen alakítsuk ki, és hogy lehetővé tegyük a játékosok teljesítményének értékelését gépi úton.

A modell több vállalatnak és a központi gazdaságirányításnak a tevékenységét fogta át. A játékosok egy-egy vállalatot képviseltek. Feladatuk az volt, hogy vállalatuk helyzetének ismeretében beruházási döntéseket hozzanak. Döntéseik következményei alapján meg kellett tanulniok a döntések és a vállalat helyzetére gyakorolt hatásuk összefüggéseit.

A beruházási döntéseket az alábbi paraméterekkel jellemeztük:

- a beruházás költsége;
- a beruházás nyereséghozama;
- a beruházás átfutási ideje;
- a beruházás célja (pl. pótlás, munkaerőhelyettesítés stb.).

A fenti jellemzők adott lehetséges kombinációihoz tartozó 144 döntésből, a tanulásnak és az összefüggések meghatározásának megkönnyítése érdekében, a Wishart-féle hierarchikus eljárással 7 osztályt képeztünk. A képződött osztályokba az alábbi típusú beruházási döntéseket soroltuk:

a) pótlás vagy munkaerőhelyettesítés,
b) kisebb felújító jellegű beruházás,
c) közepes beruházás,
d) kisebb rekonstrukció,
e) kisebb rekonstrukció és kisebb felújítás,
f) kisebb rekonstrukció és közepes beruházás,
g) közepes vagy nagy rekonstrukció.

A fenti osztályokat kvalitatív (rendezési) skálán tudtuk elhelyezni, oly módon, hogy a hozzájuk rendelt rangszámok az egyes osztályokba tartozó döntések horderejének sorrendi viszonyait tükrözték. Ily módon pl. a pótlás-munkaerőhelyettesítés az 1, a közepes vagy nagy rekonstrukció a 7 rangszámot kapta.

Az osztályozással elértük, hogy a játékosnak egyidejűleg csak 7 beruházási típus között kellett döntenie, majd további megfontolások figyelembevételével a választott típuson belül egyetlen alternatívát kiválasztania.

Következő feladatunk az volt, hogy a modell vizsgálata és a játékosok döntéseinek automatikus értékelése céljából explicitté tegyük a döntési helyzet és a helyes döntés közötti összefüggéseket.

A döntéseket a clusteranalízissel kialakított osztályok (döntéstípusok) rangszámaival jellemeztük. A döntési helyzetet leíró változók (a vállalatot jellemző állóeszköz-forgóeszköz arány, nyereség-eszköz arány, eszköz-bér arány, nyereség, fejlesztési alap, központi hitel és támogatás, a korábban megkezdett beruházások terhei stb.) több különböző kombinációját használtuk fel vizsgálatainkban. E kvantitatív változók és a döntéseket reprezentáló kvalitatív változók közötti kapcsolat meghatározására egy tanítóval működő osztályozó eljárást, az ún. potenciálfüggvényes algoritmust (*Andrews*, 1972) alkalmaztuk. Tanítási mintaként felhasználtuk a lejátszott játékok azon összetartozó döntési helyzet-döntési párjait, amelyeknél a döntések kedvező hatásúnak bizonyultak.

Az eljárás meghatározta a döntési helyzet és a jó döntés között a legvalószínűbb függvénykapcsolatot. Így választ kaptunk arra, hogy melyek a döntési helyzetnek a döntések meghozatalánál elsődlegesen figyelembe veendő paraméterei és hogy milyen értelemben és milyen súllyal kell ezeket figyelembe venni.

A döntési helyzetet leíró változók közül az alábbiak bizonyultak lényegesnek a döntések szempontjából:

– az adott évben képződő nyereség;
– a fejlesztési alap + az adott évre esedékes központi hitel és támogatás;
– a korábban megkezdett beruházások aktuális költségei;
– a lekötött eszközök és a bérköltség hányadosa.

A legnagyobb (pozitív) súllyal a nyereség és a fejlesztési alap + hitel + támogatás szerepeltek, az eszköz-bérköltség hányados és a megkezdett beruházások terhei kisebb, de nem elhanyagolható szerepet játszottak a meghozott döntésekben. (Az utóbbi változó természetesen negatív súllyal szerepelt.)

A fentiekben meghatározott összefüggés módot nyújtott arra, hogy segítségével

– ellenőrizzük, hogy modellünk összefüggései megfelelnek-e a valóságos összefüggéseknek (érvényesség vizsgálat);
– felülvizsgáljuk, hogy modellünk eléggé érzékeny-e a döntések közötti különbségekre, alkalmas-e a lényegi összefüggések megtanítására, a játék egyéb véletlen komponensei nem fedik-e el ezeket az összefüggéseket;
– automatikusan értékeljük egy újabb játék játékosainak teljesítményét, összahasonlítva az általuk hozott döntéseket az elméleti összefüggés által meghatározott döntésekkel.

Láttuk, hogy a döntések kvalitatív skálára való leképezése olyan globális vizsgálatokat tett lehetővé, amelyekre enélkül nem lett volna módunk. Természetesen az így nyert információk csak a döntések osztályaira vonatkoznak. A következő lépés az egy-egy osztályba tartozó döntések közötti finomabb különbségek vizsgálata, amely az egy-egy osztályba tartozó elemek kis száma miatt aránylag egyszerű eszközökkel elvégezhető. A cluster analízis tehát esetünkben mind a döntési, mind az elemzési és értékelési feladatok dekompozícióját elősegítette.

Végül fölhívjuk a figyelmet néhány, az alkalmazások során felmerült problémára. Az egyik – a 2. vizsgálatnál – adódott abból, hogy mind a döntési helyzeteket, mind a döntéseket különféle mértékegységekben mért paraméterekből álló vektorokkal jellemeztük. Ezek együttes kezelése csak úgy volt lehetséges, hogy a vektorokat valamennyi dimenzió szerint normáltuk.

A második problémát az egymástól statisztikusan nem független változók kezelése adta. A mintavektorok halmazára korrelációanalízist végeztünk, így a 2. vizsgálatban szereplő, a döntési helyzeteket leíró változók között számottevő korrelációt találtunk. Az analízis eredményei alapján alakítottuk ki azokat a különböző, korrelált változókat nem tartalmazó változókombinációkat, amelyekből képzett vektorokra a potenciálfüggvényes osztályozó eljárást elvégeztük. Esetünkben a változók közötti kapcsolatok aránylag egyszerűek voltak, így a változók számának csökkentéséhez nem volt szükség gépi eljárások (pl. faktoranalízis) alkalmazására.

A Diday-féle eljárás alkalmazása során merültek fel az osztályok számának előzetes meghatározásával kapcsolatos problémák, mivel az eljárás előre meghatározott számú osztállyal dolgozik. Természetesen minél több osztályt engedünk meg, annál kisebbé tehető a négyzetes eltérések összege. Ugyanakkor általában nem célszerű túlságosan sok kevéselemű osztály képzése sem. A probléma megoldására beépíthető olyan algoritmus, amely a nagy elemszámú,

ill. nagy szórású osztályok szétbontásával, és a kis elemszámú, egymáshoz közeli osztályok egyesítésével iteratív módon alakítja ki a végső osztályozást. Ilyen algoritmust tartalmaz pl. az ISODATA eljárás (*Ball–Hall*, 1965). Olyan algoritmus azonban, amely automatikusan optimális megoldásra vezet, nem ismeretes. Sőt magának az optimalitási kritériumnak a megadása is problematikus. Ráadásul a fenti típusú algoritmusok sok előzetes információ megadását igénylik (pl. maximálisan és minimálisan megengedett osztályelemszám stb.), és túlságosan mechanikusak is, ezért alkalmazásukat nem láttuk célszerűnek. Ehelyett a Diday-eljárást több különböző számú osztály esetére végrehajtottuk úgy, hogy az induló osztályokat a korábbi osztályozások során kialakult osztályok szétbontása, ill. egyesítése révén nyertük. Az eredményeket megfelelő mutatók (pl. az osztályok szórásainak átlaga) segítségével összehasonlítva, kiválasztottuk a legkedvezőbb megoldást.

Az előzőekből kitűnik, hogy az osztályozó eljárások – heurisztikus jellegüknél fogva – nem alkalmazhatók mechanikusan, megkövetelik a folyamatos emberi beavatkozást, az eredmények állandó elemzését és értékelését.

(Beérkezett: 1977. ápr. 12-én.)

IRODALOMJEGYZÉK

1. Andrews, P.: Introduction to Mathematical Techniques in Pattern Recognition. Wiley-Interscience. New York, 1972.
2. Ball, G. H. – Hall, D. J.: ISODATA, a Novel Method of Data Analysis and Pattern Classification Techn. Report, Stanford Research Inst. Menlo Park. California, 1965.
3. Diday, E. – Govaert, G.: Apprentissage et Mesures de Ressemblances Adaptatives. IRIA, Rapport de Recherche, 1974. No. 89.
4. Edwards, W. – Phillips, L. D.: Conservatism in a Simple Probability Inference Task. Journal of Experimental Psychology, 1966.
5. Forgy, E. W.: Classification so as to Relate to Outside Variables. Final Rep. Conf. Cluster Analysis of Multivariate Data, Washington, 1966.
6. MacQueen, J. B.: Some Methods for Classifications and Analysis of Multivariate Observations. Proc. Symp. Math. Statist. and Probability, 5th. Berkeley, University of California Press, 1967.
7. Millen, G. A.: The Magical Number Seven, Plus or Minus Two in: The Psychology of Communication. Penguin Books, 1967.
8. Rabár, F. – Kelemen, K.: A központi és a vállalati beruházási politika szimulációja. A Számítógépes Rendszerszimuláció Szimpozium előadása, 1975.
9. Tversky, A. – Kahneman, D. K.: Subjective Probability. A Judgement of Representativeness. Cognitive Psychology, 1972.
10. Vári, A. – Kelemen, K.: Az OKGT strukturális döntéseinek vizsgálatára készült szimulációs modell formális és verbális leírása. INFELOR tanulmány, 1974.
11. Wishart, D.: An Algorithm for Hierarchical Classifications. Biometrics, 1969.

ECONOMIC APPLICATION OF CLUSTER ANALYSIS PROCEDURES

The article is aimed at presenting the application possibilities of statistical classification procedures – first of all cluster analysis – for the reduction of the complexity of economic system-modelling as well as at reviewing two less-known cluster analysis procedures.

In the first part some complexity problems are dealt with which emerge in the different phases of system modelling (model formation, validity test of the model, sensitivity analysis, evaluation of decision rules, revealing and learning of connections, etc.). A way to their solution is the application of statistical procedures.

In the second part the main types of statistical classification procedures and the theoretical possibilities of their application are briefly reviewed. *Wishart*'s hierarchic and *Diday*'s non-hierarchic cluster analysis procedures are discussed in detail. As a matter of fact the former is suitable for carrying out six different hierarchic methods and within the latter we can vary the distance function and the objective function (expressing the anality of the classification) within wide limits.

In the third part examples are presented for the application of classification procedures in economic modelling. In one of the examples we present the results of a simulation model, which was set up for the examination of structural decisions and sensitivity analysis. Another example deals with an investment simulation game, where the above procedures facilitated the model formation, the evaluation of results and learning.

Finally, we offer potential solutions to problems often arising in the course of the application of procedures.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА ПРИ МОДЕЛИРОВАНИИ ЭКОНОМИЧЕСКИХ СИСТЕМ

Цель данной статьи заключается в том, чтобы показать возможности применения методов статистической классификации, в первую очередь кластерного анализа, в отношении ограничения комплексности экономического системного моделирования, а также изложить два относительно малоизвестных метода кластерного анализа.

В первой части затрагиваются проблемы комплексности, возникающие на различных фазах системного моделирования (разработка модели, изучение действенности модели, изучение чувствительности, оценка политики принятия решения, выявление взаимосвязей и их изучение и т. д.), одним из возможностей решения которых является использование статистических методов.

Во второй части дается краткий обзор основных типов методов статистической классификации и принципиальных возможностей их использования. Детально рассматриваются такие методы кластерного анализа как иерархический метод Вишарта и неиерархический метод Дидея. Первый из них, по существу, пригоден для применения шести различных иерархических методов, а во втором случае функция расстояния и функция, выражающая правильность классификации, могут изменяться в довольно широких пределах.

В третьей части приводятся примеры по использованию методов классификации в экономическом моделлировании. В одном из примеров излагается использование результатов анализа и исследования чувствительности симуляционной модели, составленной для изучения структурных решений. В другом примере описывается применение симуляционной игры по капитальным вложениям, когда указанные выше методы направлены на облегчение разработки модели, оценки результатов, а также и учебы.

В заключении, указываются проблемы, часто возникающие в ходе применения этих методов и излагаются некоторые возможности их решения.