TUDOMÁNYOS ÉLET

Ankét a clusteranalízis hazai alkalmazásairól

A Neumann János Számítógéptudományi Társaság Operációkutatási szakosztálya az MTA Rendszertechnikai Bizottságával és az MTA Statisztikai Bizottságával közösen 1979. április 25-én a Kossuth Klubban *Algoritmizált nomenklatúra-szerkesztés és csoportosítás* címmel ankétot rendezett.

A résztvevők széles körű betekintést nyerhettek az előadók által ismertetett divatos módszer elméleti és alkalmazási kérdéseibe, amely iránt napjainkban óriási az érdeklődés. A clusteranalízis, amely utat nyit a sokváltozós nagy minták áttekinthető numerikus értékeléséhez, számos speciális szakterületen alkalmazható a művelődéskutatástól az orvostudományig, a pszichológiától a területi tervezésig, az életminőségvizsgálatoktól a meteorológiáig . . .

Az ankéton öt előadás hangzott el. A továbbiakban ezek rövid tartalmát ismertetjük.

Az ankét első előadását *Párniczky Gábor* tartotta *Statisztikai osztályozási rendszerek inkompatibilitási problémái* címmel.

A hagyományos statisztikai módszerek körébe tartozó téma ismertetésével a résztvevők a modern clusterezési eljárások mellett bepillantást nyerhettek a klasszikus statisztika műhelyébe is.

Az előadó az osztályozási problémák gyakorlati jelentőségének megvilágítása után a témakör legfontosabb fogalmait definiálta. Az osztályozás kiindulópontját egy véges statisztikai sokaság képezi, melyen különböző logikai függvények adottak, ezeket tulajdonságoknak nevezik. Az osztályozás első lépése egy tulajdonsághalmaz, a nomenklatúra meghatározása. Adott nomenklatúra esetén az elemeket aszerint sorolhatjuk osztályokba, hogy mely tulajdonságokkal rendelkeznek. Az osztályozások legfontosabb csoportját a taxonomikus osztályozások képezik. Ezeknél a keletkező osztályok átfedésmentesek és lefedik a sokaságot.

Az osztályozási rendszer egy szerkezettel ellátott nomenklatúra. A szerkezet legáltalánosabban egy reláció a nomenklatúra elemei között. Hagyományos osztályozási rendszerekben a szerkezet irányított fával reprezentálható, ez hierarchikus szervezettséget jelent. Ilyen rendszerekben különböző osztályozási szintek jönnek létre, ezek közül a legalsó, legrészletesebben szervezett, az adatbázisban még megjelenő szint a terminális nomenklatúra. Osztályozási rendszerekre példa az ITJ, FEOR, stb. A kompatibilitás problémája a gyakorlatban a következőképpen vetődik fel: két különböző információs rendszerben tárolt adatok esetén lehetséges-e az adatok átrendezése az egyik rendszerből a másikba? Ha nem, azaz ha a rendszerek inkompatibilisek, akkor olyan harmonizáló osztályozási rendszert keresünk, mely mindkét rendszerrel kompatibilis. Bizonyítható, hogy ilyen rendszer létezik, azonban még a legszűkebb is (mely a két eredeti nomeklatúra kompozíciója) túl sok elemet tartalmaz ahhoz, hogy gyakorlatilag kezelhető legyen. A harmonizálás gyakorlati lehetetlenségét példák is alátámasztják.

Végül a harmonizálási problémák megoldására tett javaslatot az előadó. Sajnálatos módon ennek a legizgalmasabb kérdésnek a tárgyalására már kevés idő jutott, így a hallgatóság csak a vázlatos gondolatmenettel ismerkedhetett meg. A megoldást egy olyan centrális taxonómiai rendszer nyújtja, melynek középpontjában egy harmonikus tezaurusz áll, logikai szorzat helyett logikai összeggel dolgozik, valamint koordinált indexeléssel működik. Ebben az esetben lehet olyan algoritmust képezni, mely egyértelműen mutat a harmonikus rendszerből bármely terminális rendszerbe.

A második előadó, *Meszéna György* a clusteranalízis hazai alkalmazásának széles skáláját tekintette át, és kérte a jelenlevőket, hogy saját tapasztalataikkal is egészítsék ki előadását.

Bevezetőül megemlítette, hogy az első publikációk a 60-as évek elejétől jelentek meg hazánkban Csibi Sándor, Gulyás Ottó, Révész Pál, Sebestyén Gábor és Fritz József tollából.

A hetvenes évek második felében megjelent ismertetők közül a Szigma 1977. X. évf. 3. számát ajánlotta figyelmünkbe, mivel e számot teljesen a clusteranalízis bemutatására szánták. A módszertani áttekintés és továbbfejlesztés mellettt konkrét gyakorlati alkalmazásokat is olvashattunk benne.

Figyelemre méltó az MTA Szociológiai Intézet kiadványa A clusteranalízis módszerei (1977/1.), amely *Füstös László* és *Manchin Róbert* munkája.

Ezt követően szemelvényeket hallhattunk a gyakorlati alkalmazások területéről.

— *Ruzsányi Tivadar* [1] 17 tőkés cég összehasonlító vizsgálatát végezte el 9 mutató segítségével és a tapasztalatok itthoni felhasználását kereste.
— *Baksay István* és *Ruzsányi Tivadar* [2] hierarchikus clustertechnikával és dendogrammal vizsgálták a vezetők által páronkénti összehasonlítással értékelt 15 kritériumot.
— Az életminőségvizsgálat [3] mellett a művelődéskutatásban [4] közmondások, szólások vizsgálatára is alkalmas a clusteranalízis.
— *Beluszky Pál* vezetésével Borsod-Abaúj-Zemplén megye falusi településeinek tipizálását végezték el faktor- és clusteranalízis együttes alkalmazásával.
— *Andor Csaba* és *Joó András*[6] a társadalomtudományok területén végzett clusteranalízist.

Az előadó ismertette, hogy alkalmazható a clasteranalízis nagyobb számú beruházási alternatíva csoportosítására [7], elősegítve ezzel a megalapozottabb banki döntéseket.

A felsoroltakon kívül szociológiai [8, 9] és agrár jellegű alkalmazásokról is hallhattunk.

Áttérve a számítástechnikai vonatkozásokra, az előadó megemlítette, hogy hazánkban a legjobban kifejlesztett programrendszer a *Szocprog*, amely 18 clusterezési eljárást tartalmaz és az MTA Szociológiai Intézetében fejlesztették ki, CDC 3300 és R 20-as gépen működik.

Végül Meszéna György felhívta a figyelmet arra, hogy a clusteranalízis bármennyire hatékony módszere is a többváltozós statisztikának, nem csodaszer. Eredményes alkalmazhatóságát a feladat jellege határozza meg.

Csicsman József a clusterelemzés KSH-beli felhasználását ismertette. Előadásának bevezető részében a clusteranalízis céljáról, módszereiről és ezek csoportosításáról volt szó, majd a matematikai statisztikai clusterezés ismertetésére tért át. Ennél a megközelítési módnál — szemben az információtudományi clusterezéssel — feltehető, hogy a priori ismeretes a változó eloszlásfüggvénye. A KSH-ban elsősorban a Rubin-Frieman programrendszert és McQueen módszereit alkalmazzák. (Az előbbiről részletesebb információ található a Szigma már említett 1977/3. számában.) A rendszert az IBM bocsátotta közre, IBM és ESZR gépeken futtatható OS operációs rendszerben. Hátránya, hogy maximum 200—250 elemszámú adathalmazra alkalmazható, ez kétlépcsős csoportosítással küszöbölhető ki.

Ezután ismertetőt hallottunk az elmúlt évek érdekesebb alkalmazásairól, melyet a KSH-ban különböző társintézetekkel karöltve valósítottak meg. Katonaköteles fiatalokat vizsgáltak az alkalmasság élettani jellemzői szerint. A területi statisztikai felhasználások közül ismertette Csicsman József hazánk egyes városainak és községeinek fejlettségi színvonal szerinti clusterezését, és egy gazdaságstatisztikai elemzést. Elmondta, hogy ezekben az alkalmazásokban a clusterezés váltakozó sikerrel járt: míg például a kiemelt mezőgazdasági nagyüzemek és iparvállalatok igen jól csoportosíthatók voltak, addig a községek és városok fejlettség szerinti clusterezése, elsősorban a Budapest környéki községek esetében, már kevésbé volt értelmezhető. Az előadás második részében az információ tudományi clusterelemzésről esett szó. A KSH-ban kifejlesztették Futó Péter hipergráfok elméletén alapuló módszerét, mely gyakorlatilag is hatékony eljárásnak bizonyult. A módszer egy igen érdekes alkalmazása adatbázisok tervezésével kapcsolatos, segítségével igen gazdaságos szegmens struktúra határozható meg. Ehhez a témához kapcsolódott *Futó Péter* előadása, melyben a hipergráfok elméletén alapuló clusterelemzés modelljét ismertette. Beszélt az információtudományi clusterelemzés feladatáról, problémáiról (mértékegység, hasonlósági mérték megválasztása . . .)), majd a hipergráfokkal kapcsolatos legfontosabb fogalmakat definiálta. A modell és az ezen alapuló eljárás ismertetése részletesen megtalálható a clusterelemzéssel foglalkozó Szigma számban, így ettől most eltekintünk.

Az előadó a módszer több alkalmazását is ismertette, ezek közül az Építéstudományi Intézet kutatási témáinak clusterezését említjük meg.

Az eljárás számítástechnikai problémáiról szólva elmondta, hogy a futtatásokat IBM 370/145-ös gépen végezték, a legfontosabb tapasztalatuk az volt, hogy a szükséges gép-

idő a feladat struktúrájától függően tág határok között változik, előre nehezen határozható meg.

Az ötödik előadó, *Hunya Péter* a JATE Kibernetikai Laboratóriumában folyó munkákról számolt be. Korreferátumában elmondta, hogy az UNESCO által finanszírozott kutatást folytattak a kutatóhelyek hatékonyságával kapcsolatban. E tudományszociológiai vizsgálat mellett pedagógiai alkalmazásról is beszámolt, amelynek során a tudásszintet, illetve a készségeket mérő teszteket elemezték, és arra kerestek választ, hogy a tesztek belső szerkezete megfelel-e a mérni kívánt jellemzőknek.

Ideg- és elmegyógyászati adatokat is vizsgáltak clusteranalízissel. Önértékelő kérdőíveket töltettek ki a megkérdezettekkel, és elsősorban arra kerestek választ, milyen kapcsolat van a szociális helyzet, a szorongások, a neurózisok, valamint a hangulati élet között. Az adatokon faktoranalízist is végeztek és pszichológiailag jól hasznosítható eredményeket kaptak.

A clusterelemzés viszonylag ritka alkalmazási területe a régészet. Eredményes alkalmazhatóságát bizonyították be az avarkori temetőben talált csontok csoportosításával. A kapott eredmények elősegítették az avarkori társadalom szerkezetének jobb megismerését. Az említetteken kívül egyéb kutatások is folynak a Kibernetikai Laboratóriumban, ezekre azonban az előadó az idő rövidsége miatt már nem tért ki.

Az előadásokat élénk, tartalmas vita követte, melyet *Kunszt György*, az ankét elnöke vezetett. A résztvevők megemlítették a clusterelemzés hazai alkalmazásának néhány alapvető problémáját. Kevéssé ismerjük egymás eredményeit, mivel ezek többnyire csak a legszűkebb szakmai körben válnak ismertté. A különböző szakterületek közötti jobb információáramlás hasznos lenne a párhuzamosságok kiküszöbölésében, és elősegítené a széles körű tapasztalatcserét. A másik probléma a számítógépes programokkal kapcsolatos. Nincs megoldva ezek nyilvántartása, cseréje, hivatalos adásvétele. E kérdések rendezése — mely véleményünk szerint a programkészítők és az Országos Software Archívum és Követőszolgálat közös erőfeszítésével könnyen elérhető lenne — jelentősen meggyorsítaná a clusterelemzés hazai elterjedését.

Az elmondottakat összegezvén úgy véljük, az ankét hasznos volt, a résztvevők érdemi ismeretekkel gazdagodva távozhattak a TIT Múzeum utcai székházából.

Forgácsné Kovács Erzsébet
Kárpáti Zoltán

IRODALOMJEGYZÉK

1. Ruzsányi, T.: Tőkés cégek összehasonlító vizsgálata. Kőolaj és Gázipari Tájékoztató, 1978. 2. sz.
2. Baksay, I.—Ruzsányi, T.: Komplex ösztönzési-értékelési — érdekeltségi rendszer kialakításának gyakorlati-módszertani kérdései.
3. Életminőség modellek (Hankiss, E.—Manchin, R.—Füstös, L.) A magyar életminőségkutatás műhelyéből 9. füzet Tömegkommunikációs Kutatóközpont Kiadványa
4. Gondos, E.—Horváth Gaudi, I.: Clusteranalízis a művelődéskutatásban. Tömegkommunikációs Kutatóközpont VIII. évf. 3. sz.
5. Beluszky, P.: Kutatási jelentés Borsod-Abaúj-Zemplén megye falusi településeinek tipizálásáról MTA Földrajztudományi Kutató Intézet, 1978.
6. Andor, Cs.—Joó, A.: A clusteranalízis és a relációelmélet alkalmazása a társadalomtudományokban. A Tömegkommunikációs Kutatóközpont Kiadványa 1976. VII. évf. 19.
7. Füstös, L.—Meszéna, Gy.—Simonné Mosolygó, N.: Beruházási javaslatok csoportosítása, rangsorolása Szigma, 1977. X. évf. 3. sz.
8. Füstös, L.—Galusi, P.—Manchin, R.: Kísérlet a falusi társadalom szerkezetének sokváltozós empirikus — történeti elemzésére (megjelenés alatt az Akadémiai Kiadónál).
9. Füstös, L.—Galusi, P.—Manchin, R.: A magyar városok urbanizációs típusai (megjelenés alatt az Akadémiai Kiadónál).