Központi Fizikai Kutató Intézet

Módszer folytonos és bináris változókkal leirt minták osztályozására

B. Nagy András és Wolf Tamás

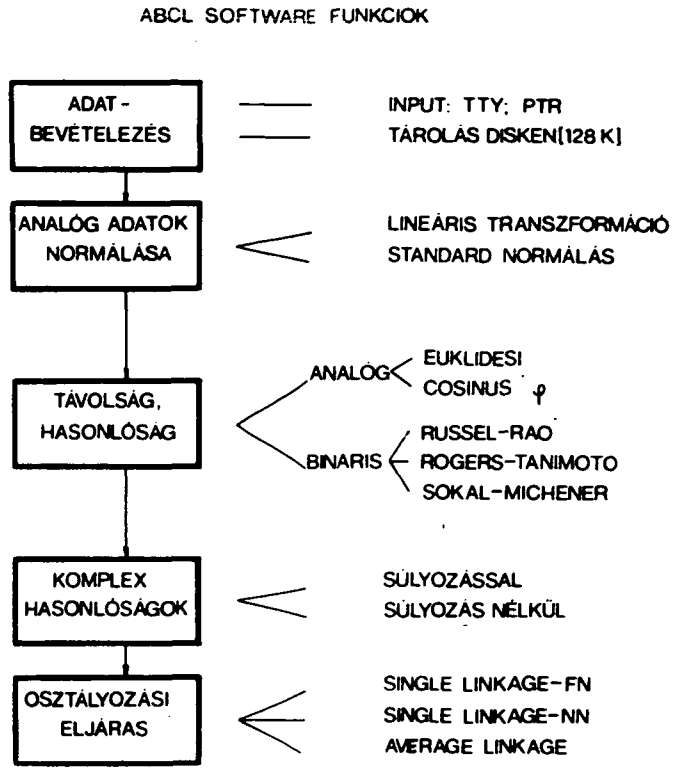
1. Bevezetés

A jelfeldolgozási feladatok jelentős részénél a vizsgált objektum analóg és bináris (általában digitális) jellemzők összességével irható le. A változók - melyek a mérés eredményeként állnak rendelkezésre - többnyire folytonos értékkészlettel birnak, mig a kérdéses objektummal kapcsolatos jegyek, tünetek megléte, illetve hiánya bináris tipusu adatot eredményez. Az ilyen, ugynevezett kevert tipusu változók feldolgozásának evidens példája az orvosi diagnózis alkotási folyamat. Ekkor a klinikai vizsgálatok eredményei (vérnyomás, hőmérséklet, bioelektromos jelek stb.) analóg jellegüek: az anamnézis során az orvos által feltett eldöntendő tipusu kérdésekre, vagy a klinikai kérdőivekre adott IGEN/NEM válaszok mind bináris tipusuak. Más területek szakemberei is gyakran találkoznak kevert változókkal leirható objektumok problematikájával (pl. a meteorológiai prognosztikánál, szociológiai vizsgálatoknál stb.).

A kevert változók feldolgozásánál is gyakran kivánatos lehet valamilyen osztályozó eljárás alkalmazása. Az irodalomból ismert cluster-algoritmusok vagy analóg (1) vagy bináris (2,3,4) változókat kezelnek. Az orvosi diagnosztikai kutatások jelenlegi szakaszában célszerünek látszott a hierarchikus cluster-algoritmusok kiterjesztése a kevert változókkal leirható modellekre.

2. Az ABCL eljárás

A kidolgozott eljárás főbb lépései az 1. ábrán láthatók. Az ADAT-BEVÉTELEZÉS lyukszalagról vagy irógépről történhet, az utóbbi esetben az adatokról lyukszalag készül az ujrafuttatás megkönnyitéséhez. Mivel a programot 16 K szavas központi tárral rendelkező TPAi gépen futtatjuk, az adatok disk-re mentése szükséges ahhoz, hogy megfelelően nagy adatbázist kezelhessünk. A 128 K szavas disk - figyelembe véve a BASIC programozási nyelv számábrázolását és a felügyelő OS/i operációs rendszer helyigényét - mintegy 100 x 50-es analóg és 100 x 50-es bináris adatmátrix feldolgozását teszi lehetővé. A disk felhasználása természetesen a futási idő megnövekedését eredményezi, de a nem rutinszerü alkalmazás időszakában ez megengedhető.



1. ábra

Az ANALOG ADATOK NORMÁLÁSÁ-t kétféle eljárással végeztük. Mindkét esetben az a cél, hogy a vektorok normálása után számított távolság-értékek kompatibilisek legyenek a bináris távolságokkal, amelyek kielégitik a szokásos távolsági kritériumokat (1). Ha az objektumok száma kicsi (néhányszor 10), LINEÁRIS TRANSZFORMÁCIO-val (LT) az adatokat a (0,1) intervallumba transzformáljuk. Amikor a minták száma nagyobb, lehetőség van STANDARD NORMÁLÁS (SN) alkalmazására.

A TÁVOLSÁG számitást az ANALOG változókra, vagy a vektorok EUKLIDESI távolsága vagy COSINUS φ -je alapján végezzük. A BINÁRIS hasonlóság meghatározására három módszer közül választjuk valamelyiket (2,3,4).

Az alkalmazott hasonlósági függvények értékkészlete a (0,1) tartományba esik, igy az analóg és bináris hasonlóságok alapján KOMPLEX HASONLOSÁG-ot számolhatunk. A gyakorlati feladatok többségénél a minták analóg és bináris változóinak száma eltérő. Igy a belőlük számolt analóg, illetve bináris hasonlóságnak is különböző a sulya az objektumok

együttes hasonlóságában. Ezért a felhasználónak lehetősége van a dimenziók figyelembevételére, illetve bizonyos heurisztikus megfontolások eredményeképpen előálló sulyozás beiktatására.

Az OSZTÁLYOZÁSI ELJÁRÁS-ok az együttes hasonlósági elemekből adódó, ugynevezett hasonlósági mátrix elemei alapján osztályoznak. A megvalósitott algoritmus ugyanakkor megengedi csak analóg vagy csak bináris tipusu minták osztályozását is. Ez egyben hasznos segédeszköz lehet annak vizsgálatára, hogy egy objektumhalmaz osztályozásakor mennyivel ad több információt a kevert minták figyelembevétele. A hierarchikus osztályozási eljárások közül - melyek az objektumok eloszlására, illetve az osztályok számára vonatkozóan nem igényelnek a priori információt - három összekapcsolási módszert programoztunk be. Ezek közös tulajdonsága, hogy az osztályozás során mindig a két leghasonlóbb elemet olvasztják össze egy osztályba. Az igy létrehozott uj osztály tulajdonságai azonban a három módszernél eltérőek. A legtávolabbi szomszéd módszernél (Furthest Neighbour: FN) az uj osztály a hozzásorolt minták hasonlóságai közül a kisebbeket tartja meg, a legközelebbi szomszéd módszer (Nearest Neighbour: NN) a nagyobbakat, az átlagos távolság szerinti összekapcsolás módszerénél (Average Linkage) pedig az uj osztály és egy másik osztály hasonlósága a hozzátartozó osztályok és a másik osztály hasonlóságainak átlagával egyenlő (1,2,3). Az egyes clusterezési lépéseknél a program kiirja a kialakult osztályokat, a hozzájuk tartozó minták sorszámait és a hasonlósági szinteket. A program addig fut, mig végül egyetlen objektumot kapunk, amelybe minden minta beletartozik.

3. Az együttes hasonlóság számitása

Az együttes hasonlóság számitása az analóg és bináris hasonlóságokon alapszik. A mért analóg adatokat előzőleg normáljuk.

Jelölje x_{ki} a k-adik minta i-edik elemét,

M a minták számát,

$$x_{imin} = \min_k \{x_{ki}\}$$

$$x_{imax} = \max_k \{x_{ki}\}$$

$$\bar{x}_i = \sum_{k=1}^{M} x_{ki}$$

$$\sigma_i = \sqrt{\frac{\sum_k^M (x_{ki} - \bar{x}_i)^2}{M-1}}$$

és x'_{ki} az x_{ki} normált értékét.

a.) Lineáris transzformálásnál

$$x'_{ki} = \frac{x_{ki} - x_{imin}}{x_{max} - x_{imin}}$$

ekkor

$$0 \leq x'_{ki} \leq 1$$

b.) Standard normálásnál

$$x_{ki} = \frac{x_{ki} - \bar{x}_i}{i}$$

Ebben az esetben a transzformált vektorokra

$$\bar{x}'_i = 0 \quad \text{és} \quad \sigma'_i = 1 \quad \text{teljesül.}$$

A normált adatok alapján számolt analóg hasonlóságokat a 2. ábrán részletezzük.

Az alkalmazott bináris távolság fogalmak értelmezése a 3. ábrán található. A hasonlósági mértéktől megkivánjuk, hogy

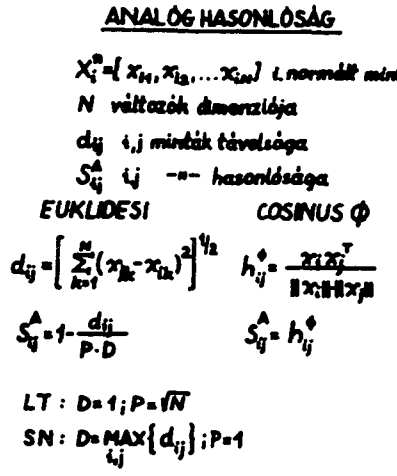
$$0 \leq S^B_{ij} \leq 1 \quad \text{ha} \quad i \neq j \qquad /i/$$

$$S^B_{ij} = 1 \quad \text{ha} \quad i = j \qquad /ii/$$

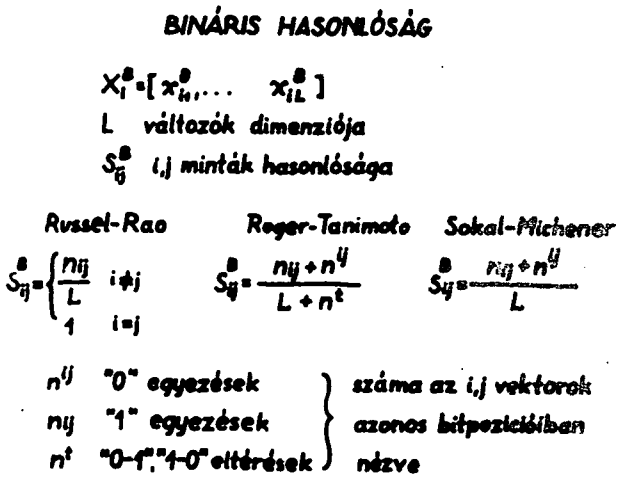
$$S^B_{ij} = S^B_{ji} \qquad /iii/$$

teljesüljön. Az /ii/ feltétel miatt a Russell-Rao hasonlósági függvényt módositottuk

$$S_{ij}^{B} = \begin{cases} \frac{n_{ij}}{L} & i \neq j = 1,2 \ldots M \\ 1 & i = j \end{cases}$$



2. ábra



3. ábra

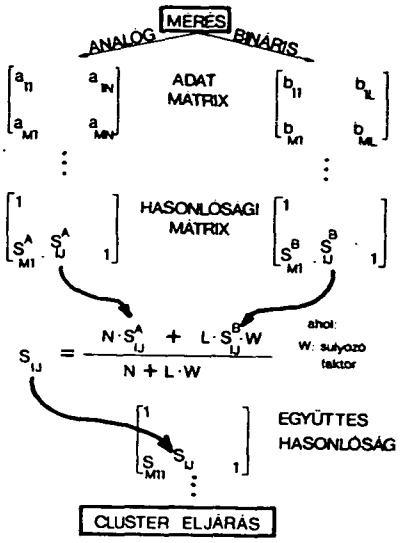
Bár a képletek formailag hasonlóak, lényeges elvi és gyakorlati alkalmazásbeli különbség van közöttük. A Russell-Rao formulát akkor célszerü választani, ha mintákban az IGEN válaszokat tekintjük fontosnak. Másszóval bizonyos tények meglétéből kivánunk következtetni az objektumok tulajdonságaira, mig a tények hiányát nem tekintjük informativnak. A Roger-Tanimoto formula alkalmazásánál az IGEN és NEM válaszok (azaz a bináris minta "1" elemei és "0" elemei) egyforma sullyal jellemzik az objektumot. A hasonlósági képlet két minta eltérő elemeit fokozottan veszi figyelembe. A Sokal-Michener eljárás szintén egyformán jellemzőnek tekinti az IGEN és NEM válaszokat, de a két minta eltérő elemeit nem hangsulyozza a hasonlóság számitásnál.

Az együttes-hasonlósági együtthatók számitását szemlélteti a 4. ábra. A módszer alkalmazásával az /i/, /ii/ és /iii/ feltételek továbbra is teljesülnek.

4. Összefoglalás

A biológiai objektumok általános esetben analóg és bináris változók együttesével irhatók le. A szerzők által javasolt eljárás az analóg és biná-

ris hasonlóság alapján bevezeti az együttes hasonlóság fogalmát, a TPAi kisszámitógépen futtatott ABCL program pedig háromféle bináris hasonlóság számitás és kétfajta analóg távolságfüggvény alapján számitott együttes hasonlósági mátrix elemein három tipusu összekapcsolási módszer szerint képes a biológiai objektumokat osztályba sorolni. Az igy előálló 18 féle clusterezési modell az első tapasztalatok szerint uj ismereteket szolgáltat a biológiai objektumok mélyebb megismeréséhez.



4. ábra

Irodalom

(1) Duran, B.S., Odell, P.L.: Cluster Analysis. A Survey, Springer-Verlag (1974)

(2) Sokal, R.R., Sneath, P.H.A.: Principles of Numerical Taxonomy, San Francisco. W.H. Freeman and Comp. (1963)

(3) Winkel, P.: Numerical Taxonomic Analysis of Cirrhosis.I. Comp. and Biomed.Res. 7, 100-110 (1974)

(4) Rogers, D.J., Tanimoto, T.T.: A Computer Program for Classifying plants, Science, Vol. 132 (oct. 21, 1960)

(5) Young, T.Y., Calvert, T.W.: Classification, Estimation and Pattern Recognition Am. Elsevier, 1974.