Kórházi beteganyag számítógépes-statisztikai elemzése szakértői rendszer kialakításához

BENYÓ ZOLTÁN—BOLLA MARIANNA—TELEGDI LÁSZLÓ—TICK JÓZSEF—BENYÓ IMRE—NAGY PÉTER

DK.: 519.24:681.3:362.11:005.008.4

A modern számítógépes statisztikai elemzések nagy jelentőséggel bírnak a preventív orvosi tevékenységek, valamint a napi gyógyító munka hatásfokának növelésénél. A szerzők célul tűzték ki nagyszámú beteganyag adatainak professzionális számítógépen történő feldolgozását, panaszok, tünetek elvégzett vizsgálatok és laborleletek elemzését, a diagnosztizált betegségek és a beavatkozás módjai közötti összefüggések, kapcsolatok feltárását a Fővárosi Jáhn Ferenc Kórház sebészeti osztályának dokumentumai alapján. A vázolt problémák céladekvát és számítástechnikailag realizálható megoldásához a többdimenziós skálázás, a diszkriminancia-, faktor- és klaszteranalízis új módszereire volt szükség. A feltárt összefüggések alapjául szolgálnak egy kialakítás alatt álló sebészeti tudásbázis alapú szakértői rendszerhez.

Orvosi-élettani kutatásoknál mind gyakrabban vetődik fel valamely matematikai-számítástechnikai alkalmazás igénye [1].

A számítástechnika és a statisztika kiemelt orvosbiológiai alkalmazása lehetővé teszi a betegségek megelőzését célzó, preventív orvosi tevékenységek hatásfokának növelését, a betegségek biztonságos felismerését, valamint a mindennapi gyógyító munkát szolgáló, hatékonyan alkalmazható módszerek kialakítását.

Ilyen irányú tevékenységünk célja egy professzionális személyi számítógépre (IBM-PC) alapozott adatbáziskezelő rendszer és statisztikai eszközrendszer kifejlesztése, valamint ennek segítségével nagytömegű beteganyag adatain számítógépes analízis elvégzése volt.

Célul tűztük ki a panaszok, a tünetek, az elvégzett vizsgálatok és a laborleletek eredményeinek elemzését, a diagnosztizált betegségek és a beavatkozás módjai közötti összefüggések, kapcsolatok feltárását, az összetartozó betegcsoportok és az összevonható változók osztályainak vizsgálatát a Fővárosi Jáhn Ferenc Kórház Sebészeti Osztályának anyaga alapján.

Ehhez a síkbeli megjelenítés, lényegkiemelés és osztályozás — részben általunk kidolgozott — módszereit használtuk fel. Megvizsgáltuk az adatbázisban tárolt információk és a statisztikai vizsgálatok eredményeinek felhasználásával kialakítandó szakértői rendszer megvalósításának lehetőségét is. Ezen kutatások a MTA Országos Tudományos Kutatási Alap (OTKA) támogatásával folynak.

Az alapadatok csoportosítása, szelekciója, jellemzése

A statisztikai feldolgozás során felhasznált a sebészeti adatbázisban tárolt alapadatok statisztikai szempontból három különböző típusba sorolhatók:

— *mennyiségi* típusú változók (mérhető és ennek következtében mérőszámokkal is kifejezhető tulajdonságra, mennyiségre vonatkozók (pl. laborleletek eredményei, életkor stb.);

— *minőségi* típusú változók: nem mérhető és így mérőszámokkal ki sem fejezhető tulajdonságokra, illetve egy esemény bekövetkeztére vagy annak hiányára vonatkoznak. Ezen változók egy fajtája a két értéket felvehető, ún. *dichotom* változó, melynek megfeleltethető egy (0,1) értékkészlettel rendelkező bináris változó (pl. dohányzás);

— *speciális* típusú változók (olyan változók, amelynek értékei WHO vagy BNO kódok).

A vizsgálatok megkezdése előtt el kellett végeznünk az alapadatok szelekcióját. Ez érintette egyrészt a változókat, másrészt az eseteket. Az orvosokkal közösen kiválasztottuk azon BNO, ill. WHO kódokat, melyeket a felvételi diagnózis, anamnézis, opus, ill. előző betegség nevű változók esetében célszerű volt az egyszerűbb kezelhetőség érdekében bináris változóvá átalakítani. Így összesen 286 olyan változó keletkezett, -amelyeken a további vizsgálatok elvégezhetők.

A szelekció másik célja azon esetek, betegek kiszűrése, amelyeknél a már kiválasztott változók hiányosan vagy hibásan voltak kitöltve. Egy erre a célra írt program segítségével az elemzést elvégezve 400 esetet tudtunk a további feldolgozásba bevonni.

Az alapadatok jellemzésekor a kiválasztott 400 esetet gyakorisági táblázatok, kereszt táblázatok, hisztogramok segítségével vizsgáltuk. Az így készített mintegy 120 táblázat és 40 hisztogram az orvosok számára igen sok információt tartalmaz.

A 400 esetből 5 nagy betegségcsoport emelhető ki. Ezen csoportok azonosítására a későbbiekben a végső diagnózisként rögzített BNO kódot használtuk. Az egyes kódok jelentése a következő:

1519 — tumor ventriculi (gyomorrák)
1533 — sigma tumor (szigma bélszakasz rák)
1539 — vastagbél rák
1541 — rectum tumor (végbél rák)
5320 — nyombélfekély

Ezen 5 csoport az esetek 65 százalékát tartalmazza. A többi eset 79-féle különböző diagnózissal rendelkezik, ahol egy diagnózis-csoporthoz 2—3 beteg tartozik.

Szignifikáns eltérést találtunk vizsgálataink során az 5 kiemelt csoportban többek között a

nemek, a vércsoport, a tünetek és a panaszok megoszlásában.

A változók közötti kapcsolatok elemzésének egyszerűbb eszközei

Ezen vizsgálatok elvégzésénél figyelembe kell vennünk az előzőekben már tárgyalt változó típusokat. Ezek között — típusoktól függően — a matematikai statisztika különböző jellegű kapcsolatokat definiál. A fontosabb kapcsolatok a következők:

Mennyiségi—mennyiségi változók kapcsolata

Két mennyiségi típusú (folytonos vagy diszkrét értéket felvevő) változó közötti kapcsolat esetén *korrelációról* beszélünk, melynek legelterjedtebb mérőszáma a *korrelációs együttható* (r).

Minőségi—minőségi változók kapcsolata

Két minőségi típusú változó esetében a köztük lévő kapcsolatot *asszociációnak* nevezzük. Ennek mértékét többnyire a C *kontingencia együtthatóval* fejezik ki.

Mennyiségi—minőségi változók kapcsolata

Egy mennyiségi ós egy minőségi típusú változó kapcsolatának vizsgálata esetén *vegyes kapcsolatról* beszélünk. A vegyes kapcsolatot azzal jellemezzük, hogy a minőségi változó milyen mértékben idézi elő az X mennyiségi változó szóródását. A vegyes kapcsolat mérőszámául a H-val jelölt *szóráshányadost* választjuk.

Az előzőekben megfogalmazott, az adatok minőségétől függő kapcsolatelemző programot — ismereteink szerint — egyik statisztikai programcsomag sem tartalmaz. Ezért készítettünk egy, ezen igényeknek megfelelő, adatvezérelt elemző programot, amely a változók nagy száma miatt egy 286X286 méretű háromszög mátrixban az összes változó lehetséges kapcsolatát adja meg.

Az esetek számából meghatározott szignifikancia szint figyelembevételével csak a lényegesebb összefüggéseket emeljük ki.

Elsőként a felvételi diagnózis, végleges diagnózis, kísérő diagnózis, előző betegségek és a műtét típusok kapcsolatát értékeltük. Ezen területen igen sok az első pillanatban szignifikánsnak tűnő (0,8—0,9) kapcsolat. Alaposabb vizsgálat után azonban kiderül, hogy a jó korrelációs együttható mögött igen kevés (1—6) esetszám húzódik meg. A vizsgált 400 betegből 107 esetet 1519-es BNO kódú (tumor ventriculi) betegséggel kezelték. A többi 293 esetnél 108 féle különböző betegséget diagnosztizáltak. Ezt a csoportot külön választva a többi betegtől a következőket állapítottuk meg:

A tumor ventriculi-val (gyomorrák) kezelt betegek 80 százalékánál kísérő diagnózis nem fordul elő. Amennyiben igen, úgy az egyenletes eloszlást mutat. Érdekesebben alakul az előző betegség kérdése. A betegek 65 százalékánál nem tapasztalható előző betegség. A fent maradó résznél azonban kiemelkedik (20 %) a 4019-es kódú hypertonia (magas vérnyomás), a 1509-es kódú tumor oesophagei (nyelőcső daganat) (10 %) és szintén 10 százalékban az egyszer már kezelt tumor ventriculi.

Figyelemre méltó, hogy míg a többi betegség esetében a férfiak, és nők aránya 44% és 56%, addig a tumor ventriculi esetében ez lényegesen módosul, 70 % és 30 %-ra. A betegek 90 százaléka 60 évnél idősebb, míg az egyéb betegségek esetében ez csak a betegek 70 százalékára igaz. A panasz idejét vizsgálva megállapítható, hogy az esetek felénél a kórházba kerülésig legalább 3 hónap telik el, ez az idő a másik esetben csupán 1 hónap [4].

A nem kód jellegű változók közötti kapcsolatok közül kiemeljük a dohányzás és az alkohol fogyasztás szoros viszonyát, ill. a véres széklet mint panasz és a melena mint tünet kapcsolatát. A változók közötti kapcsolatok vizsgálata során számos összefüggési lehetőséget ki kellett szűrni részben az adatok csekély száma, részben irrealitásuk miatt is.

A továbbiakban a matematikai statisztika lényegkiemelő módszereinek felhasználásával igyekeztünk felderíteni az összefüggéseket.

A változók struktúrájának vizsgálata faktoranalízissel

Megfigyeléseinkkel és ezek korrelációinak vizsgálatával csak a jelenségek megnyilvánulását regisztrálhatjuk, de az összefüggések mögött meghúzódó háttérváltozók rejtve maradnak. Ezek felismerése, számuk meghatározása, számszerű kifejezésük az eredeti változókkal a rendszerről meglévő ismereteink lényegi összefüggéseinek felismerését segíti.

Továbbiakban elsődlegesen a mennyiségi változók feldolgozásával foglalkozunk, a minőségi változók elemzésének ismertetésére később térünk ki.

Mint arra már korábban is kitértünk, az adatbázisban 26 mennyiségi változó található, melyek lényegében a laborleletek eredményeit rögzítik. A teljes betegcsoporton (400 eset) elvégzett analízis eredményeként a program 5 faktort alakított ki. Ezen 5 faktor a rendszerről rendelkezésre álló információ 65 százalékát tartalmazza. Az eredmények arra utalnak, hogy a változók között nincs szoros kapcsolat, struktúrájuk nem tartalmaz lényegesebb „változó csoportosulást". Olyan rendszereknél, ahol a változók információ tartalma egymást jelentősen átfedi, az első vagy az első két faktor már önmagában a rendszer információ mennyiségének 60—70 százalékát tartalmazza. Jelen esetben az aránylag egyenletesen megoszló és viszonylag kis szórás százalék értékkel rendelkező faktorok azt is mutatják, hogy a változók kiválasztása helyesen történt, mindegyik változó saját információjával járul hozzá a rendszer leírásához, vagyis az információ szempontjából kicsi a változók „átfedése".

Az esetek struktúrájának vizsgálata klaszteranalízis segítségével

Most arra keresünk választ, hogy kialakíthatók-e a betegek (esetek) vizsgálata során összetartozó betegcsoportok, ezek milyen közös jellemzőkkel rendelkeznek, ill. mi különbözteti meg őket egy másik csoporttól.

A számítástechnika széleskörű elterjedésével párhuzamosan egyre nagyobb mértékben kezdték alkalmazni a többváltozós matematikai statisztika területén az osztályozási technikákat, a megfigyelt objektumok valamilyen szempont szerinti osztályozását végző módszereket. Az osztályok meghatározása tanulási folyamat eredménye, melynek két fő típusát különböztethetjük meg:

— tanulás tanítóval;
— tanulás tanító nélkül.

A *tanítóval* történő tanulás esetében a gép egy, már előre kiértékelt tananyagot kap, és az osztályba sorolást ezen információ alapján végzi el (diszkriminancia analízis).

A *tanító nélküli* tanulásnál nincs előzetesen megadott tananyag, ott az osztályok kialakítása kizárólag a rendelkezésre álló mintából, valamilyen döntési kritérium alapján történik. Ezt az igen gyakran alkalmazott statisztikai módszert nevezzük *klaszteranalízisnek.*

A klaszteranalízis módszertana igen sokféle megoldást ismer. Ennek oka a döntési kritériumok sokfélesége, az alaphipotézisek különbözősége és a speciális szempontok kielégítésére való törekvés.

A BMDP és az SPSS/PC IBM-PC-re készült program verzióiban rendelkezésre álló, eseteket klaszterező két program közül a minőségi változók alapján történő klaszterezésre egyik sem alkalmas, a mennyiségi változók alapján történő klaszterezésre a hierarchikus klaszteranalízissel szemben a *K-közép* (K-means) programot választottuk.

A K-közép algoritmus a klasztertechnikák nem-hierarchikus csoportjának optimalizáló, iteratív módszerei közé tartozik. A csoportosítás stratégiája, hogy a létrehozott klasztereken belül a változók szórása a lehető legkisebb, míg a csoportok között a legnagyobb legyen, vagyis homogén, de egymástól határozottan elkülönülő csoportok jöjjenek létre. Az eljárás végén minden eset abba a klaszterbe kerül, amelynek a centrumához a legközelebb esik. A klaszter centrumát az adott klaszterhez tartozó esetek átlaga határozza meg.

A klaszteranalízist a teljes betegcsoport (400 eset) adataira úgy végeztük el, hogy a K-közép eljárásnál szokásos módon a klaszterek számát 2-től kezdve fokozatosan növelve alakítottuk ki az osztályokat. Figyelembe véve, hogy az elemzéshez a felhasználó szakember (orvos) számára nem túl nagy számú és nagyjából hasonló méretű osztályok a legkönnyebben értékelhetők, ezért az osztályok számát végül a futtatások alapján 20-ban határoztuk meg. Az egyes klaszterekben található esetszámok alapján megállapítható, hogy az esetek többsége 5 nagy klaszterbe sorolható. A 400 eset közül 365 az 5., 8., 9., 17., 19. klaszter valamelyikében található. A fennmaradó 35 eset, vagyis az összes esetszám 8,7 százaléka pedig a többi 15 klaszterbe került. Ezen klaszterek a valamilyen szempontból szélsőséges esetek „gyűjtő helyei".

Így például a 2. klaszterbe tartozó 1 eset vérsejtsűllyedése nulla, a 15. klaszterben lévő 1 beteg fehérvérsejtjeinek száma nulla, a 4. klaszterbe sorolt 7 beteg SGOT, SGPT, GREAT, AP értékei szintén nullák. A kialakított klaszterek jellemzése a szétválasztásban szerepet játszó változók segítségével történhet.

A klaszterközéppontok, melyek lényegében a klasztert alkotó változók átlagértékei, az öt kiemelt klaszter esetében a következőket mutatják:

1. A Haemoglobin, Haematokrit, Nátrium, Kálium, Vércukor szint, illetve az SGPT enzim és a Fehérjesejtszám az öt klaszternél közelítőleg azonos értéket vesz fel.
2. A Vérsejtsűllyedés a 8-as klaszterben kimagaslóan nagy értékű (a többi 4 klaszter értékének négyszerese).
3. Az Összfehérje száma a szérumban a 17-es klaszter esetében jelentősen csökken (a többi 4 klaszterben szereplő értékek fele).
4. A Carbamid maradók nitrogén szintje a 17-es klaszterben a többi klaszterek átlagának csupán kétharmada.
5. A Creatinin szint a 17-es klaszter esetében igen alacsony értéket mutat (a többi 4 klaszter átlagának mindössze egyharmada).
6. Az SGOT, SGPT ós AP májenzim transzferáz és foszfatáz enzimjei a 17-es klaszterben mindhárom paraméternél igen alacsony értéket vesznek fel. Míg az SGOT és SGPT a többi 4 klaszter értékének kétharmada, addig az AP esetében mindössze 3 százalék. Az AP értéke a 9-es klaszter esetében is igen alacsonynak tekinthető, hiszen az átlagértékek felét sem éri el.
7. A Serum bilirubin szint a 17-es klaszter esetében az átlagértékek kétharmada.

A klaszterközéppontok alapján végzett értékelést a klaszterek szempontjából összegezve megállapítható, hogy az 5-ös és 19-es klaszter, melyekbe együttesen a vizsgált 6 klaszter összes esetszámának 62 százaléka tartozik, kisebb eltérések mellett átlagos paraméter értékekkel rendelkező klaszterek. Ezektől a 8-as klaszter a Vérsejtsűllyedés extrém magas értékével, míg a 9-es klaszter az AP májenzim alacsony értékével különül el. A 17-es klaszter azon eseteket tartalmazza, amelyek paraméter értékei általában kissé vagy jelentősen az átlag alatt maradnak.

Az előzőekben kialakított klaszterek és a végleges diagnózis kódja kapcsolatának elemzésével azt a kérdést vizsgáltuk, hogy homogénnak tekinthető-e a teljes betegcsoport a klaszterek függvényében.

Az 5 fő klaszterben 71-féle diagnózis kód található. Értékelhető esetszámot azonban csak 5 különböző diagnózis kód ér el. Az eredmények azt mutatják, hogy az esetszámok alakulása klaszterspecifikus és egynémely diagnózis esetében jelentősen eltér a várt értéktől. Az eltérések, illetve a klaszter inhomogenitás jellemzésére használható a $!!^2$ érték. A cellák összegeként adódó $!!^2 = 31,55$ érték is mutatja, hogy a klaszterfüggetlenséget feltételező nullhipotézis elvethető.

Ezek alapján a betegek laborleleteik alapján történő csoportosításával előzetes feltételezéseket tehetünk a várható végső diagnózisra. Ennek

értelmében a 8-as klaszterben a 1519-es kódú tumor ventriculi fordul elő a legnagyobb relatív gyakorisággal, míg a 9-es klaszterben az 5320-as kódú betegség (nyombél fekély), a 17-es klaszterben az 1541-es (végbél rák), és a 19-es klaszterben az 1539-es (vastagbél rák) betegség relatív gyakorisága emelkedik ki.

Ezen módszer előnye egy jövőbeni szakértői rendszer kialakításánál jelentős, mivel az előzőekben megadott klaszter középpontokból kiindulva a beteg a labor leletei alapján klaszterbe sorolható, és az adott klaszterben minden betegségkódhoz egy a statisztika alapján megállapított súly rendelhető. Ebből az előfeltételezésből kiindulva a következtetés során a szakértői rendszer célratörőbb stratégiát követhet. Igen lényeges tulajdonsága a módszernek, hogy a minták bővülésével a statisztikai program azonnal elemezi az új minta sorozatot, elvégzi a klaszter középpontok, illetve a klaszteren belüli betegségkód-súlyok dinamikus átértékelését. Ezen öntanuló algoritmus lehetővé teszi egy adaptív modell elkészítését, mely jó hatásfokkal képes alkalmazkodni a mintaként szolgáló a sebészeti nyilvántartó rendszer adatbázisában tárolt beteganyagban bekövetkező rövididejű változásokhoz is.

A dichotom változók vizsgálata

A dichotom változók kutatásának célja hármas volt:

1. a téma elméleti kidolgozása, új, céladekvát és számítástechnikailag realizálható módszerek kialakítása;
2. a megfelelő programok elkészítése és belövése;
3. mikrogépes programcsomag ez alapján történő összeállítása, az eredmények felhasználása a Fővárosi Jáhn Ferenc Kórház Sebészeti Osztályának anyagán.

A kutatás során dichotom változók korrespondencia analízisével és többdimenziós skálázásával, továbbá ilyen változókkal jellemzett objektumok (esetek) klaszteranalízisével, ezen belül is a következő módszerek leírásával és a köztük levő kapcsolatok feltárásával foglalkoztunk:

a) korrespondenciaanalízis, kontingenciatáblák vizsgálata;

b) többszörös korrespondenciaanalízis többkomponensű, komponensenként több értéket felvevő kvalitatív változók analízise;

c) Euklideszi beágyazás dichotom vektorváltozók többdimenziós megjelenítése;

d) Bináris változók faktoranalízise;

e) Többdimenziós skálázás a változók függetlensége esetén;

f) Különböző klaszterezések konszenzusa.

Elméleti eredményeink az alábbiakban foglalhatók össze [2], [3].

a) A korrespondenciaanalízis az utóbbi évtizedben elterjedt többváltozós statisztikai módszer kontingenciatáblák vizsgálatára. A módszerről eddig nem létezett magyar nyelvű leírás. A J. P. Benzécri által szerkesztett francia alapkönyv jelölésrendszere nehézkes, leírásmódja nehezen illeszthető a többváltozós statisztikai módszerek szokásos tárgyalásába. Ezért a korrespondenciaanalízis feladatát a legnagyobb általánosságra törekedve, a feltételes várható érték operátor szinguláris felbontásával írtuk le. A megoldás így lényegében mátrixok szinguláris felbontásával adódik, ami a számítógépes algoritmus alapját képezi.

Megmutattuk, hogy a korrespondenciaanalízis nem más, mint minőségi változók kanonikus korrelációanalízise, és a kapott faktorok a feltételes kárható érték operátorának alacsonyabb rangú vözelítésére legkisebb négyzetes becslést adnak.

b) Egy közönséges kontingenciatábla két minőségi változó kapcsolatát írja le. Több minőségi változó együttes leírására blokkos szerkezetű Burt-táblát alkothatunk, melynek diagonálison kívüli blokkjai közönséges kontingenciatáblák, diagonális blokkjai pedig az egyes változók kategóriáinak gyakoriságait tartalmazzák. A szimetrikus, pozitív szemidefinit Burt-tábla spektrálfelbontásával a változó-kategóriákhoz és az objektumokhoz is a korrespondenciaanalízisben megismert módon faktor-szkórokat rendelhetünk.

Bebizonyítottuk, hogy ezek a szkórok általában nem egyeznek meg azokkal, amelyeket a részkontingenciatáblák korrespondenciaanalízisével kapnánk. Megmutattuk, hogy ez a módszer lényegében véve (a skálázástól eltekintve) ugyanaz, mint a De leeuw által kidolgozott homogenitásvizsgálat, amely objektumoknak és rajtuk megfigyelt kvalitatív változók kategóriáinak előre adott, alacsony dimenziós euklideszi térben való egyidejű ábrázolását teszi lehetővé.

c) Dichotom változók és ilyenekkel jellemzett objektumok kézenfekvő módon írhatók le hipergráfokkal n-nel jelölve a változóknak, ill. a hipergráf nekik megfelelő csúcsainak számát, a változók és az objektumok szimultán klaszterezése céljából Tusnády Gáborral olyan n-nél kisebb k egész számot és ehhez olyan k-particiót kerestünk (a hipergráf k számú komponensre történő felbontását), hogy az egyes komponenseken belül sűrűn haladjanak élek, és a komponensek közel azonos méretűek legyenek. A probléma megoldásához definiáltuk hipergráfok spektrumát, és vizsgáltuk a hipergráf euklideszi térbe való beágyazását. Az optimális k-partíciót megadó iteráció célfüggvénye olyan, hogy a k legkisebb sajátérték összege méri a k-dimenziós beágyazhatóság jóságát (valójában éppen a célfüggvény minimalizálásakor vezettük be a spektrumot definiáló mátrixot). Íly módon a legkisebb sajátértékek alapján tájékozódhatunk k nagyságáról. Ehhez alsó és felső becslést adtunk a k legkisebb sajátérték összegére a hipergráf vágásait és k-szoros összefüggőségét jellemző különböző mérőszámokkal. Ilyen becslés eddig csak közönséges gráfokra és csak a második legkisebb sajátértékre (a legkisebb mindig 0) létezett.

d) Összefüggő komponensú dichotom változók normális háttérváltozókkal történő leírására Tusnády Gáborral bináris faktormodellt dolgoztunk

ki. Eszerint a komponensek egy k-dimenziós standard normális változóra vonatkozóan feltételesen függetlenek, továbbá valószínűségeiket egy, a normális komponensektől és bizonyos paraméterektől függő logisztikus modell írja le. A paraméterek becslésére az EM (expectation and maximization) algoritmus módosításával (random EM, REM algoritmus) eljárást dolgoztunk ki.

e) A korábbi években a dichotom változókkal jellemzett objektumok klaszteranalízisére új eljárást dolgoztunk ki, a többszörös többdimenziós skálázást. Ezt akkor — szándékosan — úgy specifikáltuk, hogy általában nem veszi észre a függetlenséget. Most ez utóbbi esetet külön vizsgáltuk, mégpedig determinisztikusán. Nevezetesen olyan adatmezőt generáltunk, ahol mindegyik változópár (1,1) értékének ugyanaz a gyakorisága. Ekkor az eljárás olyan objektumklasztereket ad, amelynek mellett nem rajzolódnak ki változóklaszterek, és a változóknak megfelelő pontok — mindegyik objektumklaszter mellett — koncentrikus szabályos sokszögek csúcsain helyezkednek el. Ennek segítségével sejtést fogalmaztunk meg a következő feladat megoldására: határozzunk meg a síkon *n* számú pontot úgy, hogy páronkénti távolságaik átlaga adott, szórása minimális legyen!

f) Klaszterek többdimenziós skálázásán alapuló új módszert dolgoztunk ki, amely különböző klaszterezések konszenzusát állítja elő dichotom változók esetén. Ehhez először ugyanazon klaszterei között konstruáltunk távolságokat. Ez alapján végezzük el a klaszterek többdimenziós skálázását, és határozunk meg pontokat az egyes klaszterekhez az alacsony dimenziós euklideszi térben. Ezután az objektumokhoz rendelünk pontokat, és ezeket klaszterezzük. A klaszterek inkonzisztenciája esetén többszörös többdimenziós skálázást hajtunk végre.

Számítógépes megvalósítás

A számítógépes realizáció és felhasználás területén elért eredményeink az alábbiakban foglalhatók össze. Elkészítettük a síkbeli *meg*jelenítés, *l*énygegkiemelés és *osz*tályozás módszereire írt programjainkat tartalmazó, PASCAL nyelvű MELOSZ programcsomag alapjait. A programok többsége dichotom változókra alkalmazható. A MELOSZ tartalmaz adatkezelést, valamint folytonos változók vizsgálatára ismert többváltozós statisztikai módszereket is (klaszter- és kanonikus korrelációanalízis). Lehetőség van az objektumok és változók megjelenítésére, ill. síkbeli ábrázolására a TURBO PASCAL 3,0 grafikájának segítségével. A programok interaktívak, és mind kezdő, mind tapasztalt felhasználók kényelmesen dolgozhatnak velük.

A bejelentkezés után megjelenik a MELOSZ kezdeti menüje, amelyben a következő lehetőségek, ill. almenük közül lehet választani:

1. Adatkezelés.

2. Többdimenziós skálázás/klaszteranalízis. Ezen belül: objektumok halmazának diszjunkt klaszterekre történő particionálása a k-közép módszerrel; objektumok vagy változók távolságmátrixuk alapján történő többdimenziós skálázása; dichotom változók három különböző módszerrel történő többdimenziós skálázása és az eredmény Prokrusztészanalízise; objektumok és őket jellemző dichotom változók szimultán többdimenziós skálázása; többszörös többdimenziós skálázás; konszenzus klaszterezés létrehozása dichotom változókkal jellemzett objektumok különböző klaszterezéseiből; hipergráfok euklideszi térbe történő beágyazása. Ez utóbbi három fő lépésben minimalizálja az előző pontban említett célfüggvényt:

a) rögzítve az egyes klasztereket optimálisan beágyaz (valójában mellékfeltételes korrespondenciaanalízist hajt végre),

b) a csúcsok száma alapján optimalizál a klaszterekben,

c) az egyes objektumokat optimálisan besorolja a klaszterek valamelyikébe.

Tapasztalataink szerint az algoritmus 2—3 iterációs lépés után a célfüggvény lokális minimumára vezet.

3. Korrespondenciaanalízis. Ezen belül: közönséges kontingenciatáblák faktor-, ill. korrespondenciaanalízise, többdimenziós kontingenciatáblákon végrehajtott homogenitásvizsgálat, amely megadja több diszkrét változó együttes alacsonydimenziós kvantifikációját (többszörös korrespondenciaanalízis).

4. Kilépés a rendszerből.

Záró megjegyzések

Szerzők az eddig elvégzett kutató munkájukat nem tekintik befejezettnek. Az elméleti eredményeket, a kidolgozott számítógépes statisztikai vizsgálati módszerek alkalmazhatóságát további beteganyag adatainak feldolgozásával kívánják bizonyítani. A feltárt összefüggések további pontosítása után olyan tudásbázis alapú szakértői rendszert kívánnak létrehozni, ami az orvos számára segítséget jelent napi gyógyító munkája jobb elvégzéséhez.

Beérkezett: 1989. augusztus 1.

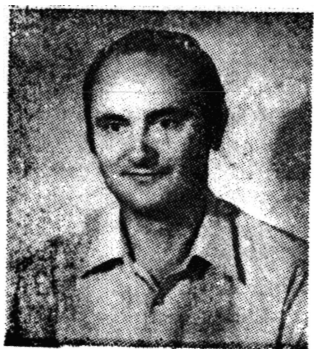
IRODALOM

[l] Z. Benyó: Computer Analysis of Biological Processes. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society New Orleans, November 4—7, 1988 pp. 1461—1463.

[2] *Telegdi L.:* Bináris változók struktúrájának vizsgálata. Alkalmazott Matematikai Lapok 13 (1987—88) pp. 17—42.

[3] *L. Telegdi:* Some Notes on MMDS and the Use of MDS for Detecting Consensus Clusters. CSQ Computational Statistcs Quarterly 4. Physica-Verlag 1989. pp. 267—280.

[4] *Tick J.:* Kórházi beteganyag számítógépes analízise. Egyetemi doktori értekezés, Budapest, 1989.



DR. BENYŐ ZOLTÁN a Budapesti Műszaki Egyetem Villamosmérnöki Kar Műszer és Irányítástechnika szakán 1961-ben villamos-mérnöki oklevelet szerzett. 1976-ben a műszaki tudomány kandidátusa fokozatot ért el. 1961 óta a Budapesti Műszaki Egyetemen dolgozik, jelenleg a Folyamatszabályozási Tanszék docense. | Szakterülete: folyamatszabályozás gyakorlati alkalmazása, digitális adatgyűjtés és feldolgozás, számítógépes szimuláció és orvosi alkalmazása.

BOLLA MARIANNA 1978-ban végzett az ELTE TTK matematikus szakán. Az egyetem elvégzése után 1988-ig az MTA Számítástechnikai és Automatizálási Kutatóintézet Biomatematikai Csoportjában (később Statisztika Osztály) dolgozott, mint tudományos munkatárs, 1988-tól az Országos Ideg- és Elmegyógyintézetben matematikus. Elsősorban matematikai statisztikával (ezen belül is többváltozós statisztikai módszerekkel) és ennek biológiai alkalmazásával foglalkozik (részt vett az Országos Közegészségügyi Intézet megbízásából az újszülöttek veleszületett rendellenességeinek feldolgozásában, jelenleg pedig öngyilkosok, epilepsziás, valamint demens betegek adatainak statisztikaivizsgálatában).

1986-tól az MTA Matematikai Kutató Intézetének levelező aspiránsa. A hipergráfok spektrumára és beágyazására vonatkozó tételeken dolgozik. A számítógépes algoritmus része a Társadalomtudományi Kutató Intézet által forgalmazott DISTAN programcsomagnak. Eredményei hazai ós nemzetközi folyóiratokban, ill. konferencia kiadványokban jelentek meg.

DR. TELEGDI LÁSZLÓ 1970-ben végezte el az ELTE TTK matematikus szakát. Az MTA SZTAKI tudományos munkatársa, majd osztályvezetője volt, 1988-tól a KSH főmunkatársa. 1986 óta a matematikai tudomány kandidátusa Fő kutatási területe az utóbbi években a többváltozós statisztikai analízis síkbeli megjelenítéssel, lényegkiemeléssel és osztályozással foglalkozó módszereinek vizsgálata volt, különös tekintettel a bináris, ill. dichotom (kétértékű vagy -kategóriájú) változók esetére. Részt vesz az „Élettani folyamatok ós szabályozások számítógépes szimulációja, elemzésre többváltozós statisztikai módszerekkel" c. OTKA-feladat teljesítésében.

TICK JÓZSEF 1977-ben a Kandó Kálmán Villamosipari Főiskola számítástechnika szakán, majd 1986-ban a BME Villamosmérnöki kar Műszer és Irányítástechnika szakán szerzett oklevelet. 1978 óta a Kandó Kálmán Villamosipari Műszaki Főiskolán dolgozik, jelenleg a Matematikai és Számítástechnikai Intézet adjunktusa. Szakterülete: számítástechnika alkalmazása az orvostudományban, valamint a szoftver technológia.

DR. BENYÓ IMRE 1954-ben a Budapesti Orvostudományi Egyetemen orvosi oklevelet szerzett. 1954—1980 között az egyetem oktatója, 1974-ben orvostudományok kandidátusa, 1983-ban az orvosdományok doktora tudományos fokozatot szerzett. Fő tudományos területei emésztőszervi megbetegedések preventív orvosi kezelése és gyógyítása. Iskolateremtő kutatási eredményeiről folyamatosan beszámol hazai és külföldi tudományos fórumokon. 1980-tól osztályvezető főorvos a Fővárosi Jáhn Ferenc Kórház Sebészeti Osztályán. Több hazai és külföldi tudományos szervevezet tagja.

DR. NAGY PÉTER 1982-ben kapott orvosi diplomát a SOTE-en. Az egyetemi évek alatt a Kórélettani Intézet TDK-sa, majd demonstrátora volt. 1980-ban az Élettani Világkongresszuson társszerzőként poszterrel szerepelt a bradykinin élettani hahatásának vizsgálatával kapcsolatban. 1981-ben Rectori pályázaton I. díjat nyert egy gyakori szívfejlődési rendellenesség patogenezisének ill. ennek operációjával elért eredmények vizsgálatával, majd a II. Belgyógyászati klinika betegeinek számítógépes adatfeldolgozásánál működött közre.

1982-től a Fővárosi Tanács Jahn F. kórház RI Sebészeti osztályán dolgozik. 1986-ban sebészeti szakvizsgát tett. A kórházban szinte minden területet átfogó számítógépes rendszer áll kidolgozás alatt, e rendszer gyógyászati, diagnosztikai, valamint statisztikai lehetőségeinek kidolgozásában részt vesz. 1989. júniusában a Kandó Kálmán Villamosipari Műszaki Főiskolán Informatikai szaküzem-mérnöki diplomát szerzett. Jelenleg az SZKI segítségével sebészeti szakértői rendszer kifejlesztésén dolgozik.

Benyó, Z., Bolla, M. (Ms.), Telegdi, L., Tick, J., Benyó, I.., Nagy, P.; **Statistical analysis of hospital patients-by computer, for expert system**

Modern statistical analysis by computer are of great significance in preventive medicine and in the increase of the effectivity of daily therapeutic work. The authors are aiming at the processing of data collected from a high number of patients by a professional computer, including the analysis of laboratory findings, complaints, symptoms and accomplished examinations, the connections between diagnosed illnesses and the methods of interference, the revelation of relations, on the basis of the documents of the Surgical Wards of the Budapest Hospital „Jáhn Ferenc". New methods of multidimensional scaling, discriminance-, factor- and claster analysis were needed to solve the problem appropriately. The explored relationships can be the basis for an expert system of surgical knowledge.