RUZSÁNYI TIVADAR—LELKES PÉTER

A RULE–3 automatikus osztályozási eljárás

1. Bevezetés

Napjainkban az automatikus osztályozási eljárásoknak nagyon változatos és folyamatosan bővülő szakirodalma áll rendelkezésünkre, sőt a számítógépekkel szállított software-k sem nélkülözik a szélesebb körben is ismert eljárások egyikét-másikát. Annak ellenére azonban, hogy ma már a legkülönfélébb problémák megoldására alkalmas, számítógépi alkalmazásra javasolt algoritmusok széles tárházából válogathatunk, megismerhetjük a különböző eljárások alkalmazásának feltételeit, előnyeit, hátrányait (PÁRNICZKY, 1976; FÜSTÖS—MESZÉNA—SIMONNÉ, 1977; S. BENEDIKT—VÁRI, 1977; SVÁB, 1979), a gyakorlat folyamatosan olyan újabb és újabb problémákat vet fel, melyek megoldásához célszerű a már ismert módszerek valamilyen ésszerű ötvözetét kidolgozni, létrehozni egy a korábbiaktól eltérő, viszonylag új eljárást. A RULE–3 eljárás kifejlesztésekor is lényegében erről volt szó.

Az automatikus osztályozási eljárásokra vonatkozó elméleti elemzések (GOWER, 1967; SOKAL, 1974), az eljárások által nyújtott eredményeknek, ill. maguknak az eljárásoknak összehasonlítása (DUBES—SAIN, 1976; ATTINGER et alia, 1978), továbbá a témákhoz kapcsolódó saját tapasztalataink (RUZSÁNYI 1979; 1980) alapján ugyanis nyilvánvaló számunkra az, hogy az automatikus osztályozási eljárások nem alkalmazhatók mechanikusan. Az alkalmazás során számos, a vizsgált probléma jellegéből, a rendelkezésre álló adatok tartalmából, a jellemzők (tényezők) kiválasztásából, a változók típusaiból adódó kérdést kell részletesen elemezni. Problémát okozhat a hasonlósági mérték, az osztályok kialakulását meghatározó döntési kritérium értelmezése, figyelembevételük az eredmények interpretálásakor. A gyakorlati alkalmazások során sűrűn előfordul az is, hogy a különböző eljárásokat felhasználó szakemberek — mivel csak részlegesen ismerik a bonyolultabb eljárások elméleti apparátusát — bizonyos kétkedéssel, bizalmatlansággal fogadják az eredményeket.

A RULE–3 kidolgozásakor igyekeztünk figyelembe venni az előzőekből fakadó tanulságokat, de kihasználtuk a korszerű hardware és software által kínált lehetőségeket is. A kidolgozott programcsomagok ugyanis lehetővé teszik az osztályozás kiinduló mátrixának, az osztályozandó objektumoknak (elemeknek), ill. tulajdonságoknak (tényezőknek) interaktív szűrését.

2. Az eljárás célja, az alkalmazás előkészítése

Az eljárás eredetileg a korszerű döntési módszerek alkalmazásához, az összegyűjtött tényanyag elemzéséhez kapcsolódott (RUZSÁNYI—VÁRI, 1980). Ezen probléma kutatása külföldön már a korábbiakban is folyt és egy hasonló

vizsgálat során olyan kérdőíves módszert használtak, melyet mi is adaptáltunk a következő előnyei miatt:

— a probléma átgondolása után a kérdőív rövid idő alatt kitölthető és
— viszonylag nagy számú, az aktuális problémára vonatkozó állítás („kérdés") érvényessége, illetve érvénytelensége tárható fel (SCHULTZ—SLEVIN, 1975).

A kutatás első fázisában kellett tehát megoldanunk a kérdőívek feldolgozását (RUZSÁNYI—LELKES, 1980). A kérdés az volt, hogy milyen homogén csoportokba rendeződnek a kérdőívek, figyelembe véve tartalmukat, azaz az egyes állítások érvényességére vonatkozó ítéleteket, továbbá, milyen csoportokba rendeződnek az állítások. A probléma annak kapcsán merült fel, hogy vajon elkülönülnek-e a sikeres, illetve sikertelen döntéselőkészítésekre vonatkozó kérdőívek, és az állítások milyen specifikus csoportja kapcsolódik ezekhez.

Az eljárást tehát valamely kérdőíves felmérés feldolgozására, a kérdőívek (vizsgálati elemek) és a kérdőívbe foglalt „kérdések" (tényezők) automatikus és hierarchikus osztályozására alakítottuk ki. Figyelembe vettük a komplex rendszerek elemzésével foglalkozó szakemberek azon igényét is, hogy olyan automatikus osztályozásra van szükség (KINDLER—PAPP, 1977), melyek lehetővé teszik

— sorrendi,
— intervallum és
— arányskálán

rendelkezésre álló adatok felhasználását egyazon eljárás keretében. A RULE—3 ezen túlmenően lehetőséget nyújt nominális skálán rendelkezésre álló adatoknak a hasznosítására is, természetesen a szükséges skálatranszformációkat követően, együtt az előbbiekkel.

Mielőtt részleteznénk a különböző módszerek, megközelítési módok összekapcsolását, kitérünk arra, hogy az input adatokat milyen meggondolások alapján lehet előállítani, ugyanis a programcsomag többek között olyan, a gyakorlatban nagyon elterjedt kérdőíves adatok feldolgozására is alkalmas, amikor a résztvevők különböző tényezők (tulajdonságok, minőségi jellemzők, kritériumok stb.) szerint *öt osztálybasorolási lehetőség* felhasználásával értékelnek egy vagy több, a vizsgálatba bevont dolgot (objektumot). (Az értékelők és a vizsgált objektumok számának szorzata adja a vizsgálati elemek maximális számát.)

Az automatikus osztályozás bázisa mindig a kiinduló mátrix. Ilyen kiinduló mátrixot mutat be az 1. táblázat, a 8 vizsgálati elemet és 7 tényezőt tartalmazó példa a későbbiekben is szerepel majd.

Az 1. táblázatban feltüntetett kiinduló mátrixot tehát megkaphatjuk úgy, hogy mind a 8 értékelő ugyanazt az objektumot értékeli a 7 tényező szerint, de az is előfordulhat, hogy 4 értékelő 2 objektumot értékel. Gyakori lehet az is, hogy egy-egy kérdőívet tekintünk vizsgálati elemnek, azaz az első kérdőívre adott „válaszok" az előző táblázat szerint sorrendben a következők: 5, 1, 1, 2, 4, 1, 3. Előfordulhat, hogy a „tényezők az értékelők", azaz ekkor 7 értékelő értékel — külön-külön — 8 objektumot, stb. Nagyon lényeges tehát az, ha a kiinduló mátrix adatai diszkrét értékkészlettel rendelkező intervallum skálán

1. táblázat

Osztályozás kiinduló mátrixa

Elem \ Tényező	T—1	T—2	T—3	T—4	T—5	T—6	T—7
E—1	5	1	1	2	4	1	3
E—2	1	2	4	5	2	3	5
E—3	5	1	3	3	5	2	3
E—4	4	5	3	1	3	4	1
E—5	4	2	2	3	5	1	4
E—6	1	3	5	4	1	4	5
E—7	3	5	4	1	2	5	2
E—8	2	3	2	5	1	3	4

adottak, ekkor a mérési skálából adódó lehetőség miatt ugyanazzal az eljárással és közvetlenül osztályozhatók a kiinduló mátrix sorai és oszlopai.

Az alkalmazás előkészítésével kapcsolatban célszerű felhívnunk a figyelmet arra, hogy a jellemzők (tényezők) kiválasztásakor — különösen bonyolult esetekben — speciális szervezési technikák is használhatók. Vizsgálatunk során gazdasági mérnökhallgatók körében az NCM (nominal group technique) módszerrel (KINDLER, 1978) tártuk fel a korszerű döntési módszerek sikerességét befolyásoló tényezőket. Ez kettős célt szolgált. Egyfelől lehetőséget nyújtott az első, kísérletinek tekintett kérdőív módosítására, másfelől pedig azt is lehetővé tette, hogy a hallgatók egységes kulcsfogalmak felhasználásával készíthessenek esettanulmányt egy-egy konkrét döntési probléma megoldásáról.

Az öt osztályhoz a következő minősítéseket rendelhetjük:

Osztály kód	Minősítés
5	kiváló
4	jó
3	közepes
2	megfelelő
1	rossz

Alkalmas az öt osztály bizonyos állítások (tényezők) igaz, vagy hamis voltának megítélésére is (kérdőíves felmérésünk során ezt alkalmaztuk), pl. a következő módon:

Osztály kód	Az állítás jellege
1	az állítás pontosan tükrözi a valóságot
2	az állítás közelítően megfelel a valóságnak
3	ismerete szerint nincs összefüggésben az állítás a valósággal
4	az állításnak inkább az ellenkezője fogadható el
5	az állításnak pontosan az ellenkezője fogadható el

Az öt osztály lehetőséget teremt egyfelől arra, hogy az értékelést szakértők végezzék el, mégpedig kvalitatív tulajdonságok alapján, de alkalmas arra is, hogy egyazon automatikus osztályozás keretében ne csupán kvalitatív, hanem közvetlenül mérhető tényezők is szerepeljenek, amennyiben ilyen adatok sorrendi, intervallum, vagy arányskálán állnak rendelkezésre. Az említett skálákon mért adatokat természetesen át kell alakítani. Ez pl. oly módon történhet, hogy az adatok terjedelme, eloszlása stb. ismeretében egy szakértő, ill. szakértői csoport meghatározza az adott tényező szempontjából megfelelő öt osztályt, majd az egyes adatokhoz rendelt osztály kódja lesz az input adat. Más módozatok is elképzelhetők az osztályozás kiinduló mátrixának előállítására. Pl. elterjedt az a módszer, hogy az adott téma szakembere az egyes vizsgálati elemeket több jellemző szerint, arányskálán rendelkezésre álló adatokból minősíti. Ebben az esetben az eredeti változók számánál jóval kevesebb tényező marad a további vizsgálatokhoz — kérdéses, hogy célszerű-e elveszteni a minősítés során az információ egy részét, bár a szakértők szerint erről szó sincs.

3. Az eljárás áttekintése

Előljáróban ismét érdemes felhívnunk a figyelmet arra, hogy az osztályozás kiinduló mátrixa olyan adatokat (kódokat) tartalmaz, melyek célszerű előkészítés után megteremtik a lehetőséget mind a vizsgálati elemek, mind a tényezők automatikus és hierarchikus osztályozásának. Ezért az eljárásra épülő program a kiinduló mátrix sorai és oszlopai szerint is elvégezheti az osztályozást. A továbbiakban tehát osztályba sorolandó vizsgálati elem alatt az osztályozás kiinduló mátrixának sorait értjük, de emlékeznünk kell arra is, hogy osztályba sorolandó elemnek tekinthetjük a kiinduló mátrix oszlopait is. A kiinduló mátrix **P**, eleme P_{ij}.

A hierarchikus (többszintű) módszer adta osztályozás egy irányított fával reprezentálható, melyet dendogramnak szokás nevezni és amely nagyon jó eszköz a hierarchia és a kritikus szintek illusztrálására.

Az eljárást úgy építettük fel, hogy öt különböző hasonlósági mérték alkalmazására nyíljék lehetőség, továbbá ellenőrizhető legyen a hasonlóság, ill. a dendrogramhoz kiszámított „hasonlósági szint szignifikanciája" is. Természetesen ez közvetett módon történik, mégpedig úgy, hogy illeszkedésvizsgálatot végzünk a hasonlónak tekintett elemekre. A hasonlósági mutatók mindegyikét egy általánosan ismert fogalom, a gyakoriság alapján határozzuk meg. A gyakoriságok kiszámításához, ill. az illeszkedés vizsgálatához egy 5×5 típusú kontingencia táblát alkalmazunk — kapcsolódva az elemeknek előzetes öt osztályba sorolásához az egyes tényezők szerint. Ha a kiindulási mátrix i és j vizsgálati elemét hasonlítjuk össze, akkor ez a következő, 5×5 típusú tábla alapján végezhető el (1. tábla)

A táblában n_{ij} a gyakoriságok számát jelöli, mégpedig attól függően, hogy milyen kapcsolatban áll egymással a P_{ik} és a P_{jk}. Ha pl. $P_{ik} = 1$, $P_{jk} = 1$, akkor ez az n_{11} cellában levő gyakoriságok számát növeli. A P_{ik} és P_{jk} által felvehető értékekhez tehát a következő gyakoriságokat összegező cellaindexeket rendeljük:

$$\text{HA} \begin{cases} P_{ik} = 4, & P_{jk} = 1 \\ P_{ik} = 1, & P_{jk} = 2 \\ \cdot & \cdot \quad \cdot \\ P_{ik} = 2, & P_{jk} = 2 \\ \cdot & \cdot \quad \cdot \\ P_{ik} = l & P_{jk} = m \\ P_{ik} = 5, & P_{jk} = 5 \end{cases} \quad \text{AKKOR A CELLA INDEXE} \quad \begin{cases} 11 \\ 12 \\ \cdot \\ 22 \\ \cdot \\ lm \\ 55 \end{cases}$$

1. tábla

E_i \ E_j	1	2	3	4	5
1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}
2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}
3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}
4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}
5	n_{51}	n_{52}	n_{53}	n_{54}	n_{55}

N számú tényező (ill. a kiindulási mátrix transzponáltjának vizsgálatakor M elem) esetén

$$\sum_{l=1}^{5} \sum_{m=1}^{5} n_{lm} = N.$$

A kontingencia tábla elemeiből (az i és j elem összehasonlítása esetén) a következő összevont gyakoriságok képezhetők:

1) $n_1^{ij} = n_{11} + n_{22} + n_{33} + n_{44} + n_{55}$,
az n_1^{ij} tehát a $P_{ik} = P_{jk}$ feltétel teljesülésének gyakoriságát adja;

2) $n_2^{ij} = n_{12} + n_{21} + n_{23} + n_{32} + n_{34} + n_{43} + n_{45} + n_{54}$,
az n_2^{ij} tehát a $P_{ik} - P_{jk} = \pm 1$ feltétel teljesülésének gyakoriságát adja;

3) $n_3^{ij} = N - n_1^{ij} - n_2^{ij}$,
az n_3^{ij} tehát a $P_{ik} - P_{jk} \geq \pm 2$ feltétel teljesülésének gyakoriságát adja;

4) $n_4^{ij} = n_{13} + n_{24} + n_{35} + n_{31} + n_{42} + n_{53}$;

5) $n_5^{ij} = n_{14} + n_{25} + n_{41} + n_{52}$;

6) $n_6^{ij} = n_{15} + n_{51}$.

Az eljárás alkalmazása esetén felhasználható hasonlósági mértékek az alábbiak (h_{ij}-vel jelölve az i és j elem hasonlóságát, **H**-val az elemek páronkénti összehasonlításával keletkező szimmetrikus hasonlósági mátrixot):

I. $$h_{ij} = \sum_{k=1}^{N} |P_{ik} - P_{jk}| = n_2^{ij} + 2n_4^{ij} + 3n_5^{ij} + 4n_6^{ij}.$$

A hasonlósági mátrix (**H**) legkisebb elemének indexei adják azt a két sorindexet, melynek alapján a leghasonlóbb vizsgálati elemek kiválaszthatók, azaz a hasonlóságot a következő alapján keressük: min $\{h_{ij}, h_{ij}$ eleme **H**$\}$-nak. Ez a hasonlósági mérték egyébként a legegyszerűbbnek tekinthető. Minimális értéke 0 lehet. A súlyozás „enyhén büntető" jelleget ad az összevont gyakorisági celláknak.

II. $$h_{ij} = \sum_{k=1}^{N} (P_{ik} - P_{jk})^2 = n_2^{ij} + 4n_4^{ij} + 9n_5^{ij} + 16n_6^{ij}.$$

Ez a hasonlósági mérték az osztályok közötti eltérések négyzetösszegére épül. Minimális értéke 0 lehet. A hasonlóságot a min $\{h_{ij}, h_{ij}$ eleme **H**-nak$\}$ alapján, azaz a legkisebb négyzetösszeg segítségével határozzuk meg. Ez a hasonlósági mérték az előzőhöz viszonyítva már „erősen büntető" jellegű, mivel a megfelelő összevont gyakorisági cellák súlyai négyzetesen növekednek. Természetesen más súlyozás is elképzelhető.

III. $$h_{ij} = \frac{n_1^{ij}}{n_1^{ij} + n_3^{ij}}.$$

Ez a hasonlósági mérték szintén közismert. Lényegét tekintve arról van szó, hogy figyelmen kívül hagyjuk mind a nevezőben, mind a számlálóban a viszonylag bizonytalan és kis eltéréseket rögzítő cellákat, azaz az n_2^{ij}-t. A hasonlósági mutató max. értéke 1, minimális értéke 0 lehet. Az $n_3^{ij} = 0$ esetén 1 a mutató értéke. A hasonlóságot a max $\{h_{ij}, h_{ij}$ eleme **H**-nak$\}$ alapján keressük.

IV. $$h_{ij} = \frac{n_1^{ij}}{N}.$$

Ez a formula közismert nevén a *Russel–Rao* mérőszám. Értéke szintén 0 és 1 között változik, azonban a nevezőben figyelembe veszi az n_2^{ij}-t is. Ezért a nevezője ennyivel nagyobb az előző mutató nevezőjénél, azaz egy „szigorúbb" mértékről beszélhetünk. A hasonlóságot a max $\{h_{ij}, h_{ij}$ eleme **H**-nak$\}$ alapján keressük.

V. A következő mutató jelentősen eltér mind az I. és II. típusú, mind a III. és IV. típusú mutatótól, mivel az $r_{lm}^{ij} = n_{lm}^{ij}/N$ relatív gyakoriságból indulunk ki a *Shannon* által bevezetett (Farag, 1979) kölcsönös információs mértéken alapuló mutató kiszámításához. Ekkor figyelembe véve, hogy

$$r_{\cdot m}^{ij} = \sum_{l=1}^{5} r_{lm}^{ij}$$

és

$$r_{l\cdot}^{ij} = \sum_{m=1}^{5} r_{lm}^{ij},$$

a hasonlósági mutató a következő:

$$h_{ij} = \sum_{l=1}^{5} \sum_{m=l-1,\, m\neq 0}^{m=l-1,\, m\neq 6} \alpha_{lm} \cdot r_{lm}^{ij} \cdot \log \left\{ \frac{r_{lm}^{ij}}{r_{\cdot m}^{ij} \cdot r_{l\cdot}^{ij}} \right\},$$

$$\text{ahol: } \alpha_{lm} = \begin{pmatrix} 1 \text{ ha } l = m \\ \frac{1}{2} \text{ ha } l \neq m \end{pmatrix}.$$

Legkisebb értéke 0 lehet, amikor ugyan az i és j vizsgálati elem lehet hogy nem független, de a mérték alapján nem hasonló egymáshoz.

Miután meghatároztuk a **H**-t, kiválasztottuk a megfelelő i és j vizsgálati elemet, akkor kerül sor az első ciklus lezárására, az első osztály kialakítására, melyet a dendrogramon az első szint képvisel. Ezt az osztályt a továbbiakban a **P**-ben a z sor képviseli az i és j sor törlése után. Ehhez azonban meg kell határoznunk a P_{zk} értékeket, melyhez az ún. „legtávolabbi szomszéd" elnevezésű technika alapgondolatát felhasználva jutunk el. Azonban nem a hasonlósági mátrixot (távolsági mátrixot) használjuk közvetlenül fel, hanem visszatérünk a P-hez. Meghatározzuk a z sort, töröljük az i és j sort. Erre a redukcióra a következő két döntésfüggvény (redukciós mérték) valamelyikének felhasználásával kerülhet sor, jelölve M-mel a **P** sorainak számát (a döntési kritériumokat is a gyakoriságon alapulónak tekinthetjük az I. és a II. hasonlósági mértéknél bemutatottak szerint):

I. Ha $P_{ik} = P_{jk}$, akkor $P_{zk} = P_{ik}$;

ha $P_{ik} \neq P_{jk}$, akkor $P_{zk} = P_{ik}$, abban az esetben,

$$\text{ha } \sum_{m=1}^{M} |P_{ik} - P_{mk}| \geq \sum_{m=1}^{M} |P_{jk} - P_{mk}|,$$

egyébként $P_{zk} = P_{jk}$, $(k = 1, 2 \ldots N)$

(lásd az I. hasonlósági mértéket!).

II. Ha $P_{ik} = P_{jk}$, akkor $P_{zk} = P_{ik}$;

ha $P_{ik} \neq P_{jk}$, akkor $P_{zk} = P_{ik}$, abban az esetben,

$$\text{ha } \sum_{m=1}^{M} (P_{ik} - P_{mk})^2 \geq \sum_{m=1}^{M} (P_{jk} - P_{mk})^2,$$

egyébként $P_{zk} = P_{jk}$, $(k = 1, 2 \ldots N)$

(lásd a II. hasonlósági mértéket!).

Ezzel zárult le az első ciklus, mivel meghatároztuk a redukált **P**-t. Sorok szerinti osztályozás esetén $M - 1$, oszlopok szerintinél $N - 1$ ciklus szükséges az összes elemnek egy osztályba egyesítéséhez. Az egyes redukciók alapjául szolgáló h_{ij}-k ismeretében, az elemek megfelelő rendezését követően a dendrogram már megrajzolható.

Mivel a RULE–3 automatikus osztályozási eljárás a gyakoriságokra épül, ezért megvizsgálhatjuk, hogy az osztályba sorolás szignifikánsnak tekinthető-e. Tehát abban az esetben, ha valamelyik ciklusnál az i és j elemet válasz-

tottuk ki a h_{ij} alapján, akkor az i és j illeszkedését *Kolmogorov–Szmirnov*-próbával elemezhetjük. A próba előnye, hogy igen kicsiny minták esetében is használható. A kumulált gyakoriságok összegezésén alapul. Először kiszámítjuk a kumulált elméleti gyakoriságokat, majd ezeket összehasonlítjuk a kumulált tapasztalati gyakoriságokkal. A próba azt vizsgálja, hogy ez a legnagyobb eltérés tulajdonítható-e véletlennek. Az elméleti elosztást — a kontingencia tábla alapján — egyenletes eloszlásnak tételezzük fel, tehát ha az i és j sor illeszkedése véletlenszerű, akkor

$$n_{11} = n_{12} = n_{13} = \ldots = n_{55},$$

azaz esetünkben

$$r_{lm} = 1/25 = 0{,}04.$$

A próba elvégzéséhez az n_1^{ij}-t, n_2^{ij}-t és az n_3^{ij}-t használjuk fel:

n_1^{ij}-hez rendelt elméleti gyakoriság: $E_{\mathrm{I}} = 5/25 = 0{,}2$,
n_2^{ij}-hez rendelt elméleti gyakoriság: $E_{\mathrm{II}} = 8/25 = 0{,}32$.
n_3^{ij}-hez rendelt elméleti gyakoriság: $E_{\mathrm{III}} = 12/25 = 0{,}48$.

A kumulált elméleti gyakoriságok:

$$KE_{\mathrm{I}} = 0{,}2 \qquad KE_{\mathrm{II}} = 0{,}52 \qquad KE_{\mathrm{III}} = 1{,}00.$$

Tényleges gyakoriságok:

$$T_{\mathrm{I}}^{ij} = n_1^{ij}/N \qquad T_{\mathrm{II}}^{ij} = n_2^{ij}/N \qquad T_{\mathrm{III}}^{ij} = n_3^{ij}/N.$$

A kumulált tényleges gyakoriságok:

$$KT_{\mathrm{I}}^{ij} = T_{\mathrm{I}}^{ij} \qquad KT_{\mathrm{II}}^{ij} = T_{\mathrm{I}}^{ij} + T_{\mathrm{II}}^{ij} \qquad (KT_{\mathrm{II}}^{ij} =$$

Az osztályok képzésénél figyelembe vett hasonlósági mértékek mellett megadjuk a következő mutatókat is, melyek szignifikanciája az N (ill. M) ismeretében ellenőrizhető, a próba elvégezhető:

$$D_{\mathrm{I}}^{ij} = KE_{\mathrm{I}} - KT_{\mathrm{I}}^{ij}$$
$$D_{\mathrm{II}}^{ij} = KE_{\mathrm{II}} - KT_{\mathrm{II}}^{ij}.$$

A kidolgozott programcsomag segítségével tehát 10 különböző technika (5 hasonlósági mérték, 2 döntésfüggvény) felhasználásával elemezhetjük a kiinduló mátrixot vagy transzponáltját, vizsgálhatjuk a kialakuló osztályokat.

4. A programok ismertetése

Napjainkban reneszánszukat élik a kis- és középgépek, egyre inkább elterjednek a számítógépes hálózatok. A kis- és középgépek alkalmazásának nagy előnye, hogy szoros ember-gép kapcsolat alakítható ki, amely — megítélésünk szerint — nagyon fontos egy adathalmaz (kiinduló mátrix) elemeinek automatikus osztályozásánál. De az is nagyon fontos, hogy a szükséges módosítások kisebb költséget és rövidebb időt igényelnek, a számítógép közvetlenül beépülhet a szervezeti egységek munkájába, az elemzési munkafolyamatba. Az ember-

gép kapcsolat igénye és lehetősége vezetett bennünket akkor, amikor programcsomagunkat a hazai gyártású R—10 és TPA—1140 típusú számítógépekre dolgoztuk ki.

A programcsomag két programból áll. Az első, a STAT nevű, az adatokat mágnesszalagra másolja és ellenőrzés céljából listázza. A program segítségével két szinten egymásba ágyazott részcsoportok is kijelölhetők. A program az egyes részcsoportok végén kinyomtatja az adatok csoporton belüli százalékos megoszlását. A csoportok a dendrogram alapján jelölhetők ki. A program nagyvonalú blokksémáját az 1. ábra mutatja be.

A program blokksémájából kitűnik, hogy futás közben a folyamat konzol üzenetekkel irányítható. Ezt a kapcsolatot úgy alakítottuk ki, hogy a téves utasításokat a program figyelmen kívül hagyja. A mágnesszalagra átmásolt adatokat a RULE—3 törzseljárás programjának segítségével dolgozzuk fel. A program saját futtatási opcióit a futás elején a printeren nyomtatja ki. Az eljárás interaktív módon konzolról irányítható. A program nagyvonalú blokksémáját a 2. ábra mutatja be.

A két program egymással összekapcsolható. A STAT lehetővé teszi a disc-re írt átrendezett adatfile ismételt beolvasását, így a következő feldolgozási ciklus a már átrendezett adatokból indul. Ez eredményezi az eredménymátrixnál a jellegzetes almátrix felbontást. A két program együttes futásának blokksémáját a 3. ábra mutatja be.

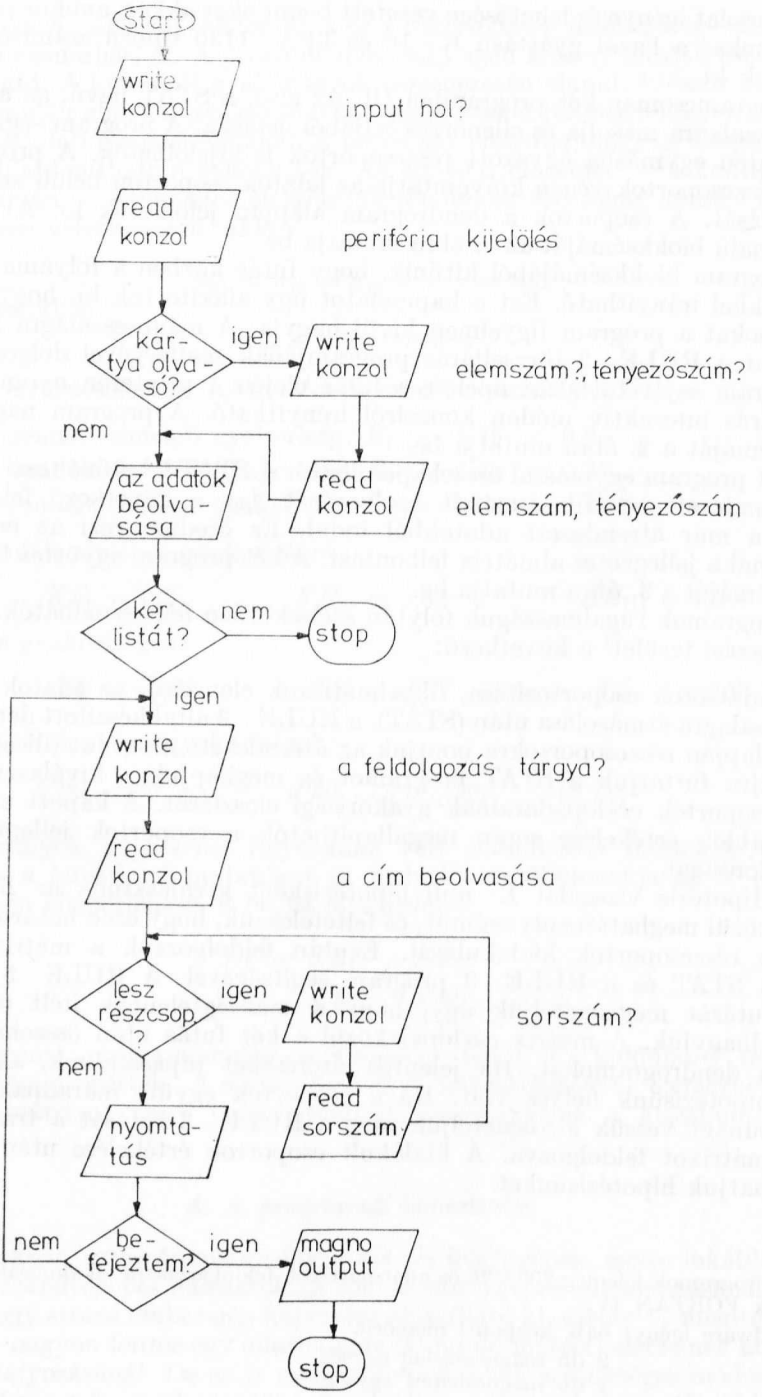
A programok rugalmasságuk folytán széleskörűen felhasználhatók. Néhány alkalmazási terület[1] a következő:

— Adatsorok csoportosítása, ill. almátrixok elemzése: az adatok mágnesszalagra átmásolása után (STAT), a RULE—3 által készített dendrogram alapján részcsoportokra bontjuk az átrendezett mátrixot (disc). Ezután újra futtatjuk a STAT programot és megkapjuk a kiválasztott részcsoportok oszlopadatainak gyakorisági eloszlását. A kapott eredménytablók értékelése során megállapíthatók a csoportok jellemző tulajdonságai.

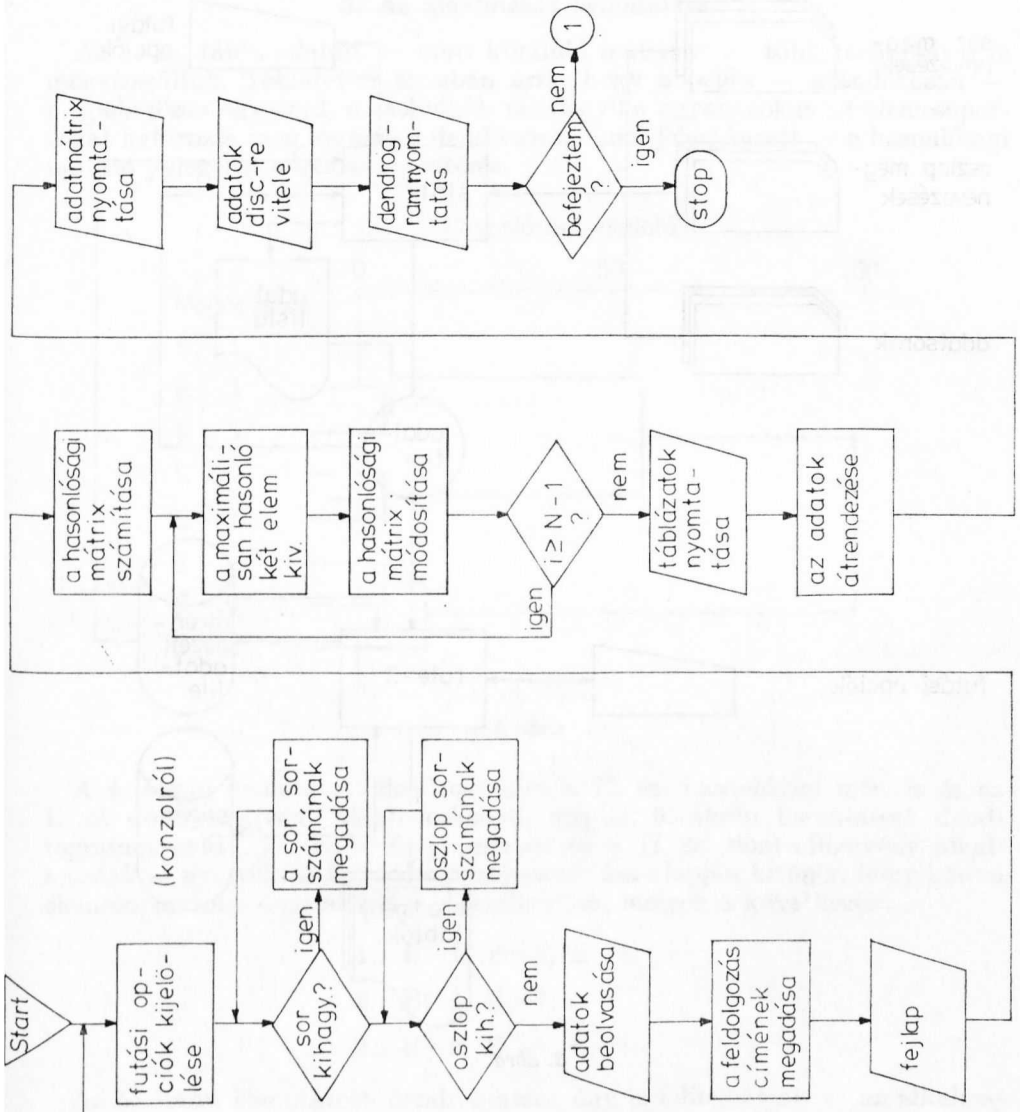
— Hipotézis vizsgálat I.: null hipotézisként kiválasztunk az N tényező közül meghatározott számút, és feltételezzük, hogy ezek határozzák meg a részcsoportok kialakulását. Ezután feldolgozzuk a mátrix adatait a STAT és a RULE—3 program segítségével. A RULE—3 program futását megismételjük úgy, hogy a szükségtelennek ítélt oszlopokat elhagyjuk. A mátrix oszlopai közül a két futás után összehasonlítjuk a dendrogramokat. Ha jelentős eltéréseket tapasztalunk, akkor nullhipotézisünk helyes volt. Ha a csoportok együtt maradnak, feltevésünket vessük el, ismételjük meg a RULE—3 futását a transzponált mátrixot feldolgozva. A kialakult csoportok értékelése után módosíthatjuk hipotézisünket.

[1] A programok jelenleg 200×25-ös adatmátrixok feldolgozására alkalmasak, programnyelvük FORTAN IV.

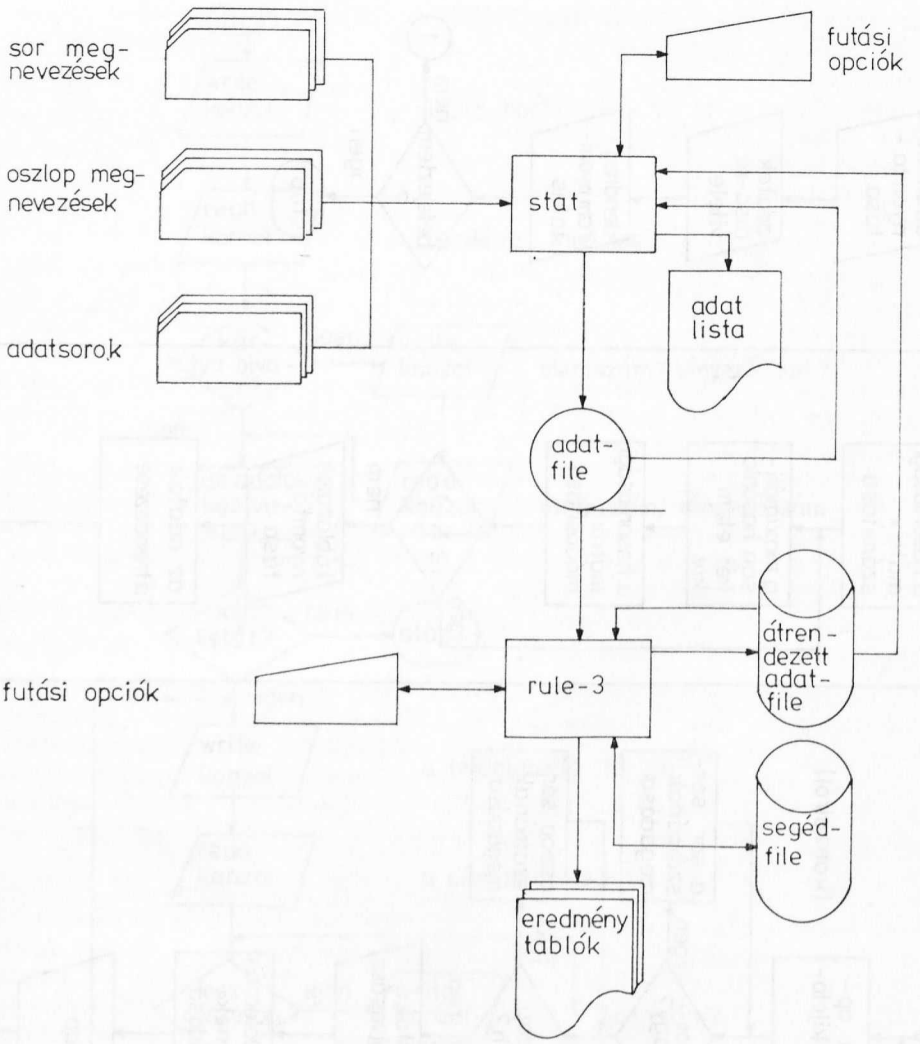
Hardware igény: 64K központi memória
2 db mágnesszalag egység
1 db mágneslemez egység
1 db printer
konzol display


Start
write konzol
input hol?
read konzol
periféria kijelölés
kár-tya olva-só?
igen
write konzol
elemszám?, tényezőszám?
nem
az adatok beolva-sása
read konzol
elemszám, tényezőszám
kér listát?
nem
stop
igen
write konzol
a feldolgozás tárgya?
read konzol
a cím beolvasása
lesz részcsop ?
igen
write konzol
sorszám?
nem
nyomta-tás
read sorszám
nem
be-fejeztem?
igen
magno output
stop

1. ábra


Start
1
futási opciók kijelölése
(konzolról)
sor kihagy?
igen
a sor, sor számának megadása
oszlop kih.?
igen
oszlop, sor számának megadása
nem
adatok beolvasása
a feldolgozás címének megadása
fejlap
a hasonlósági mátrix számítása
a maximálisan hasonló két elem kiv.
a hasonlósági mátrix módosítása
i ≥ N - 1 ?
igen
nem
táblázatok nyomtatása
az adatok átrendezése
adatmátrix nyomtatása
adatok disc-re vitele
dendrogram nyomtatás
befejeztem ?
nem
1
igen
stop

2. ábra

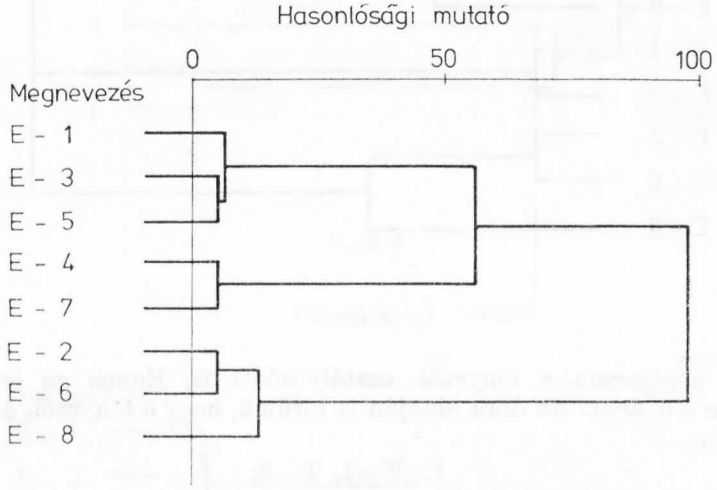


3. ábra

— Hipotézis vizsgálat II.: ha feltételezzük, hogy kiinduló mátrixunk redundáns oszlopokat tartalmaz, a RULE—3 futása után töröljük a RULE—3 segítségével a fölöslegesnek vélt oszlopokat és újra futtatjuk a programot. Ezután összehasonlítjuk a két dendrogramot. Ha a kettő megegyezik, vagy csak „kicsit" tér el egymástól, az elhagyott oszlopok valóban fölöslegesek voltak.

5. Az alkalmazás bemutatása

Az 1. sz. tábla adatait — mint kiinduló mátrixot — több technikával is megvizsgáltuk. Tekintettel azonban arra, hogy a példa — szándékosan — meglehetősen egyszerű, a technikák mindegyike ugyanazokat az elemcsoportokat határozta meg, csupán a dendrogramoknál jelentkezett — a hasonlósági mutató jellegéből adódóan — eltérés.



1. ábra

A 4. ábrán bemutatott dendrogramot a II. sz. hasonlósági mérték és az I. sz. döntésfüggvény alkalmazásával, míg az 5. ábrán bemutatott dendrogramot a III. sz. hasonlósági mérték és a II. sz. döntésfüggvény alkalmazásával nyertük. A két ábra összehasonlítása alapján kitűnik, hogy három elemcsoportról — osztályról — beszélhetünk, melyek a következők:

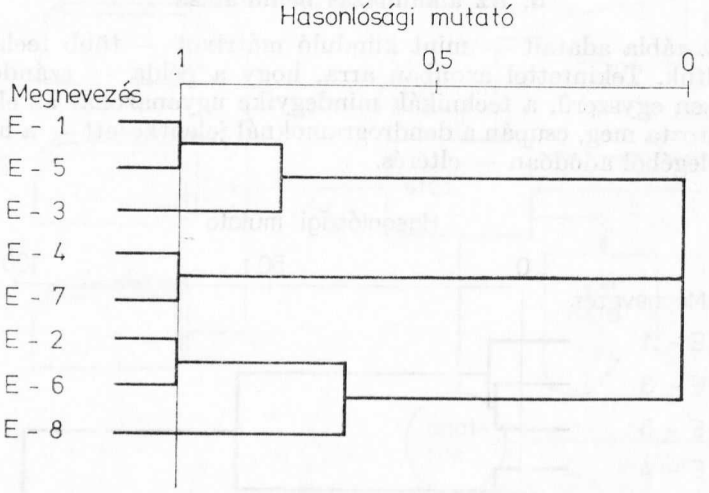
1.: E—1, E—3, E—5;

2.: E—4, E—7;

3.: E—2, E—6, E—8.

Az 5. ábrán bemutatott dendrogramot úgy is felfoghatjuk — az alkalmazott technika sajátosságából adódóan —, mintha a 4. ábrán látható dendrogram azon részét nagyítanánk ki, amely az 5-től 56-ig terjedő tartományban van. Az 5. ábra dendrogramja az osztályokat „automatikusan" határozza meg, bár a *Kolmogorov—Szmirnov*-próba szerint az E—8-nak a 3. osztályba való besorolása már eléggé kockázatos.

Az osztályozás kiinduló mátrixának sorait az 5. ábra dendrogramjának megfelelő rendezésben a 2. táblázat tartalmazza. A táblázatban elhatároltuk egymástól a három osztály sorait. Az azonos osztályba sorolt elemek adatsorainak hasonlósága magáért beszél. A 2. tábla alapján — egy konkrét példa esetén a tényezők (tulajdonságok) ismeretében — az osztályok jellemző ismérvei leírhatók, azonban ezt az elemzést lényegesen megkönnyíthetjük



5. ábra

azzal, ha elvégezzük a tényezők osztályozását is. Ennek az eredményét mutatja be a 6. ábra. Az ábra alapján is kitűnik, hogy a tényezők 4 osztályba sorolhatók:

1.: T—1, T—5;

2.: T—2, T—6;

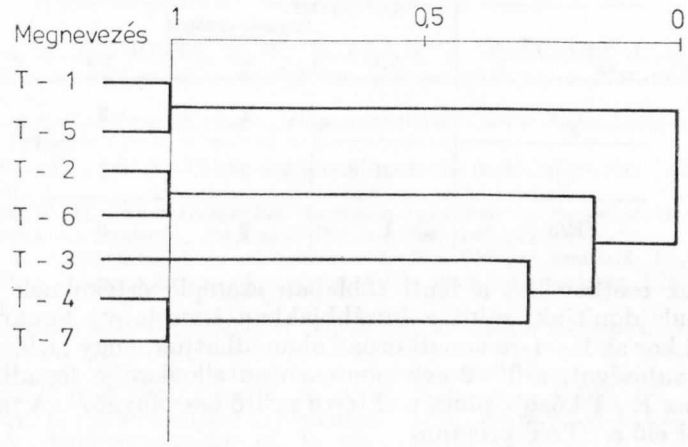
3.: T—3;

4.: T—4, T—7.

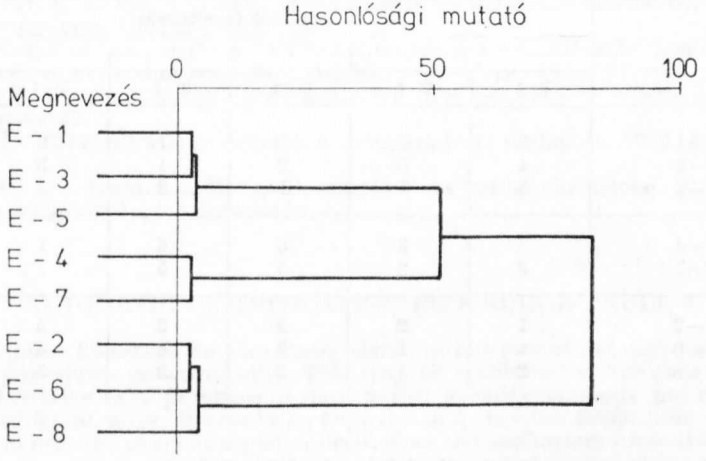
2. tábla

Elem	Tényező						
	T—1	T—2	T—3	T—4	T—5	T—6	T—7
E—1	5	1	1	2	4	1	3
E—5	4	2	2	3	5	1	4
E—3	5	1	3	3	5	2	3
E—4	4	5	3	1	3	4	1
E—7	3	5	4	1	2	5	2
E—2	1	2	4	5	2	3	5
E—6	1	3	5	4	1	4	5
E—8	2	3	2	5	1	3	4

A T—3 besorolását a 4. osztályba a *Kolmogorov—Szmirnov*-próba alapján is el kell vetnünk. A T—3 értékeinek áttekintése, ill. önálló osztályként való szereplése alapján arra a következtetésre juthatunk, hogy a T—3 elhagyása esetén nem változnak meg a kialakult osztályok. Ha osztályozzuk az elemeket



6. ábra



7. ábra

a II. mutató és az I. döntésfüggvény alapján, akkor a 7. ábrát kapjuk. Ha most összevetjük a 4. és a 7. ábrát, akkor megállapíthatjuk, hogy következtetésünk helyes volt, az osztályok és a dendrogram lényegében változatlanok maradtak.

Az eddigiek alapján már osztályokba rendezhetjük a kiinduló mátrix oszlopait és sorait is — ekkor mind az elemeket, mind a tényezőket osztályba soroljuk, azaz „kétdimenziós osztályozásról" beszélhetünk. Mivel az elemek osztályozásánál a T—3 nem játszik szerepet, ezért ezt akár el is hagyhatjuk, mint ahogyan azt a végleges elemzéshez alapuló szolgáló 3. táblánál tettük. A táblán elhatároltuk egymástól az elemek és a tényezők osztályait, hogy ezzel a kialakult almátrixokat kerekített értékkel jellemezhessük, mégpedig a következő módon:

Elemek osztályai	Tényezők osztályai		
	T/1	T/2	T/3
E/1	5	1	3
E/2	3	5	1
E/3	1	3	5

Abban az esetben, ha a fenti táblában szereplő értékeknek ugyanazt a jelentést tulajdonítjuk, mint a korábbiakban ismertetett konkrét vizsgálatunknál, akkor az E—1-re vonatkozóan elmondhatjuk, hogy „a T—1 pontosan tükrözi a valóságot, a T—2-nek pontosan az ellenkezője fogadható el, míg a T—3 és az E—1 között nincs említésre méltó összefüggés". A tábla kialakítását segíti elő a STAT program.

3. tábla

Osztály	Elem	Tényezők és osztályaik					
		1		2		3	
		T—1	T—5	T—2	T—6	T—7	T—4
1	E—1	5	4	1	1	2	3
	E—5	4	5	2	1	3	4
	E—3	5	5	1	2	3	3
2	E—4	4	3	5	4	1	1
	E—7	3	2	5	5	1	2
3	E—2	1	2	2	3	5	5
	E—6	1	1	3	4	4	5
	E—8	2	1	3	3	5	4

A továbbfejlesztés lehetőségei

Az eljárás és a programcsomag használata alapján három fejlesztési irányt körvonalazhatunk:

— Volumen fejlesztés. A jelenlegi 200×25-ös kiinduló mátrixméretnél nagyobb, így 1000×100-as mátrixok feldolgozására is célszerű felkészülni, pl.: pszichológiai, szociológiai, szervezeti vizsgálatokhoz.
— Az alkalmazható hasonlósági mutatók és döntésfüggvények körének bővítése. A jelenlegi 5 féle hasonlósági mértéken felül még nagyon sok fellelhető a szakirodalomban. Minimális fejlesztéssel bővíthető a lehetőségek száma.
— Célszerű lenne olyan adatelőkészítő szubrutinokat is kidolgozni, melyek megkönnyíthetik a skálatranszformációt.

(Beérkezett: 1981. május 27-én)

IRODALOM

1. Attinger, E. O.—Henry, D. T.—Attinger, F. M.—Adams, J. M.—Anné, A.: Biological control hierarchies. IEEE Transactions on Systems, Man and Cybernetics, 1978/1.
2. S. Benedikt, V.—Vári, A.: Egyes clusteranalíziseljárások és gazdasági alkalmazások. Szigma, 1977/3.
3. Dubes, R.—Jain, A. K.: Clustering techniques: the user's dilemma. Pattern Recognition, 1976.
4. Farag, R. F. H.: An information theoretic approach to image partitioning. IEEE Transactions on Systems, Man and Cybernetics, 1978/11.
5. Füstös, L.—Meszéna, Gy.—Simonné, M. N.: Cluster analízis. Szigma, 1977/3.
6. Gower, S. C.: A comparison of some methods of cluster analysis. Biometrics, 1976. december.
7. Kindler, J.—Papp, O.: Komplex rendszerek vizsgálata Budapest, 1977. Műszaki Könyvkiadó.
8. Kindler, J.: A csoportos döntések korszerű módszerei, különös tekintettel a névleges csoport módszerre. BME Ipari Üzemgazdasági Tanszék, 1978.
9. Párniczky, G.: A statisztikai informatika alapjai. Statisztikai Kiadó, 1976.
10. Rényi, A.: Valószínűségszámítás. Tankönyvkiadó, 1973.
11. Ruzsányi, T.: Preferencia, szervezet, döntés. Tudományszervezési Tájékoztató, 1979/5.
12. Ruzsányi, T.: Vállalatok összehasonlító elemzésének módszere. Ipargazdaság, 1980/10.
13. Ruzsányi, T.—Vári, A. (szerk.): A döntéselméleti kutatások és alkalmazások helyzete Magyarországon. OMFB—REI, 1980.
14. Ruzsányi, T.—Lelkes, P.: A RULE-3 eljárás és program. OMFB—REI kézirat, 1980.
15. Schultz—Slevin: Implementing OR/MS. New York, 1975. Elsevier.
16. Sokal, R. R.: Classification: purposes, principles, progress, prospects. Science, 1974. szeptember 27.
17. Sváb, J.: Többváltozós módszerek a biometriában. Budapest, 1979. Mezőgazdasági Kiadó.
18. Yule, G. U.—Kendall, M. G.: Bevezetés a statisztika elméletébe. Budapest, 1964. Közgazdasági és Jogi Könyvkiadó.

AUTOMATIC CLASSIFICATION PROCEDURE "RULE-3"

The technical literature on automatic classification procedures and the number of suggested procedures grows steadily. This may be attributed to the fact that various fields of utilization raise problems of most varied technical contents and of data to be treated in different ways. It should be kept in mind that the application of automatic classification procedures is not a goal in itself, since the application must always be fitted into a concrete process of problem solution, that is input and output of the procedure are largely determined by the users. The elaboration of RULE-3 may be attributed to similar reasons.

The essence of the procedure is the following: columns and rows of the initial matrix of classification are first classified, then the sub-matrices of the rearranged matrix are characterized — expediently — by a (rounded) numerical value. This way a „two-dimensional" classification is produced and a reduced matrix obtained that is suitable for the determination of specific relationships between the classes of observation elements and those of observation parameters. By means of the programme package the initial matrix and its transpose may be analyzed by using 10 various techniques (5 similarity measures, 2 decision functions). The procedure is suitable also for the simultaneous handling of data measured on various scales after an appropriate scale transformation. An example demonstrates the simplicity and many-sided practical utility of the procedure.

МЕТОД АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ РУЛЕ-3

Специальная литература по методам автоматической классификации, вся совокупность разработанных методик все время расширяется, что может относиться за счет того, что в различных сферах использования возникают проблемы, содержащие самые различные специальные аспекты, т. е. данные, к которым нельзя подходить по одинаковому. Во внимание следует принимать и то, что использование методов автоматической классификации не является самоцелью в связи с тем, что само использование каждый раз должно увязываться с конкретным процессом решения проблемы, т. е. вход и выход метода в значительной мере уже определен самыми потребителями. Причина разработки РУЛЕ-3 также может сводиться к — аналогичным вышеизложенному-причинам.

Суть метода заключается в том, что классифицируются колонки и строки исходной матрицы, а потом подматрицы упорядоченной матрицы характеризуются — целесообразно — одним (округленным) цифровым значением, т. к. проводится «двух диапазонная» классификация, в результате которой получаем такую редуцированную матрицу, которая пригодна для определения специфических связей между классами рассматриваемых элементов и классами рассматриваемых параметров. С помощью разработанных пакетов программ и при использовании 10 различных техник (5 зависимостей по аналогии и 2 по принятию решений) можно анализировать и транспонированную исходную матрицу. Этот метод пригоден также и для того, чтобы после соответствующей трансформации одновременно обрабатывать данные, имеющиеся на различных шкалах измерения. Приводимый в статье пример указывает на простоту метода и многогранность его практического применения.