STATISZTIKUS ALAKFELISMERÉSI MÓDSZEREK ÉS ALKALMAZÁSUK AZ ORVOSI DIAGNOSZTIKÁBAN

Egyetemi doktori értekezés

Boda Krisztina

Szeged, 1983.

B 2332

József Attila Tudományegyetem Könyvtára
SZEGED

Tartalom

Jelölések

$\underline{\xi} = (\xi_1, \dots, \xi_N)^T$	alak-vektor, tünet-vektor
E_N	N dimenziós euklideszi tér
R	osztályok száma
$C_i, (i = 1, \dots, R)$	az i-edik osztály v. alak
$\underline{\xi} \sim C_i$	a $\underline{\xi}$ -vel jellemzett objektum a C_i osztályba tartozik
$B_i, (i = 1, \dots, R)$	az egy osztályba tartozó objektumoknak megfelelő E_N-beli halmaz
D_i	döntési tartomány
$d = d(\underline{x})$	döntésfüggvény
$d^*(\underline{x})$	Bayes döntésfüggvény
p_i	a C_i osztály a priori valószínűsége
$f_i(\underline{x})$	a $\underline{\xi}$ C_i osztályra vonatkozó feltételes sűrűségfüggvénye
$f(x)$	a $\underline{\xi}$ sűrűségfüggvénye
$\hat{p_i}(\underline{x})$	a C_i osztály a posteriori valószínűsége
c_{ij} $i \neq j$	annak a hibás osztályozásnak a költsége, ha $\underline{\xi} \sim C_i$ és $\underline{\xi} \sim C_j$-t döntjük
$L_i = L_i(d)$	a $d(\underline{x})$ döntésfüggvényhez tartozó feltételes veszteség
$L = L(d)$	a $d(\underline{x})$ döntésfüggvényhez tartozó kockázat
$P_e(d^*)$	a Bayes döntés hibavalószínűsége
$d_{jk}(\underline{x})$	a C_j és C_k osztályt elválasztó felület
$\underline{\mu}_i$	a $\underline{\xi}$ várható értéke a C_i osztály esetén
$\underline{\Sigma}_i$	a kovariancia mátrix a C_i osztály esetén
$(\underline{\xi}_1, \vartheta_1), \dots, (\underline{\xi}_n, \vartheta_n), \dots$	tananyag

ϑ_n	tanítás
$d^{(n)}(\underline{x})$	a döntésfüggvény közelítésére alkalmazott függvénysorozat
$\underline{x}_i^{(j)}=(x_{i1},\dots,x_{iN})^T$, $i=1,\dots,n_j$, $j=1,\dots,R$	a j-edik osztály i-edik pontja a tananyagban
n_j	a j-edik osztályhoz tartozó tanulópontok száma
$\underline{m}^{(j)}$	a $\underline{\mu}_j$ becslése, átlagvektor a j-edik osztályban
$\underline{S}$	a $\underline{\Sigma}$ kovariancia mátrix becslése
log	e alapú logaritmus
I	egységmátrix
E(.)	a várható érték

1. Bevezetés

Az alakfelismerés olyan módszerek, eljárások összefoglaló neve, amelyek segítségével bizonyos objektumok jellemzőik alapján adott kategóriák valamelyikéhez egyértelműen hozzárendelhetők. A kategóriákról (osztályokról) feltesszük, hogy eleve (a priori) léteznek. Az objektumokat általában nem tudjuk direkt módon azonosítani, hanem jellemzőik (mérések, megfigyelési adatok) alapján.

A probléma a karakterek felismeréséből indult ki. Itt egy kategóriát egy adott karakter "elméleti" képe jelent, a jellemzők megállapítása pedig egy karakter képének valamilyen módon történő digitalizálásával lehetséges. Ma már számos tudományágban használják az alakfelismerési módszereket: képek felismerése kapcsán pl. kromoszómák azonosítására (sejt-szövet analízis), ujjlenyomatok (biológia, kriminalisztika), hangfelismerés, görbék osztályozására (EEG, EKG). Több ezer cikk tanúskodik az alakfelismerés orvos-diagnosztikai alkalmazásairól: itt a feladat egy paciens tünetei alapján a betegség felismerése.

E dolgozat célja egyes statisztikus alakfelismerési módszerek - különösen a lineáris diszkriminancia analízis és általánosításainak - áttekintése, valamint a módszerek orvos-diagnosztikai felhasználási lehetőségeinek bemutatása a külföldi és hazai irodalom áttekintése révén és saját esettanulmányok alapján.

Az alakfelismerés alapproblémája jellemzőik alapján

objektumok hozzárendelése adott kategóriák valamelyikéhez. Egy objektumot N számú jellemzője alapján egy N dimenziós vektorral adhatunk meg. A hozzárendelést általában úgy adják meg, hogy a téves döntés valószínűsége, az ún. hibavalószínűség minél kisebb legyen. A minimális hibavalószínűségű döntést Bayes döntésnek nevezzük. Az első fejezetben definiáljuk a döntésfüggvényt, amelynek segítségével egy vektor hozzárendelése elvégezhető, és megadjuk a Bayes döntésfüggvény alakját általánosan és többváltozós normális eloszlás esetén.

A második fejezetben a statisztikus alakfelismerés modelljét tárgyalva megmutatjuk, hogy a döntésfüggvény megadásához szükség van az ún. tananyagra, amely nem más, mint egy olyan statisztikai minta, amelyben minden egyes mintaelemnek ismerjük a kategóriáját. A feladat első részeként meg kell határoznunk a döntésfüggvényt (tanulás), majd a kapott döntési szabályt kell alkalmaznunk ellenőrzésre (tesztelés), majd ezután ismeretlen kategóriájú egyedek osztályozására. Kitérünk a tananyag méretének megválasztásával kapcsolatos problémákra, azaz arra, hogy a jellemzők számát úgy kell meghatározni, hogy minél kevesebb számú jellemző minél több információt hordozzon. Az alakfelismerési feladat része lehet egy ún. lényegkiemelési transzformáció, amely az eredeti adatokat a legkönnyebben osztályozható formára hozza.

Az alakfelismerésben eddigi ismereteink szerint nincsenek "mindenható" módszerek, egy adott feladat esetén

gyakran keresni kell azt a döntésfüggvényt, amely a legjobb osztályozást adja. A harmadik fejezet így számos statisztikus alakfelismerési módszert ismertet és utal e módszereknek a Bayes döntéshez való kapcsolatára. Megadjuk Anderson (1974) alapján a Bayes döntésfüggvényt normális eloszlás esetén a paraméterek becslésével, amely egyenlő kovariancia mátrixok esetén a következő alakú:

$$d(\underline{x}) = (\underline{m}^{(1)} - \underline{m}^{(2)})^T \underline{S}^{-1} \underline{x} - \frac{1}{2}(\underline{m}^{(1)} + \underline{m}^{(2)})^T \underline{S}^{-1} (\underline{m}^{(1)} - \underline{m}^{(2)}) \qquad (1)$$

ahol $\underline{m}^{(1)}, \underline{m}^{(2)}$ a két osztály átlagvektorai, $\underline{S}$ pedig a közös kovariancia mátrix becslése. Igen egyszerű és szemléletes elv alapján működik a legközelebbi szomszéd módszer (Cover és Hart, 1967), amely egy adott tananyag esetén egy ismeretlen kategóriájú pontot a hozzá legközelebb eső mintapont osztályába sorol be. Különféle lineáris és szakaszonként lineáris algoritmusokkal zárjuk az ún. tanítóval való tanulás módszereinek bemutatását.

A lineáris diszkriminancia analízis néven emlegetett alakfelismerési módszerekkel foglalkozik a harmadik fejezet. Bemutatjuk, hogy a sűrűségfüggvény-becsléskor kapott (1) formulához egy optimalizálási transzformáció útján is eljuthatunk (Fisher, 1936). Ennek általánosításaként ismertetjük a Day-Kerridge (1967), illetve J. A. Anderson (1972) által tárgyalt logisztikus diszkriminancia analízist, amely nemcsak normális, hanem diszkrét eloszlású valószínűségi változókra is alkalmazható.

Attila Tudományegyetem SZEGED

Ezek további változata az Albert (1981) által közölt probit diszkriminancia analízis.

A negyedik fejezet a módszerek orvos-diagnosztikai alkalmazási lehetőségeiről ad áttekintést az általunk feldolgozott nagyszámú nemzetközi és hazai közlemény alapján. Saját tapasztalatainkat az újszülöttkori sárgaság, a hasnyálmirigy gyulladás és a gyermekkori légúti megbetegedések diagnosztizálása kapcsán mutatjuk be, melyeket a SZOTE Gyermekgyógyászati Klinika és az I. sz. Belgyógyászati Klinika orvosaival együttműködve végeztünk, illetve végzünk ma is. (Boda, 1971., 1972, Boda és Pap , 1983).

A hasnyálmirigy gyulladás diagnosztizálásakor a feladat az volt, hogy egy paciens bizonyos enzim értékeinek ismeretében következtessünk a betegség fennállására. A problémát itt az jelentette, hogy egy beteg embernek is csak bizonyos enzim értékei mutatnak kóros elváltozást, míg más enzimek normálisak maradhatnak. Ezért nem lehet egyetlen enzim alapján képet alkotni a hasnyálmirigy működéséről, és ezért kézenfekvő az egyszerre több jellemzőt figyelembe vevő alakfelismerési módszerek alkalmazása. Különböző előkészítő tevékenységek és több alakfelismerési módszer lefuttatása után a lineáris diszkriminancia analízis adta a legjobb, 93 %-os helyes döntést, melyet az utólagos tesztek is megerősítettek. Ezek az eredmények orvosi szempontból is értékesek, és a diagnosztikai munkában segítséget nyújtanak az orvosnak.

2. Az alakfelismerés alapproblémája, módszerei

2.1. Az alakfelismerés alapproblémája

Az alakfelismerés témakörébe a gyakorlatban az olyan feladatok tartoznak, amelyekben korábban, tapasztalati úton megadott esetek, objektumok, helyzetek újabb előfordulásának felismerésére kell eljárást konstruálnunk. A felismerés úgy történik, hogy az objektumot adott kategóriák valamelyikébe besoroljuk. A kategóriákról feltesszük, hogy véges számúak és létezésük eleve, a priori adott. Az objektumokról is feltesszük, hogy valamely kategóriához tartoznak. A probléma ott merül fel, hogy az objektumokat nem lehet általában direkt módon azonosítani, csak a róla végzett megfigyelések, mérések alapján. A besorolás egy döntési szabály segítségével történhet, amelyet előzetes megfigyelések alapján alkothatunk meg.

Tekintsük az orvosi diagnózis felállításának problémáját: az orvosnak egy vizsgálatra jelentkező paciensről a panaszai, tünetei, laborvizsgálatok, stb. alapján el kell döntenie, hogy milyen betegsége van. Az azonosítandó objektum tehát a paciens, a kategóriák pedig a felmerülő lehetséges betegségek. A betegségek száma és a létezése ismert. Az azonosítás azonban csak a paciens vizsgálata után, a mérési eredmények, megfi-

gyelések birtokában lehetséges. A gyakorlatban az orvos ismereteire és tapasztalataira támaszkodik a döntésnél, ezek segítségével ismeri fel a betegséget a pillanatnyi mérési és megfigyelési eredmények alapján.

Látható, hogy az orvosi diagnózis felállítása pontosan úgy formalizálható, ahogyan az alakfelismerési probléma.

A továbbiakban megadjuk az eddig vázolt probléma matematikai tárgyalását.

Legyen adott egy (Ω, A, P) valószínűségi mező, és $C_1, \dots, C_R$ ezen egy teljes eseményrendszer, azaz

$$C_i \subset \Omega; \quad C_i \neq \emptyset \qquad i = 1, \dots, R;$$

$$\bigcup_i C_i = \Omega; \qquad C_i \cap C_j = \emptyset, \quad i \neq j, \quad i, j = 1, \dots, R.$$

A C_i-ket <u>osztályoknak, kategóriáknak</u> vagy hipotéziseknek nevezzük. Legyen adott továbbá objektumoknak egy halmaza, amely objektumok a priori valamely kategóriához hozzátartoznak; mi azonban nem tudjuk, melyikhez, mivel ezekről az objektumokról csak közvetett adatok, mérések, megfigyelések állnak rendelkezésre. Egy objektumot így egy valószínűségi vektor-változóval jellemzünk, amelyről feltesszük, hogy az E_N N dimenziós euklideszi tér egy vektora:

$$\underline{\xi} = (\xi_1, \dots, \xi_N)^T$$

A $\underline{\xi}$ szokásos elnevezései: <u>megfigyelés, tünet-vektor, alakvektor</u>.

Az alakfelismerési feladat a $\underline{\xi}$-vel jellemzett objektum hozzárendelése a C_i kategóriák valamelyikéhez.

Egy objektumról tett méréssel tulajdonképpen az

$$\Omega \to E_N \qquad (2.1)$$

leképezést hajtottuk végre, azaz egy adott $\omega \in \Omega$-hoz egy $\underline{\xi} = \underline{\xi}(\omega)$ valószínűségi vektorváltozót rendeltünk hozzá.

Feltesszük, hogy ezzel a leképezéssel olyan $B_1 \ldots, B_R$ halmazok állnak elő E_N-en, hogy $\omega \in C_i$ akkor és csak akkor teljesül, ha $\underline{\xi}(\omega) \in B_i$ $(i = 1, \ldots, R)$.

2.1. Megjegyzés. Könnyen belátható, hogy ekkor a B_i halmazok is diszjunktak, azaz

$$B_i \cap B_j = \emptyset \quad , \quad i \neq j \quad , \quad i, j = 1, \ldots, R .$$

Ha ugyanis lenne olyan $\underline{\xi}$, amely $\underline{\xi}(\omega) \in B_i \cap B_j$, akkor

$\underline{\xi}(\omega) \in B_i$ és $\underline{\xi}(\omega) \in B_j$ miatt $\omega \in C_i$ és $\omega \in C_j$ lenne, ami ellentmond annak a feltevésnek, hogy a C_i-k diszjunktak.

E transzformációt figyelembe véve jelölje

$$\underline{\xi} \sim C_i$$

azt a hozzárendelést, hogy a $\underline{\xi}$ -hez tartozó objektum a C_i osztályba tartozik, azaz $\omega \in C_i$.

A $B_1, \ldots, B_R$ halmazok alapján az alakfelismerési feladat, egy ismeretlen osztályú objektum adott kategóriához való rendelése megoldható oly módon, hogy $\underline{\xi} \in B_i$ esetén azt mondjuk, hogy $\underline{\xi} \sim C_i$, i= 1, ...,R.

A gyakorlatban a B_i halmazok megkeresése nem köny-

nyű feladat. Ismert kategóriájú objektumokról nyert mérések (az ún. tananyag, pontos definiciója 2.2-ben) szolgáltatják az alapot ehhez. Ha a mérések diszjunkt halmazokat eredményeznek az euklideszi térben, nehéz a nagy dimenziószám és a nagy elemszám miatt e halmazokat megkeresni.

Az alakfelismerési feladat megoldása a gyakorlatban úgy történik, hogy valamilyen módon (a tananyag alapján) az euklideszi téren $D_1,\ldots, D_R$ diszjunkt részhalmazokat, úgynevezett döntési tartományokat hozunk létre, és $\underline{\xi} \in D_i$ esetén a $\underline{\xi} \sim C_i$ döntést hozzuk (i=1,...,R).

2.1. Definició. Egy olyan mérhető valós $d(\underline{x})$ függvényt, amely az euklideszi teret az $\{1,\ldots,R\}$ halmazra képezi le egyértelműen, döntésfüggvénynek nevezzük:

$$d(\underline{x}): E_N \rightarrow \{1,\ldots,R\}$$

Az alakfelismerési feladat megoldását adott döntésfüggvény ismeretében a következő döntési szabály adja: a $d(\underline{\xi}) = i$ esetén a $\underline{\xi} \sim C_i$ döntést hozzuk. (i=1,...,R).

2.2. Megjegyzés. A döntésfüggvény és a döntési tartományok kapcsolatát a következő összefüggések szemléltetik:
a/ Adott $d(\underline{x})$-hez a döntési tartományok:

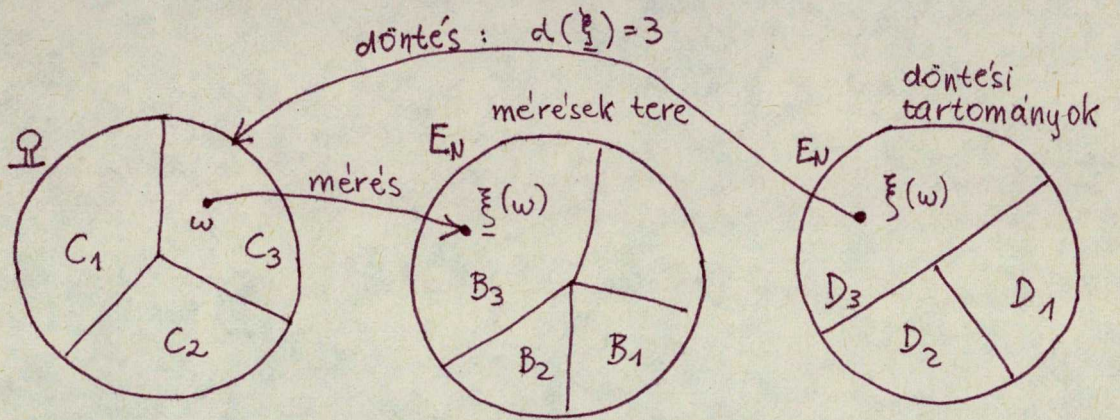
$$D_i = \{\underline{x} \mid \underline{x} \in E_N,\ d(\underline{x}) = i\} \qquad i = 1,\ldots,R \quad (2.2)$$

b/ Adott döntési tartományokhoz a $d(\underline{x})$:

$$d(\underline{\xi}) = i, \text{ ha } \underline{\xi} \in D_i \qquad i = 1,\ldots,R. \quad (2.3)$$

Az objektumok, a megfigyelések és a döntési tartományok kapcsolatát szemlélteti a 2.1. ábra

A döntésfüggvényt - vagy a döntési tartományokat - úgy igyekeztünk meghatározni, hogy a döntésük a lehető legjobban megegyezzen az objektum eredeti kategóriájával. Éppen ebben rejlik a gyakorlati alkalmazás számos problémája, egyrészt az E_N megfigyelési tér, másrészt a döntésfüggvény olyan megválasztására törekszünk, ahol a döntési tartományok a lehető legjobban megközelítik a megfigyelések E_N-beli csoportjait.



2.1. ábra. Az objektumok, a róluk készített mérések és a döntési tartományok kapcsolata

Eddig lényegében az úgynevezett determinisztikus modellt ismertettük: egy adott E_N-beli pontról a B_i-k ismeretében egyértelműen megadható, hogy melyik osztályhoz rendelhető hozzá.

Sok gyakorlati esetben a mérések alapján az E_N-ben az egy csoportba tartozó pontok halmazai nem diszjunktak. Ez úgy lehetséges, hogy a mérésekkel nem kaptunk meg

minden információt, vagy kevés jellemzőt mértünk, vagy pedig nem elég informatív jegyeket. Ugyancsak nem kapunk szükségképpen diszjunkt B_i halmazokat, ha a (2.1)-beli $\Omega \to E_N$ leképzésről nem tesszük fel, hogy kölcsönösen egyértelmű. Ekkor a B_i halmazok közötti átfedés a véletlentől függ. Egy E_N-beli pontról nem állíthatjuk egyértelműen, hogy melyik osztályhoz tartozik, csak bizonyos valószínűséggel. A döntésünk tehát hibával jár. Ez a sztochasztikus eset. A statisztikus alakfelismerés feladata olyan döntési tartományok meghatározása, amelyek alapján a döntés hibájának a valószínűsége a lehető legkisebb. A következő fejezetben erre a problémára ad egy lehetséges megoldást.

2.1.1. A Bayes döntés

Legyen adott egy (Ω, A, P) valószínűségi mező, $\underline{\xi}$ ezen a mezőn egy valószínűségi vektorváltozó, $C_i (i=1,\ldots,R)$ az (Ω, A, P)-n egy teljes eseményrendszer. Olyan $d(\underline{x})$ döntésfüggvényt keresünk, amellyel $\underline{\xi}$ osztályba sorolása, azaz valamely C_i-hez való hozzárendelése bizonyos értelemben optimális, pl. a legkisebb hibával jár. Hibás az osztályozás akkor, ha $d(\underline{\xi})=j$ és $\underline{\xi} \sim C_i$, $i \neq j$, $i,j=1,\ldots,R$.

Vezessük be a következő jelöléseket:

Legyen $p_i = P(C_i)$, $i=1\ldots R$ az i-edik osztály a priori valószínűsége. Nyilván $p_i \geq 0$ és $\sum_{i=1}^{R} p_i = 1$. Jelölje $f_i(\underline{x})$ a $\underline{\xi}$ megfigyelés C_i-re vonatkozó feltételes sűrűségfüggvényét, azaz

$f_i(\underline{x}) \geq 0$, $\int_{E_N} f_i(\underline{x})\,d\underline{x} = 1$ és $\int_A f_i(\underline{x})\,d\underline{x} = P(\underline{\xi} \in A | C_i)$, $A \subset E_N$.

Az $f_i(\underline{x})$ sűrűségfüggvények létezését általában semmi sem garantálja, de a legtöbb gyakorlati esetben ezt feltételezhetjük.

A fenti jelölésekkel a hibás osztályozás valószínűsége a következőképpen adható meg:

$$P_{ji} = \int_{D_j} f_i(\underline{x})\,d\underline{x}\,, \quad j \neq i\,, \quad i,j = 1,\dots,R\,, \tag{2.4}$$

ahol D_j a $d(\underline{x})$ döntésfüggvényhez tartozó döntési tartomány. A hibás döntésekhez hozzárendelhetünk egy számot, a hibás döntés költségét. Jelölje $c_{ij} = c_{ij}(d) \geq 0$, $i \neq j$, $c_{ii} = 0$ azt a költséget, (vagy veszteséget), amelyet a fenti esetben kapunk.

2.2. Definició. A $d = d(\underline{x})$ döntésfüggvényhez tartozó feltételes veszteség a C_i osztály esetén (tehát $\underline{\xi} \sim C_i$ esetén)

$$L_i = L_i(d) = \sum_{\substack{j=1 \\ j \neq i}}^{R} c_{ij} P_{ji} = \sum_{\substack{j=1 \\ j \neq i}}^{R} \int_{D_j} c_{ij} f_i(\underline{x})\,d\underline{x}\,, \quad i = 1,\dots,R\,. \tag{2.5}$$

Az $\underline{L} = (L_1,\dots,L_R)$ vektort veszteségvektornak nevezzük.

Két döntésfüggvény a hozzájuk tartozó veszteségvektor segítségével összehasonlítható; ehhez tekintsük a következő két definiciót:

2.3. Definició. Legyen $d_1(\underline{x})$ és $d_2(\underline{x})$ két döntésfüggvény, és $(L_{11},\dots,L_{R1})$ ill. $(L_{12},\dots,L_{R2})$ a hozzájuk tartozó veszteségvektorok. A d_1 döntésfüggvénnyel megadott döntési szabályt jobbnak nevezzük, mint a d_2-vel megadottat, ha

$$L_{i1} \leq L_{i2}\,, \quad i = 1,\dots,R\,,$$

és legalább egy i-re $L_{i1} < L_{i2}$. Ha minden i-re egyenlőség áll, a két döntési szabály ekvivalens.

2.4. Definició. Egy d döntésfüggvénnyel megadott döntési szabályt megengedettnek nevezünk, ha nem létezik nála - a 2.3. Definició értelmében - jobb döntés.

Célunk bizonyos kritériumok alapján egy megengedett döntési szabály megtalálása. Aszerint, hogy az a priori valószínűséget ismerjük-e, két módon járhatunk el:

a/ Ismert a priori valószínűségek esete

Egyszerű a megoldás abban az esetben, ha az a priori valószínűségeket ismerjük. Ekkor definiálhatjuk a feltétel nélküli várható veszteséget:

2.5. Definició. A d döntésfüggvényhez tartozó kockázat, vagy várható veszteség:

$$L = L(d) = \sum_{i=1}^{R} p_i L_i = \sum_{i=1}^{R} p_i \sum_{\substack{j=1 \\ j \neq i}}^{R} c_{ij} P_{ji} \qquad (2.6)$$

2.6. Definició. A (2.6)-ot minimalizáló - azaz a minimális veszteséget hozó - döntést a $\underline{p} = (p_1, \ldots, p_R)$ a priori valószínűségekhez tartozó Bayes döntésnek, a hozzátartozó döntésfüggvényt Bayes döntésfüggvénynek nevezzük. A Bayes függvényt a továbbiakban $d^*(\underline{x})$-gal, a hozzátartozó döntési tartományokat D_i^* -gal fogjuk jelölni.

2.1. Tétel. Adott $\underline{p}$ és c_{ij} mellett a (2.6) minimális, ha a döntési tartományok a következő alakúak:

$$D_j^* = \{\underline{x} \mid \underline{x} \in E_N, \sum_{\substack{i=1 \\ i \neq j}}^{R} p_i c_{ij} f_i(\underline{x}) \leq \sum_{\substack{i=1 \\ i \neq k}}^{R} p_i c_{ik} f_i(\underline{x})\}, \quad j = 1, \ldots, R, \; j \neq k. \qquad (2.7)$$

Ha a (2.7)-ben valamely indexekre az egyenlőség valószínűsége 0 minden k és j -re, akkor a minimalizálás egyértelmü egy 0 mértékü halmaz kivételével.

Bizonyitás

Először bizonyitjuk, hogy a (2.7)-ben megadott D_j tartományok minimalizálják (2.6)-ot. Helyettesítsük (2.6)-ban a P_{ji} helyébe (2.4)-et:

$$L=\sum_{i=1}^{R} p_i L_i=\sum_{i=1}^{R} p_i \sum_{\substack{j=1\\ j\neq i}}^{R} c_{ij} \int_{D_j} f_i(\underline{x})\, d\underline{x} \tag{2.8}$$

ahol most D_j , $j=1,\dots,R$ valamely $d(\underline{x})$ döntésfüggvényhez tartozó döntési tartomány. (2.8)-at alakítsuk át a következőképpen:

$$L=\sum_{i=1}^{R} p_i \sum_{\substack{j=1\\ j\neq i}}^{R} \int_{D_j} c_{ij} f_i(\underline{x})\, d\underline{x}=\sum_{i=1}^{R}\sum_{\substack{j=1\\ j\neq i}}^{R} \int_{D_j} p_i c_{ij} f_i(\underline{x})\, d\underline{x}=$$

$$=\sum_{j=1}^{R} \int_{D_j} \sum_{\substack{i=1\\ i\neq j}}^{R} p_i c_{ij} f_i(\underline{x})\, d\underline{x} \,. \tag{2.9}$$

Ebből az átalakításból látható, hogy adott p_i és c_{ij} $(i,j=1,\dots,R)$ mellett L akkor lesz minimális, ha $D_j=D_j^*$, $j=1,\dots,R$.

Hátra van még az egyértelműség bizonyítása.

Legyen

$$h_j(\underline{x})=\sum_{\substack{i=1\\ i\neq j}}^{R} p_i c_{ij} f_i(\underline{x}) \quad , \quad j=1,\dots,R \tag{2.10}$$

Ekkor

$$L=\sum_{j=1}^{R} \int_{D_j} h_j(\underline{x})\, d\underline{x}=\int_{E_N} h(\underline{x})\, d\underline{x} \,,$$

ahol $h(\underline{x})=h_j(\underline{x})$, ha $\underline{x}\in D_j$. Mivel a Bayes döntés

szerint $h^*(\underline{x}) = \min_i h_i(\underline{x})$, igy egy tetszőleges $d(\underline{x})$ -hez és a $d^*(\underline{x})$ -hez tartozó veszteség-különbség:

$$\int_{E_N} (h(\underline{x}) - h^*(\underline{x}))\, d\underline{x} = \sum_{i=1}^{R} \int_{D_j} (h_j(\underline{x}) - \min_i h_i(\underline{x}))\, d\underline{x} \geq 0 .$$

Egyenlőség csak akkor áll, ha $h_j(x) = \min_i h_i(x)$ a mindenek 0 a valószínűsége a feltétel miatt. □

2.3. Megjegyzés. Ha (2.7)-ben egyenlőség áll fenn valamely indexekre, akkor tetszés szerint osztályozhatunk.

A továbbiakban vizsgáljuk azt a speciális esetet, amikor $c_{ij} = 1$ minden i,j -re $(i \neq j)$, $i,j = 1, \dots, R$.
Ekkor (2.7) a következő alakú:

$$D_j^* = \{\underline{x} \mid \underline{x} \in E_N, \sum_{\substack{i=1 \\ i \neq j}}^{R} p_i f_i(\underline{x}) \leq \sum_{\substack{i=1, \\ i \neq k}}^{R} p_i f_i(\underline{x})\} \quad j \neq k, \; k,j = 1, \dots, R \qquad (2.11)$$

amely $\sum_{\substack{i=1 \\ i \neq j,k}}^{R} p_i f_i(\underline{x})$ kifejezéssel való egyszerűsítés után

$$D_j^* = \{\underline{x} \mid \underline{x} \in E_N, \; p_k f_k(x) \leq p_j f_j(\underline{x})\} \quad k,j = 1, \dots, R \, , \; k \neq j \qquad (2.12)$$

alakú. Ebben az esetben tehát a D_j^* tartományba azok a pontok tartoznak, amelyekre $p_j f_j(\underline{x})$ maximális.

Irjuk fel a Bayes döntéshez tartozó kockázatot ebben az esetben! (2.6) szerint

$$L^* = L(d^*) = \sum_{i=1}^{R} p_i \sum_{\substack{j=1 \\ j \neq i}}^{R} \int_{D_j^*} f_i(\underline{x})\, d\underline{x} = \sum_{i=1}^{R} \sum_{\substack{j=1 \\ j \neq i}}^{R} \int_{D_j^*} p_i f_i(\underline{x})\, d\underline{x} . \qquad (2.13)$$

Ez nem más, mint a Bayes döntés hibavalószínűsége.

Mivel a Bayes döntésről van szó, $p_i f_i(\underline{x})$ minimumát írhatjuk, így (2.13) tovább alakítható:

$$L^* = \sum_{i=1}^{R} \int_{D_j^*} \min_i p_i f_i(\underline{x})\, d\underline{x} = \int_{E_N} \min_i p_i f_i(\underline{x})\, d\underline{x} =$$

$$= 1 - \int_{E_N} \max_i p_i f_i(\underline{x})\, d\underline{x} . \qquad (2.14)$$

Vezessük be az $f(\underline{x}) = \sum_{i=1}^{R} p_i f_i(\underline{x})$ jelölést a $\underline{\xi}$ megfigyelés sűrűségfüggvényére, és jelölje

$$\hat{p_i}(\underline{x}) = \frac{p_i f_i(\underline{x})}{f(\underline{x})} = P(C_i \mid \underline{\xi} = \underline{x})$$

a C_i osztály a posteriori valószínűséget. Ekkor a Bayes döntés hibavalószínűsége a következőképpen is felírható:

$$P_e(d^*) = L(d^*) = \int_{E_N} \min_i \hat{p_i}(\underline{x}) f(\underline{x})\, d\underline{x} = 1 - \int_{E_N} \max_i \hat{p_i}(\underline{x}) f(\underline{x})\, d\underline{x} \qquad (2.15)$$

b/ <u>Ismeretlen a priori valószínűségek esete</u>

Abban az esetben, ha az a priori valószínűségek nem ismertek, egy megengedett döntési szabály megtalálásához más utat kell követni. Megengedett döntési szabályok halmazából való kiválasztás elve a <u>minimax elv:</u> válasszuk ki azt a d^M döntési szabályt, amelynél a várható veszteségek maximuma mimimális, azaz

$$L_i(d^M) \leq \max_{1 \leq i \leq R} L_i(d) \qquad i = 1, \ldots, R$$

vagy

$$\max_i (L_i(d^M)) = \min_d \max_i L_i(d) \qquad i = 1, \ldots, R .$$

A megengedett döntési szabályok halmazát, tulajdonságait és a Bayes döntéssel való kapcsolatát vizsgálta részletesen T. V. Anderson (1974), Rao (1973) és Wald (1950).

Fő eredményeik a következők:

(i) A megengedett döntési szabályok osztálya teljes (sőt minimálisan teljes).

2.4. Megjegyzés. A teljesség és a minimális teljesség definiciói:

2.7. Definició. Döntési szabályok egy C osztálya teljes, ha minden C-n kivüli d döntési szabályhoz tartozik egy olyan C-beli elem, amely jobb, mint d.

2.8. Definició. Döntési szabályok olyan teljes osztálya, amely nem tartalmaz más teljes osztályt, minimálisan teljes.

(ii) Minden megengedett döntési szabály egyúttal egy Bayes döntési szabály is.

(iii) Ha $p_i > 0$, $i = 1, \dots, R$, akkor a hozzá - mint a priori valószínűségeloszláshoz tartozó - Bayes döntés létezik és megengedett.

(iv) Létezik egy legkedvezőtlenebb $\underline{p}^0$ a priori valószínűségeloszlás és egy minimax döntési szabály, amely megengedett és a $\underline{p}^0$ -hoz tartozó Bayes döntési szabály.

(v) Ha $R = 2$, létezik egy egyértelmű minimax és megengedett döntési szabály, melyre $L_1(d^M) = L_2(d^M)$.

Egyes esetekben tehát az egyébként nehéz választás könnyen végrehajtható. Konkrétan két osztály esetén a következő fejezetben mutatjuk be a minimax megoldás megkeresését.

2.1.2. A Bayes döntés normális eloszlás esetén

Tegyük fel, hogy $\underline{\xi}$ normális eloszlású, és legyen

$$f_i(\underline{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\underline{\Sigma}_i|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \underline{\Sigma}_i^{-1}(\underline{x}-\underline{\mu}_i)\right] \quad , i = 1, \dots, R \tag{2.16}$$

ahol $\underline{\Sigma}_i$ jelöli a kovariancia mátrixot, $\underline{\mu}_i$ a $\underline{\xi}$ várható értékét a C_i osztály esetén, $|\Sigma_i|$ a kovariancia mátrix determinánsát, $\underline{\Sigma}_i^{-1}$ pedig az inverzét.

Az előző fejezethez hasonlóan most is két esetet vizsgálunk.

a/ Ismert a priori valószínűségek esete

Tegyük fel először, hogy az a priori valószínűségek ismertek. Legyen továbbá $c_{ij}=1$ minden i,j -re. Ekkor (2.7)-nek megfelelően a feladat olyan D_j tartományok keresése, $(j=1,\dots,R)$ melykre $p_i f_i(x)$ - vagy ami ezzel ekvivalens - $\log p_i f_i(\underline{x})$-maximális.

Legyen

$$g_i(\underline{x}) = \log p_i f_i(\underline{x}) \qquad i = 1, \dots, R.$$

(2.16)-t behelyettesítve

$$g_i(\underline{x}) = \log p_i - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|\underline{\Sigma}_i| -$$

$$-\frac{1}{2}\left(\underline{x}^T\underline{\Sigma}_i^{-1}\underline{x} - \underline{x}^T\underline{\Sigma}_i^{-1}\underline{\mu}_i - \underline{\mu}_i^T\underline{\Sigma}_i^{-1}\underline{x} + \underline{\mu}_i^T\underline{\Sigma}_i^{-1}\underline{\mu}_i\right) . \tag{2.17}$$

Mivel $\frac{N}{2}\log 2\pi$ i-től független, elhagyható; és

$$\underline{x}^T\underline{\Sigma}_i^{-1}\underline{\mu}_i = \underline{\mu}_i^T\underline{\Sigma}_i^{-1}\underline{x} \quad ,$$

(2.17) a következőképpen alakul:

$$g_i(\underline{x}) = -\frac{1}{2}\underline{x}^T\underline{\Sigma}_i^{-1}\underline{x} + \underline{\mu}_i^T\underline{\Sigma}_i^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_i^T\underline{\Sigma}_i^{-1}\underline{\mu}_i - \frac{1}{2}\log|\Sigma_i| + \log p_i \tag{2.18}$$

Ekkor

$$D_j^* = \{\underline{x} \mid \underline{x} \in E_N,\ g_k(\underline{x}) \leq g_j(\underline{x})\} \quad k \neq j,\ k,j = 1,\dots,R,$$

vagy

$$D_j^* = \{\underline{x} \mid \underline{x} \in E_N,\ d_{jk}(\underline{x}) = g_j(\underline{x}) - g_k(\underline{x}) \geq 0\} \tag{2.19}$$

itt $d_{jk}(\underline{x})$ a j és k-adik osztályt elválasztó felület egyenlete, amely esetünkben egy másodrendű felület.

Ha még feltesszük, hogy $\underline{\Sigma}_i = \underline{\Sigma}$, $i = 1,\dots,R$, azaz a kovariancia mátrixok megegyeznek, akkor (2.19)-ből a másodfokú tag és az $\frac{1}{2}\log\underline{\Sigma}$ kifejezés is elhagyható:

$$g_i(\underline{x}) = \underline{\mu}_i^T\underline{\Sigma}^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_i^T\underline{\Sigma}^{-1}\underline{\mu}_i + \log p_i = \underline{w}_i^T\underline{x} + w_0$$

ahol

$$\underline{w}_i = \underline{\Sigma}^{-1}\underline{\mu}_i$$

$$w_0 = -\frac{1}{2}\underline{\mu}_i^T\underline{\Sigma}^{-1}\underline{\mu}_i + \log p_i$$

Bármely két osztályt elválasztó felület ekkor egy hipersik:

$$d_{jk}(\underline{x}) = (\underline{\mu}_j - \underline{\mu}_k)^T\underline{\Sigma}^{-1}\underline{x} - \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)^T\underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k) + \log\frac{p_j}{p_k} = \tag{2.20}$$

$$= \underline{w}^T\underline{x} + w_0$$

ahol most $\underline{w} = \underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k)$

$$w_0 = -\frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)^T\underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k) + \log\frac{p_j}{p_k}$$

Ha még feltesszük, hogy $p_1 = \dots = p_R$, akkor

$$D_j^* = \{\underline{x} \mid \underline{x} \in E_N,\ (\underline{\mu}_j - \underline{\mu}_k)^T\underline{\Sigma}^{-1}\underline{x} \geq \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)^T\underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k)\}$$

$$j,k = 1,\dots,R,\ j \neq k \tag{2.21}$$

a C_j osztályba soroláshoz a legjobb osztályozást adja.

b/ <u>Ismeretlen a priori valószínűségek esete</u>

Tegyük fel, hogy az a priori valószínűségek nem ismertek. Ekkor a minimax elv szerint kell a megfelelő döntést kiválasztanunk. Tekintsük két osztály esetét, tehát legyen $R=2$, és legyen $\underline{\xi}$ megfigyelés adott.

Tekintsük az

$$u = (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} \underline{\xi} - \frac{1}{2} (\underline{\mu}_1 + \underline{\mu}_2)^T \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \qquad (2.22)$$

valószínűségi változót. Ha $\underline{\xi}$ $N(\underline{\mu}_1, \underline{\Sigma})$ eloszlású, akkor u is normális eloszlású, és

$$E(u) = (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} \underline{\mu}_1 - \frac{1}{2} (\underline{\mu}_1 + \underline{\mu}_2)^T \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) =$$

$$= \frac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) = \frac{1}{2} \alpha \qquad (2.23)$$

ahol $$\alpha = (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \qquad (2.24)$$

Az u varianciája

$$Var(u) = E(u - E(u))^2 = E\left[(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} \underline{\xi} - (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} \underline{\mu}_1\right]^2 =$$

$$= E\left[(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} (\underline{\xi} - \underline{\mu}_1)(\underline{\xi} - \underline{\mu}_1)^T \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2)\right] =$$

$$= (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} E\left[(\underline{\xi} - \underline{\mu}_1)(\underline{\xi} - \underline{\mu}_1)^T \underline{\Sigma}^{-1}\right](\underline{\mu}_1 - \underline{\mu}_2) = (\underline{\mu}_1 - \underline{\mu}_2)^T \underline{\Sigma}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) = \alpha \qquad (2.25)$$

Tehát ha $\underline{\xi}$ $N(\underline{\mu}_1, \underline{\Sigma})$ eloszlású, akkor u $N(\frac{1}{2}\alpha, \alpha)$ eloszlású. Hasonlóan látható be, hogy ha $\underline{\xi}$ a másik osztályhoz tartozó objektum jellemző vektora, tehát $N(\underline{\mu}_2, \underline{\Sigma})$ eloszlású, akkor u eloszlása $N(-\frac{1}{2}\alpha, \alpha)$. A $\underline{\xi}$ -hez tartozó objektumot így a $\underline{\xi}$ vektorváltozó helyett az u egydimenziós valószínűségi változóval vizsgálhatjuk.

Ha $\underline{\xi} \sim C_1$, akkor a hibás osztályozás valószínűsége

$$P_{21} = \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2\alpha}\left(z - \frac{\alpha}{2}\right)^2} dz = \int_{-\infty}^{\frac{c - \alpha/2}{\sqrt{\alpha}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \qquad (2.26)$$

ahol c alkalmasan választott konstans, és az $y = \frac{z - \frac{\alpha}{2}}{\sqrt{\alpha}}$ helyettesítést alkalmaztuk.

Ha $\underline{\xi} \sim C_2$, akkor

$$P_{12} = \int_{c}^{\infty} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2\alpha}\left(z + \frac{\alpha}{2}\right)^2} dz = \int_{\frac{c + \alpha/2}{\sqrt{\alpha}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy . \qquad (2.27)$$

A minimax megoldáshoz c -t úgy kell választanunk, hogy

$$c_{21} \int_{(c+\frac{\alpha}{2})/\sqrt{\alpha}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy = c_{12} \int_{-\infty}^{(c-\frac{\alpha}{2})/\sqrt{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \qquad (2.28)$$

teljesüljön.

Ha feltesszük, hogy a hibás osztályozás költségei egyenlők, azaz $c_{21}=c_{12}$ és $c=0$, akkor a hibavalószínűség

$$\int_{\sqrt{\alpha}/2}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy , \qquad (2.29)$$

amely a normális eloszlás táblázatából könnyen meghatározható.

2.2 A statisztikus alakfelismerés modelljei, módszerei

2.2.1. A statisztikus alakfelismerés modellje

Az előző fejezetben a Bayes döntéssel megadtuk az alakfelismerési feladat egy lehetséges megoldását. Eszerint a feltételes sűrűségfüggvények - és esetleg az a priori valószínűségek - ismeretében egy $\underline{x}$ megfigyelést a legkisebb hibával tudjuk a megfelelő osztályhoz hozzárendelni.

A feltételes sűrűségfüggvények pontosan a legritkábban ismertek, szerencsés esetben az eloszlás típusáról van feltételezésünk. Ezért a döntésfüggvény konstruálásához szükségünk van egy mintára, az ún. tananyagra.

2.6. Definició. A $(\underline{\xi}_1,\vartheta_1),\dots,(\underline{\xi}_n,\vartheta_n),\dots$ azonos eloszlású, teljesen független valószínűségi változósorozatot tananyagnak nevezzük, ahol $\underline{\xi}_n \in E_N$ és $\vartheta_n = i$ $(i=1,\dots,R)$, ha $\underline{\xi}_n$ az i-edik osztályból származik. A $\underline{\xi}_n$-eket tanulópontoknak, a ϑ_n -eket tanításnak nevezzük.

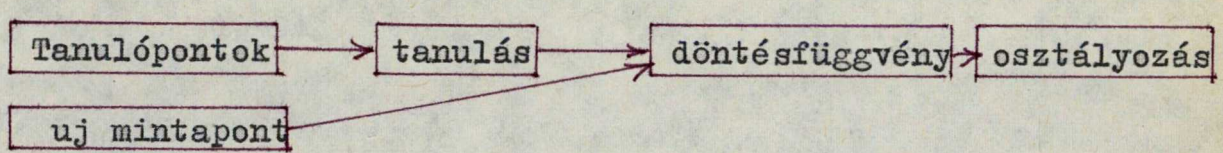
A ϑ_n -ről feltételezzük, hogy úgy veszi fel az $1,2,\dots,R$ értékeket, hogy

$$E(\vartheta_n|\underline{\xi}_n) = d(\underline{\xi}_n) \tag{2.30}$$

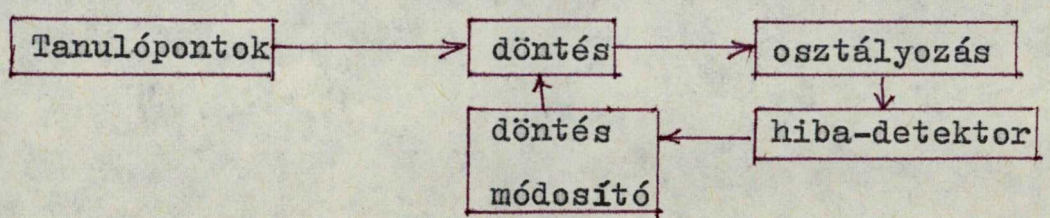
ahol $d(\underline{\xi}_n)$ a $d(x)$ döntésfüggvény $\underline{\xi}_n$ pontbeli értéke. A tananyag tehát ismert kategóriájú vektorokból álló statisztikai minta, a döntésfüggvényt konkrétan ennek segítségével lehet meghatározni.

A statisztikus alakfelismerő módszerek célja olyan döntésfüggvény konstruálása a tananyag alapján, hogy a döntéshez tartozó hibavalószínűség a legkisebb legyen, azaz a döntés a lehető legjobban közelítse meg a Bayes döntést.

A döntésfüggvény meghatározása és alkalmazása többféleképpen történhet. Az egyik lehetőség a "felismerés előtti tanulás" (2.2. ábra). Eszerint adott tananyag alapján megalkotjuk a döntésfüggvényt (tanulás), és egy új pontot a döntésfüggvény segítségével osztályozunk. A döntés "jóságának" ellenőrzésére szükség van egy újabb ismert kategóriáju ponthalmazra, a teszthalmazra, amelyre a döntési szabályt alkalmazva és összehasonlítva az ismert kategóriával, a tévedések számából következtetni lehet a hibás döntés valószínűségére. A másik lehetőség az "egyidejű tanulás és felismerés" (2.3. ábra). A döntésfüggvény itt a hiba-detektor szerint módosul úgy, hogy minden egyes tanulópont beérkezése után a döntés javul vagy változatlan marad.



2.2. ábra



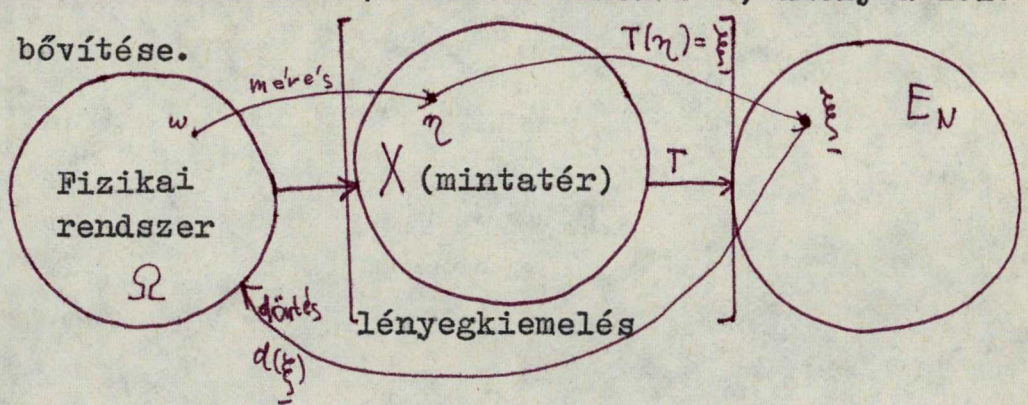
2.3. ábra

A döntésfüggvény meghatározásakor természetesen minél nagyobb méretű tananyagra törekszünk a megbízható döntés elérése érdekében.

A tananyag méretére vonatkozóan pontos ismereteink nincsenek. Fritz szerint az "N dimenziós térben történő tanuláshoz hozzávetőlegesen $A \cdot 2^N$ nagyságú tananyagra van szükség". (Fritz /1974, 38.o./), ahol A konstans a tanuló algoritmustól és az eloszlásoktól függ. Ilyen nagy elemszámra tett követelmény a gyakorlatban ritkán teljesíthető. Ezért érdekes Hughes (1968) elméleti tanulmánya, mely szerint véges a mintaelemszám esetén létezik egy optimális N dimenziószám, amely felett az osztályozás sikeressége átlagosan csökkenhet, sőt Lachenbruch (1969) szerint bizonyos módszereknél elegendő osztályonként a koordinátaszám háromszorosa. Tapasztalataink szerint különösen az orvosi diagnosztikában kényes kérdés a mintaelemszám. Kellő méretű tananyag előállítása gondot okozhat egyrészt bizonyos betegségek ritkasága miatt, másrészt a kontroll esetek számának növelése is nehéz, ha a szükséges vizsgálatok kockázatosak vagy akár csak kellemetlenek. A gyakorlatban általában arra törekszünk, hogy a mintaelemszám sokkal nagyobb legyen, mint a dimenziószám.

Egy másik, a döntést jelentősen befolyásoló előkészítő munka a mintaelemszám meghatározása mellett a figyelembe veendő jellemzők (koordináták) megválasztása, az ún. lényegkiemelés.

Az osztályozandó objektumok sokszor olyanok, hogy a róluk származó megfigyelések esetleg nem is euklideszi tér pontjai. Lehetnek pl. függvények - EEG vagy EKG görbék: azaz az agyról illetve a szívről elvezetett elektromos jelek grafikonjai, vagy betűk. Tehát a vizsgált valószínűségi változó eredetileg egy $(X, \mathcal{X}, \mu)$ mérhető térből veszi az értékeit. Az osztályozási feladat annál könyebben oldható meg, minél kevésbé "absztrakt" a X tér. Ezért a tulajdonképpeni és az eddig tárgyalt alakfelismerési feladatot ilyen esetben megelőzheti egy olyan transzformáció, amely a megfigyelést egy egyszerű struktúrájú, leggyakrabban euklideszi térbe transzformálja. Ez a <u>transzformáció a lényegkiemelés.</u> Az alakfelismerő rendszert tágabb értelemben a 2.4. ábra szemlélteti, amely a 2.1. ábra bővítése.



2.4. ábra

Ha az X eredetileg maga is véges dimenziós euklideszi tér, a lényegkiemelés célja lehet pl. a dimenziószám csökkentése. A lényegkiemelésnél többnyire a következő szempontokat veszik figyelembe:

a/ az osztályozás alacsony dimenziójú térben történjen lehetőleg.

b/ a jellemzők elegendő információt hordozzanak az alakzatra vonatkozólag, a dimenzió redukálás ne járjon tehát számottevő információveszteséggel. (Ez utóbbin itt az eredeti és a transzformált megfigyelésekhez tartozó Bayes döntések hibavalószínűségeinek eltérését értjük.)

c/ a lényegkiemelést rendszerint olyan kritérium alapján végzik, hogy az egyes osztályokba tartozó pontok a transzformáció után "közelebb" kerüljenek egymáshoz, az osztályok közötti "távolság" pedig növekedjék.

A lényegkiemeléssel a továbbiakban nem kívánunk részletesen foglalkozni. Csupán néhány leggyakrabban használatos módszert említek, ezek:

- korrelációszámítás
- faktoranalízis (Harman /1976/)
- információelméleti módszerek (pl. Watanabe /1967/, Tou és Heydorn /1967/)

A konkrét lényegkiemelési módszer megválasztása az adott feladattól függ. A 4. fejezetben, amelyben saját tapasztalatainkról számolunk be, kitérünk az esetenként alkalmazott (igen egyszerű) lényegkiemelési módszerekre.

2.2.2. Statisztikus alakfelismerési módszerek

A 2.1.1. pontban a Bayes döntés az alakfelismerési feladat megoldását adja: az a priori valószínűségek és a feltételes sűrűségfüggvények ismeretében a Bayes döntés végrehajtása igen egyszerű. Általában azonban ezek nem ismertek, Rendelkezésre áll viszont egy minta, a tananyag, amelynek segítségével becsülhetjük e valószínűségeket, a sűrűségfüggvényeket, vagy közelíthetjük magát a döntésfüggvényt. Az egyes alakfelismerési eljárások abban különböznek egymástól, hogy a döntésfüggvényt hogyan állítják elő. Ebben a részben bemutatunk néhány statisztikus alakfelismerési módszert, és kapcsolatukat a Bayes döntéssel.

a/ A legközelebbi szomszéd módszer

Az a priori valószínűségek és a sűrűségfüggvények ismeretének hiányában heurisztikus eljárások is konstruálhatók egy pontnak egy adott tananyag alapján való osztályozására. Természetesen az a szerencsés, ha a heurisztikus algoritmusokról megállapítható a Bayes döntéshez való kapcsolatuk.

A Cover és Hart (1967) által leírt "legközelebbi szomszéd" (Nearest Neighbour - NN) algoritmus elve igen egyszerű és szemléletes:

Legyen adott $\zeta_n = \{(\underline{\xi}_1, \vartheta_1), \dots (\underline{\xi}_n, \vartheta_n)\}$ tananyag, $\vartheta_i \in \{1, \dots, R\}$, $i = 1, \dots, n$, azaz ϑ_j a $\underline{\xi}_j$ tanulópont osztályának a sorszámát jelöli. Ez azt is jelenti, hogy $P(\vartheta_j = i \mid \underline{\xi}_j = \underline{x}) = \hat{p}_i(\underline{x})$, $j = 1, \dots, n$,

ahol $\hat{p}_i(\underline{x})$ az i-edik osztály a posteriori valószínűsége. Legyen $\underline{\xi}$ az osztályozandó vektor, és jelölje $\underline{\xi}$ valódi osztályának sorszámát ϑ.

2.6 Definíció. Azt a $\underline{\xi}_n^* \in \{\underline{\xi}_1, \dots, \underline{\xi}_n\}$ pontot, amelyre

$$\min_{1 \le i \le n} \operatorname{dist}(\underline{\xi}_i, \underline{\xi}) = \operatorname{dist}(\underline{\xi}_n^*, \underline{\xi}), \tag{2.31}$$

ahol dist $(.,.)$ az E_N-en értelmezett tetszőleges távolságfüggvény, a $\underline{\xi}$ legközelebbi szomszédjának nevezzük.

Az NN döntés eszerint: $\underline{\xi}$ -t abba a kategóriába sorolja, amelynek a sorszáma éppen ϑ_n^*.

Hibás a döntés akkor, ha $\vartheta \neq \vartheta_n^*$.

Adott véges elemszámú tananyag esetén a hibás döntés valószínűsége maga is valószínűségi változó, amely a tananyagtól függ:

$$P(H_{NN}^n) = \sum_{i=1}^{R} P(\vartheta = i, \vartheta_n^* \neq i \mid \zeta_n). \tag{2.32}$$

(2.32) feltételes várható értéke rögzített n mellett:

$$E(P(H_{NN}^n \mid \zeta_n)) = E\Big(\sum_{i=1}^{R} P(\vartheta = i, \vartheta_n^* \neq i \mid \zeta_n, \underline{\xi}) \mid \zeta_n\Big). \tag{2.33}$$

A feltételes valószínűség definíciója szerint

$$P(\vartheta = i, \vartheta_n^* \neq i \mid \zeta_n, \underline{\xi}) = P(\vartheta_n = i \mid \vartheta_n^* \neq i, \zeta_n, \underline{\xi}) \cdot P(\vartheta_n^* \neq i \mid \zeta_n, \underline{\xi}) =$$
$$= P(\vartheta_n = i \mid \underline{\xi}) \cdot P(\vartheta_n^* \neq i \mid \underline{\xi}_n^*) = \hat{p}_i(\underline{\xi})(1 - \hat{p}_i(\underline{\xi}_n^*)). \tag{2.34}$$

(2.33)-at a (2.34) felhasználásával a következőképpen alakíthatjuk:

$$P_{NN}^n = E(P(H_{NN}^n \mid \zeta_n)) = E\Big[\sum_{i=1}^{R} (\hat{p}_i(\underline{\xi}) - \hat{p}_i(\underline{\xi})\hat{p}_i(\underline{\xi}_n^*)) \mid \zeta_n\Big] \tag{2.35}$$

Cover és Hart (1967) bizonyította a következő tételt:

Tétel: Ha az a posteriori valószínűségek folytonosak,

akkor (i) a (2.35) feltételes várható értéknek

létezik határértéke, és ez

$$P_{NN} = \lim_{n\to\infty} P_{NN}^{n} = E\left(1 - \sum_{i=1}^{R} p_i^2(\underline{\xi})\right) \qquad (2.36)$$

(ii)

$$P_e(d^*) \le P_{NN} \le P_e(d^*)\left(2 - \frac{R}{R-1} P_e(d^*)\right), \qquad (2.37)$$

ahol $P_e(d^*)$ a (2.15)-ben megadott Bayes hiba.

A P_{NN} : az NN módszer aszimplotikus átlagos hibája. Bebizonyítható, hogy a P_{NN} határérték valóban jól becsüli a hibavalószínűség tényleges értékét. /Wagner (1971)/.

A tételben adott egyenlőtlenség áll, tehát a NN módszer akkor is alkalmazható, ha a tanulópontok tere nem végesdimenziós, hanem tetszőleges szeparábilis metrikus tér. Ezért az NN módszert többféle metrikával is végrehajthatjuk. A nem euklideszi metrikák használatának a konkrét gyakorlati alkalmazások során van jelentősége.

(2.37) azt is kifejezi, hogy $P_{NN} \le 2P_e(d^*)$, tehát az NN módszer aszimplotikus átlagos hibája a Bayes hiba kétszeresénél nem lehet nagyobb. Az NN módszer tehát kis Bayes hiba esetén jó eredményt ad.

Ezen előnyei indokolják, hogy az NN módszer mind elméleti vizsgálatokban (Patrick, Fisher, (1979), Wagner (1971)) mind gyakorlati alkalmazásokban (lásd.a 4. fejezetet) széleskörüen elterjedt annak ellenére, hogy számítógépes realizáláskor az egész tananyagot tárolni kell.

Az NN módszer változatai közül a k-NN módszer a legismertebb. Ez a $\underline{\xi}$ -hez legközelebbi $k = 2k_0 + 1$ pontot vizsgálja $(k_0 = 1, 2, \ldots)$ és $\underline{\xi}$ -t abba a kategóriába sorolja, ahova a többség tartozik. A k-NN módszer nem mindig jobb a NN-nél, lehet olyan példát találni, (Fritz (1974)), melyre $P_{NN} < P_{k\text{-}NN}$.

b/ Súlypont módszer

2.1.2-ben megmutattuk, hogy a Bayes döntésfüggvény normális eloszlás és egyenlő kovariancia mátrixok esetén lineáris. Más lineáris döntésfüggvények is konstruálhatók a sűrűségfüggvények felhasználása nélkül. Ilyen az ún. súlypont módszer, amelynek elve a következő.

Legyen adott egy tananyag, melyről tegyük fel, hogy n számú pontot tartalmaz, és n_j számú tanulópontot az egyes kategóriákból.

Jelölje $\underline{x}_i^{(j)}$, $i=1,\dots,n_j$, $j=1,\dots,R$

a j-edik kategóriából származó i-edik tanulópontot. A továbbiakban ezt a jelölést fogjuk alkalmazni a tananyagra. Legyen $\underline{m}_j$ a j-edik osztály (B_j) súlypontja (átlagvektora). Legyen a döntés az, hogy egy $\underline{x}$ pontot ahhoz az osztályhoz sorolunk, amelynek a súlypontjához a legközelebb esik, azaz amelyre a

$$d_k=(\underline{x}-\underline{m}^{(k)})^T(\underline{x}-\underline{m}^{(k)})=\underline{x}^T\underline{x}-2\,\underline{m}^{(k)T}\underline{x}+\underline{m}^{(k)T}\underline{m}^{(k)}\to\min \qquad (2.38)$$

Mivel a minimalizálásnál az első tag nem játszik szerepet, elhagyható, így lineáris függvényt kapunk. A j-edik és k-adik osztályt elválasztó hipersik egyenlete:

$$d_{jk}(\underline{x})=(\underline{m}^{(j)}-\underline{m}^{(k)})^T\underline{x}-\frac{1}{2}(\underline{m}^{(j)T}\underline{m}^{(j)}-\underline{m}^{(k)T}\underline{m}^{(k)})=$$

$$=(\underline{m}^{(j)}-\underline{m}^{(k)})^T\underline{x}-\frac{1}{2}(\underline{m}^{(j)}+\underline{m}^{(k)})^T(\underline{m}^{(j)}-\underline{m}^{(k)}) \qquad (2.39)$$

Ez az elválasztási módszer akkor ad jó eredményt, ha a szórások megegyeznek. Pontosabban (2.39) a Bayes döntés speciális esete normális eloszlás, egyenlő a priori valószínűségek és $\underline{\Sigma}=\sigma^2\underline{I}$ esetén. Ugyanis

$\underline{\Sigma}^{-1}=\frac{1}{\sigma^2}\underline{I}$ miatt (2.17) most

$$g_i(\underline{x})=-\frac{1}{2\sigma^2}(\underline{x}-\underline{m}^{(i)})^T(\underline{x}-\underline{m}^{(i)})$$

alakú. Ezt a módszert Duda és Hart (1973) minimum-távolság osztályozásnak nevezi.

c/ Két osztály lineáris döntésfüggvény segítségével való szétválasztásához (szeparálásához) az együtthatók meghatározására alkalmas iteratív algoritmus a következő; (Rosenblatt, 1961):

Adott a tananyag, $\{\underline{x}_i^{(j)}\}$, $i=1,\dots,n_j$, $j=1,2$.

Feladat a $d(\underline{x})=\underline{w}^T\underline{x}+w_0$ alakú döntésfüggvény együtthatóinak meghatározása. $d(x)>0$ esetén C_1 egyébként C_2 a döntés. Az iteratív algoritmus a következő:

1. Legyen $d^{(0)}(\underline{x})=\underline{w}^{(0)T}\underline{x}+w_0^{(0)}$ tetszőleges kezdeti döntésfüggvény (pl. $\underline{w}$ azonosan 0 vektor)

2. Tegyük fel, hogy eljutottunk n lépésben a döntésfüggvény n-edik közelítéséhez:

$$d^n(\underline{x})=\underline{w}^{(n)T}\underline{x}+w_0^{(n)}$$

Ekkor a $d^{(n+1)}(\underline{x})$ együtthatóit a következő "tanulási szabállyal" kapjuk:

$$\underline{w}^{(n+1)}=\underline{w}^{(n)}+c_n\underline{x}_n\,,$$
$$w_0^{(n+1)}=w_0^{(n)}+c_n\,,$$

ahol $\underline{x}_n$ a soron következő tanulópont, osztálytól függetlenül, és

$$c_n=\begin{cases}0, & \text{ha } d^{(n)}(\underline{x}) \text{ az } \underline{x}_n \text{ pontot jól osztályozta,}\\ c, & \text{ha } d^{(n)}(\underline{x}_n)<0 \text{ és } \underline{x}_n\in B_1\,,\\ -c, & \text{ha } d^{(n)}(\underline{x}_n)>0 \text{ és } \underline{x}_n\in B_2\,,\end{cases}$$

és c tetszőleges konstans.

<u>Állítás:</u> ha az $\{\underline{x}_i^{(1)}\}_{i=1}^{n_1}$ és $\{\underline{x}_i^{(2)}\}_{i=1}^{n_2}$ halmazok

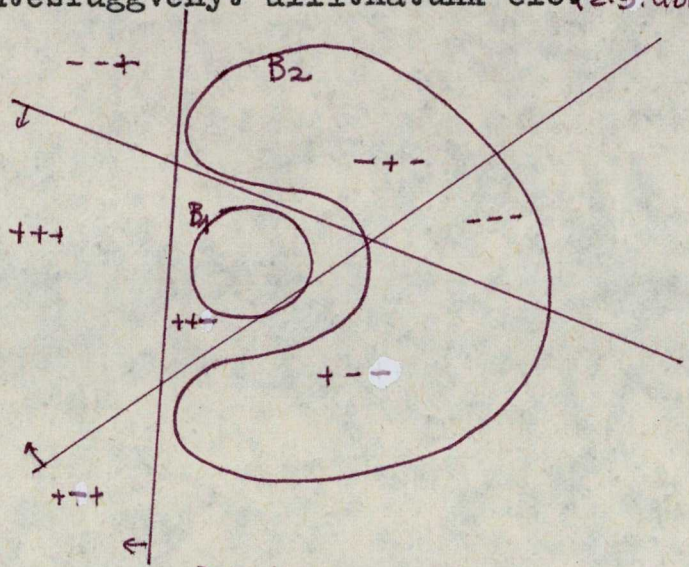
lineárisan szétválaszthatók, akkor a fenti algoritmus véges számú lépésben véget ér, azaz van olyan k_{max} szám, hogy

$$d^{(k+1)}(\underline{x})=d^{(k)}(\underline{x}) \qquad \text{minden } k>k_{max}\text{-ra.}$$

Attila József Tudományegyetem Központi Könyvtára SZEGED

A bizonyítás, pld. Meyer-Brötz, Schürman (1972) könyvében megtalálható.

d/ Az előbbi algoritmus általánosítása az ún. Commitee Machine algoritmus (Nilsson, 1963), amellyel szakaszonként lineáris döntésfüggvényt állíthatunk elő (2.5. ábra)



2.5. ábra.

Adott két osztály, a feladat a két halmaz $2L+1$ számú hipersíkkal történő szétválasztása, ahol $L = 0, 1, 2, \ldots$ előre rögzített szám:

$$d_1(\underline{x}) = \underline{w}_1^T \underline{x} + w_{10}$$
$$d_2(\underline{x}) = \underline{w}_2^T \underline{x} + w_{20}$$
$$\vdots$$
$$d_{2L+1}(\underline{x}) = \underline{w}_{2L+1}^T \underline{x} + w_{(2L+1)0} .$$

A döntés "többségi elv" szerint történik: egy ismeretlen $\underline{x}$ pontot abba az osztályba sorolunk, amelybe $d_1(\underline{x}), \ldots, d_{2L+1}(\underline{x})$ közül a többség döntötte:

$$d(\underline{x}) = \operatorname{sign}\left[\sum_{j=1}^{2L+1} \operatorname{sign} d_j(\underline{x})\right]$$

A döntésfüggvények együtthatóinak meghatározása iteratív módon, a tanulópontok beérkezése utáni - az előző algoritmushoz hasonló - módosításokkal történik. Az algoritmus gyakorlati tapasztalatok szerint jól alkalmazható annak

ellenére, hogy a konvergenciára nincs ismert bizonyítás.

e/ Cluster algoritmusok

Az eddig ismertetett módszereket szokás a "tanulás tanítóval" (supervised learning) közös néven emlegetni, mivel az osztályba sorolást, illetve a döntésfüggvény megalkotását egy tananyag alapján lehet megadni. Az alakfelismerési módszerek egy másik csoportjánál, a "tanító nélküli algoritmusok"-nál (unsupervised learning) a kiinduló-pont is más: adott adathalmaz elemeiből kell valamilyen módon "osztályokat", ún. clusterokat képeznünk. Az adathalmazról semmi előzetes ismeretünk nincs.

A 2.2 pontbeli tananyag definícióval összhangban tekintsük az adatokat most is független, azonos eloszlású valószínűségi változósorozatnak:

$$\underline{\xi}_1, \underline{\xi}_2, \dots, \underline{\xi}_n, \dots \in E_N$$

(Ez a sorozat a valóságban mindig véges). A $\{\underline{\xi}_i\}_{i=1}^{\infty}$ változók közös sűrűségfüggvényét jelölje $f(\underline{x})$. Tekinthetjük clustereknek pl. az E_N tér olyan részhalmazait, amelyekben a $\underline{\xi}_n$ pontok koncentrációja viszonylag nagy, pl. ha megadhatók olyan diszjunkt $D_1, \dots, D_s$ halmazok az E_N-ben, hogy rajtuk az $f(\underline{x})$ viszonylag nagy, rajtuk kívül pedig mindenütt nagyon kicsi. A cluster algoritmusok nagy része az elemek között definiált "távolság" vagy "hasonlóság" fogalomra épül. Így egy clusterbe kerülnek azok a pontok, amelyek a "leghasonlóbbak", vagy a "legközelebb" vannak egymáshoz, mint más clusterbeli pontokhoz.

Nagyon sok cluster algoritmust írtak már le, a legismertebbek Hartigan (1975), Anderberg (1973), illetve Duran-Odell (1974) monográfiákban megtalálhatók.

A cluster algoritmusok két nagy csoportja a hierarchikus és dinamikus módszerek. A hierarchikus módszerek elve: kezdetben minden pont önmagában egy cluster. E pontok "hasonlósága" alapján összevonjuk a "leghasonlóbb" clustereket. Az összevonást addig folytatjuk, míg végül egy clustert kapunk, amelybe az összes pont beletartozik. Az eredmény függ a hasonlósági mérték megválasztásától és az összevonás módjától.

A dinamikus módszerek lényege az, hogy kezdeti konfigurációból indulnak ki - azaz előzetes feltevésünk van pl. a clusterek számáról, középpontjairól stb. Az - általában iterativ - algoritmusok a $\{\xi_n\}$ halmaz pontjai alapján módosítják a kezdeti konfigurációt.

A cluster módszerekkel nemcsak az adatok, hanem a változók is osztályozhatók, így a cluster analízis hasznos lényegkiemelési módszer is lehet. Ilyen értelemben rokon a faktoranalízissel.

Nem célunk a cluster algoritmusok részletes leírása. E rövid kis ismertetéssel csak az algoritmusok sokaságát, sokrétűségét kívántuk érzékeltetni és azt a bizonytalanságot, amely a helyes módszer megválasztását és ennélfogva az eredmények interpretálását jellemzi.

3. Sűrűségfüggvénybecslés és a lineáris diszkriminancia analízis

Ismeretlen a priori valószínűségek és sűrűségfüggvények esetén a Bayes döntésfüggvény közelítésének kézenfekvő módja adott tananyag alapján a feltételes sűrűségfüggvények becslése. Erre igen sok eljárás ismeretes, mint pld. az egyszerű hisztogram-becslések vagy pedig a Parzen (1962) illetve Meisel (1972) által javasolt módszerek, illetve ezek általánosításai. Számos eljárás, továbbá a Bayes döntésfüggvényt közelítő iteratív tanuló algoritmus született ezek alapján, a legismertebb a Wagner-Wolverton (1967) illetve Specht (1967) féle polinomális diszkrimináló módszer. Ezek ismertetésére nem térünk ki, mivel alkalmazásukhoz igen nagy tananyag szükséges, amely a mi orvosi diagnosztikai célú vizsgálatainkban eddig nem állt rendelkezésre.

Részletesen foglalkozunk viszont azzal az esettel, amelyben feltesszük, hogy az eloszlás típusa ismert, pl. normális eloszlás. Ekkor a sűrűségfüggvény-becslés az eloszlás paramétereinek becslésére korlátozódik. Ez az eset azért is érdekes, mert ha még a kovariancia mátrixok egyenlőségét is feltesszük, akkor a Bayes döntésfüggvény - mint azt a 2.1.2-ben megmutattuk - lineáris és a sűrűségfüggvény paramétereinek becsléseit behelyettesítve a kapott függvény megegyezik a Fisher féle lineáris diszkriminancia analízis döntésfüggvényével.

Ennek bizonyítása kis eltérésekkel több helyen megtalálható. /Anderson (1974), Kendall (1977)/. Azonban bizonyos, a Fisher féle diszkriminancia analízist részletesen tárgyaló könyvek erről nem tesznek említést. /Weber E. (1969), Nollau (1975)/. A Fisher féle diszkriminancia analízis kétféle módon való megközelítése igen tanulságos volt a hasnyálmirigy gyulladás diagnosztizálásával kapcsolatos vizsgálatainkban. (4.3 fejezet)

3.1 A sűrűségfüggvénybecslés normális eloszlás esetén

Legyen adott egy tananyag. Jelölésére használjuk a 2.2.2 b/-ben bevezetett jelölést, azaz legyen

$\underline{x}_i^{(j)} \quad i=1,\dots,n_j, \quad j=1,\dots,R$

a j-edik osztályból származó i-edik tanulópont. Tegyük fel, hogy $\{\underline{x}_i^{(j)}\}_{i=1}^{n_j}$ egy $N(\underline{\mu}_j, \underline{\Sigma})$ eloszlású populációból vett statisztikai minta $(j=1,\dots,R)$.

A $\underline{\mu}_j$ és $\underline{\Sigma}$ becslésére az

$$\underline{m}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j} \underline{x}_i^{(j)} \tag{3.1}$$

és

$$S = \frac{1}{\left(\sum_{j=1}^{R} n_j - R\right)} \sum_{j=1}^{R}\sum_{i=1}^{n_j} \left(\underline{x}_i^{(j)} - \underline{m}^{(j)}\right)\left(\underline{x}_i^{(j)} - \underline{m}^{(j)}\right)^T \tag{3.2}$$

formulákat használjuk. A 2.1.2. fejezetben láttuk, hogy a

$D_j^* = \{\underline{x} \mid \underline{x} \in E_N,\ d_{jk}(\underline{x}) > 0\} \quad j,k=1,\dots,R,\ j \neq k$

a legkisebb hibával járó döntési tartomány, ahol (2.19) illetve (2.20) szerint

$$d_{jk}(\underline{x}) = (\underline{\mu}_j - \underline{\mu}_k)^T \underline{\Sigma}^{-1} \underline{x} - \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)^T \underline{\Sigma}^{-1}(\underline{\mu}_j - \underline{\mu}_k) + \log \frac{p_j}{p_k} \tag{3.3}$$

Ha a (3.1) és (3.2) becsléseket (33)-be helyettesítjük, egyenlő a priori valószínűségek esetén a

$$\tilde{d}_{jk}(\underline{x}) = (\underline{m}^{(1)} - \underline{m}^{(2)})^T \underline{S}^{-1} \underline{x} - \frac{1}{2}(\underline{m}^{(j)} + \underline{m}^{(k)})^T \underline{S}^{-1} (\underline{m}^{(j)} - \underline{m}^{(k)}) \qquad (3.4)$$

közelítést kapjuk. Ebben az első tag a Fisher által javasolt <u>lineáris diszkriminancia függvény.</u>

Felmerül a kérdés, vajon mennyit hibázunk, ha (3.3) helyett (3.4)-et használjuk döntésre. Mivel a (3.4)-gyel egy olyan transzformációt adtunk meg, amely az E_N teret egy dimenzióba transzformálja, a hibát megadhatjuk, ha ismerjük a

$$v_{jk} = (\underline{m}^{(j)} - \underline{m}^{(k)})^T \underline{S}^{-1} \underline{\xi} - \frac{1}{2}(\underline{m}^{(j)} + \underline{m}^{(k)})^T \underline{S}^{-1} (\underline{m}^{(j)} - \underline{m}^{(k)}) \qquad (3.5)$$

valószínűségi változó eloszlását. Ehhez először az

$$u_{jk} = \left[\underline{\xi} - \frac{1}{2}(\underline{\mu}_j + \underline{\mu}_k)^T\right] \underline{\Sigma}^{-1} (\underline{\mu}_j - \underline{\mu}_k) \qquad (3.6)$$

eloszlását célszerű meghatározni. Hasonlóan a 2.1.2-beli levezetéshez, belátható, hogy u_{jk} $N(\frac{1}{2}\alpha_{jk}, \alpha_{jk})$ eloszlású, ahol $\alpha_{jk} = (\underline{\mu}_j - \underline{\mu}_k)^T \underline{\Sigma}^{-1} (\underline{\mu}_j - \underline{\mu}_k)$

Mivel a (3.1) és (3.2) a $\underline{\mu}$ és $\underline{\Sigma}$ konzisztens becslései, megmutatható /Anderson (1974)/, hogy v_{jk} eloszlása határesetben megegyezik u_{jk} eloszlásával. Így a $\tilde{d}_{jk}(x)$ függvénnyel való osztályba sorolásnál csak kis hibát követünk el.

3.2 A Fisher féle lineáris diszkriminancia függvény és kapcsolata a Bayes döntésfüggvénnyel

Két osztály szeparálására Fisher (1936) javasolt egy ugynevezett lineáris diszkriminancia függvényt. Megmutatható, hogy ez megegyezik (3.4)-gyel, bár Fisher más úton jutott el hozzá (olyan lineáris függvényként, amely maximalizálja az "osztályok közötti" és az "osztályon belüli" varianciák hányadosát). A Fisher féle diszkriminancia analízis rendkívül gyakori például az orvosi diagnosztikában alkalmazott módszerek között; magát az eljárást számos tanulmány, könyv tárgyalja. Ezek egy része a Fisher féle vagy ahhoz hasonló koncepcióból indul ki, és nem is említi a Bayes döntéssel való kapcsolatát, amelyik pedig említi /Anderson (1974), Kendall (1976), Young (1974)/ ott a bizonyítások különböznek. Mi ezektől kissé eltérő módon mutatjuk be a két módszer egyezését, magát a diszkriminancia analízist pedig Weber (1969) könyve alapján, az ott tárgyalt esetnél általánosabban mutatjuk be.

Legyen adott tehát két osztály és egy tananyag: $\{\underline{x}_1^{(1)}, \dots, \underline{x}_{n_1}^{(1)}\}$, illetve $\{\underline{x}_1^{(2)}, \dots, \underline{x}_{n_2}^{(2)}\}$, amelyről feltesszük, hogy nem csupa azonos elemből áll osztályonként. Jelölje $\underline{m}^{(j)} = (m_1^{(j)}, \dots, m_N^{(j)})^T$ az átlagvektorokat:

$$\underline{m}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} \underline{x}_i^{(j)} \quad (j=1,2) \tag{3.7}$$

Tegyük fel, hogy $\underline{m}^{(1)} \neq \underline{m}^{(2)}$.

Keressük a két osztályt elválasztó függvényt

$$d(\underline{x}) = \underline{b}^T \underline{x} \tag{3.8}$$

alakban, ahol $\underline{b} = (b_1, \dots, b_N)^T$ meghatározása a feladat. A $d(\underline{x})$ -szel tulajdonképpen az E_N-et egy dimenzióba transzformáljuk. Jelöljük az $\underline{x}_i^{(j)}$ vektorból kapott számot $y_i^{(j)}$ -vel:

$$y_i^{(j)} = d(\underline{x}_i^{(j)}) = \underline{b}^T \underline{x}_i^{(j)} \tag{3.9}$$

Célunk (3.8)-ban a $\underline{b}$ vektor olyan megválasztása, hogy $d(\underline{x})$ "jó" elválasztó függvény legyen, azaz a transzformált értékek $(y_i^{(j)} - k)$ osztályonként a lehető legjobban elkülönüljenek. Az elkülönülés annál jobb, minél távolabb esnek az átlagok és a szórások minél kisebbek.

Tekintsük a

$$Q = \frac{D^2}{T} = \frac{\left(\bar{y}^{(1)} - \bar{y}^{(2)}\right)^2}{\sum_{i=1}^{n_1} \left(y_i^{(1)} - \bar{y}^{(1)}\right)^2 + \sum_{i=1}^{n_2} \left(y_i^{(2)} - \bar{y}^{(2)}\right)^2} \tag{3.10}$$

kifejezést, ahol

$$\bar{y}^{(j)} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i^{(j)} , \quad j = 1,2 .$$

Nyilván $Q = Q(b)$, és a két osztály között elkülönülés annál nagyobb, minél nagyobb a Q. Feladat tehát olyan $d(\underline{x}) = \underline{b}^T \underline{x}$ függvény keresése, melyre

$$Q(b) = \frac{D^2(b)}{T(b)}$$

maximális.

Természetesen a $\underline{b} \neq \underline{0}$ megoldások érdekelnek, mivel $\underline{b} = \underline{0}$ esetén transzformáció minden vektort a 0-ba transzformál.

3.1 Definíció A (3.10)-et maximalizáló függvényt Fisher féle diszkriminancia függvénynek nevezzük.

A (3.10) szélső-értékhelyének megkeresésében a b_ℓ -ek szerint differenciálva a következőt kapjuk:

$$\frac{\partial Q}{\partial b_\ell} = \frac{1}{T^2}\left[2D\frac{\partial D}{\partial b_\ell}\cdot T - D^2\frac{\partial T}{\partial b_\ell}\right] = \frac{D}{T}\left(2\frac{\partial D}{\partial b_\ell} - \frac{D}{T}\frac{\partial T}{\partial b_\ell}\right) \quad (\ell=1,\dots,N) \qquad (3.11)$$

(3.11)-at 0-val egyenlővé téve elegendő a második tényezőt vizsgálni, mivel feltételeink szerint $\frac{D}{T}$ csak $\underline{b}=\underline{0}$ esetén lehet egyenlő 0-val.

(3.11)-ből rendezés után kapjuk:

$$\frac{T}{D}\cdot\frac{\partial D}{\partial b_\ell} = \frac{1}{2}\frac{\partial T}{\partial b_\ell}, \quad \ell=1,\dots,N. \qquad (3.12)$$

Kiszámítjuk (3.12) bal- és jobboldalát:

$$\frac{\partial D}{\partial b_\ell} = \frac{\partial}{\partial b_\ell}\left[\sum_{k=1}^{N} b_k\left(m_k^{(1)} - m_k^{(2)}\right)\right] = m_\ell^{(1)} - m_\ell^{(2)}$$

$$\frac{\partial T}{\partial b_\ell} = \frac{\partial}{\partial b_\ell}\left[\sum_{j=1}^{2}\sum_{i=1}^{n_j}\sum_{k=1}^{N}\left[b_k\left(x_{ik}^{(j)} - m_k^{(j)}\right)\right]^2\right] =$$

$$= 2\sum_{k=1}^{N} b_k \sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(x_{ik}^{(j)} - m_k^{(j)}\right)\left(x_{i\ell}^{(j)} - m_\ell^{(j)}\right).$$

Ha bevezetjük a

$$c_{k\ell} = \sum_{j=1}^{2}\sum_{i=1}^{n_j}\left(x_{ik}^{(j)} - m_k^{(j)}\right)\left(x_{i\ell}^{(j)} - m_\ell^{(j)}\right)$$

jelölést, (3.12)-be behelyettesitve a következőt

kapjuk:

$$\sum_{k=1}^{N} b_k c_{kl} = \frac{T}{D}\left(m_l^{(1)} - m_l^{(2)}\right) \qquad l = 1, \ldots, N \qquad (3.13)$$

A (3.13) egyenletrendszert megoldva kapjuk a $\underline{b}$ vektort. A $C = \{c_{kl}\}_{k,l=1}^{N}$ mátrix bevezetésével (3.13) a következő alakba írható: $\underline{C}\,\underline{b} = \underline{m}^{(1)} - \underline{m}^{(2)}$

amiből

$$\underline{b} = \underline{C}^{-1}\left(\underline{m}^{(1)} - \underline{m}^{(2)}\right) \qquad (3.14)$$

Mivel $\underline{C}$ a közös kovariancia mátrix maximum likelihood becslésének $(n_1 + n_2 - 2)$ -szerese/(lásd 3.2) vagy (3.12)/, így a megoldás létezik és egyértelmű.

Hátra van még annak belátása, hogy a (3.16)-tal adott $\underline{b}$ vektor a (3.10)-nek valóban szélső értéke, mégpedig maximuma. Ettől a számítások terjedelmessége miatt eltekintünk, megjegyezve, hogy végül a kovariancia mátrix pozitív definítségén alapul.

Ezzel megmutattuk, hogy ha $\underline{m}^{(1)} \neq \underline{m}^{(2)}$, akkor a (3.10)-et maximalizáló $\underline{b} \neq \underline{0}$ vektor

$$\underline{b} = (n_1 + n_2 - 2)\,\underline{S}^{-1}\left(\underline{m}^{(1)} - \underline{m}^{(2)}\right) \qquad (3.15)$$

összefüggéssel adott, ahol

$$S = \frac{1}{n_1 + n_2 - 2} \sum_{j=1}^{2} \sum_{i=1}^{n_j} \left(\underline{x}_i^{(j)} - \underline{m}^{(j)}\right)\left(\underline{x}_i^{(j)} - \underline{m}^{(j)}\right)^T . \qquad (3.16)$$

Látható, hogy (3.15) - konstanstól eltekintve - megegyezik a (3.4)-beli döntésfüggvényben szereplő együtthatóval. Egyúttal tehát megmutattuk a Fisher féle diszkriminancia analízis és a Bayes döntésfüggvény kapcsolatát is.

A gyakorlatban $d(\underline{x})$ segítségével nyert döntéseknél egy konstans is szerepet játszik: az $y_i^{(j)}$ értékek eloszlása alapján a minimax eljáráshoz hasonlóan most is meg kell határoznunk azt a c küszöbszámot, amelynél a hibás osztályozás valószínűsége a legkisebb. A gyakorlatban ezt a számot az $y_i^{(j)}$ értékből csoportonként készített két hisztogram alapján állapítják meg. A c küszöbszám ismeretében $\sum_{k=1}^{N} b_k x_k \geq c$ esetén az $\underline{x}$ ponttal képviselt objektumot a C_1, egyébként a C_2 osztályba soroljuk.

Ha ezt a konstanst például a következőképpen választjuk: $$c = -\frac{1}{2}(\underline{m}^{(1)} + \underline{m}^{(2)})^T \underline{b} \,,$$

akkor ezt $d(\underline{x})$-hez adva

$$\tilde{d}(\underline{x}) = (\underline{m}^{(1)} - \underline{m}^{(2)})^T \underline{C}^{-1} \underline{x} - \frac{1}{2}(\underline{m}^{(1)} + \underline{m}^{(2)})^T \underline{C}^{-1} (\underline{m}^{(1)} - \underline{m}^{(2)}) \qquad (3.17)$$

alakú döntésfüggvényt kapjuk, amely a (3.4) két osztályra érvényes esete. Tehát ugyanahhoz a diszkriminancia függvényhez jutottunk, mint a Bayes döntésfüggvény normális eloszlás és egyenlő kovariancia mátrixok esetén. Természetesen ezen az alapon történő osztályozás nem garantálja az osztályozás hibájának minimalizálását.

3.3 Logisztikus diszkriminancia analízis

Az eddig ismertetett módszerek többnyire folytonos valószínűségi változókra alkalmazhatók, vagy még inkább azt is feltételezik, hogy a mintacsoportok többdimenziós normális eloszlású populációból származnak. A gyakorlati életben azonban meglehetősen gyakran fordulnak elő olyan

esetek, amikor diszkrét vagy éppenséggel lineáris változók alapján kell osztályozni, illetve a változók egyik része diszkrét, másik része folytonos (pl. kérdőívek adatai).

Diszkrét változókra vonatkozó diszkrimináló módszerekkel Linhart (1959), Cochran és Hopkins (1961), majd Martin és Bradley (1972) foglalkozott. Ezek csak diszkrét illetve dichotóm változókra alkalmas módszereket adtak meg. Általánosabb a Cox (1966) illetve Day és Kerridge (1967) által bevezetett logisztikus diszkrimináció, amely diszkrét és folytonos változókra egyaránt alkalmazható. A Cox-Day-Kerridge modellnek számos általánosítása ismeretes. /Anderson (1972), Albert (1981)/

Az a posteriori valószínűségek logisztikus formában való megadását, valamint az optimális elválasztó függvényre az általános elméleti modellt Day és Kerridge (1967) a következőképpen tárgyalja:

Legyen adott két osztály, C_1 és C_2, legyenek $\underline{\mu}_1$ és $\underline{\mu}_2$ a két osztály átlagvektorai, $\underline{A}$ mátrix egy $N \times N$ -es mátrix, $\Phi(\underline{x})$ pedig tetszőleges integrálható nemnegatív skalárfüggvény.

Ekkor $\underline{\xi}$ megfigyelés feltételes sűrűségfüggvényének általános alakja:

$$f_i(\underline{x}) = P(\underline{x} \mid C_i) = \alpha_i \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \underline{A}^{-1}(\underline{x}-\underline{\mu}_i)\right\}\Phi(\underline{x}) \qquad (3.12)$$

ahol α_i normalizáló konstans, melyet úgy kell választani, hogy $\int_{E_N} f_i(\underline{x})\,d\underline{x} = 1$ teljesüljön.

A (3.12) általános modellből a következő speciális esetek adódnak:

1. (i) $\Phi(\underline{x}) \equiv 1$ és

   (ii) $\underline{A}$: kovariancia mátrix

   Ebben az esetben $f_i(\underline{x})$ a normális eloszlás sűrűségfüggvénye.

2. (i) $\Phi(\underline{x}) = \begin{cases} 1, & \text{ha x minden komponense 0 vagy 1,} \\ 0 & \text{egyébként} \end{cases}$

   (ii) $\underline{A}$: az identitás mátrix

A 2. fejezetben a Bayes döntésfüggvényt a posteriori valószínűségek segítségével adtuk meg, azaz:

$$\hat{p_i}(\underline{x}) = P(C_i \mid \underline{\xi} = \underline{x}) = \frac{p_i f_i(\underline{x})}{\sum_{i=1}^{R} p_i f_i(\underline{x})} \tag{3.13}$$

R=2 esetén

$$\hat{p_1}(\underline{x}) = \frac{\frac{p_1 f_1(\underline{x})}{p_2 f_2(\underline{x})}}{1 + \frac{p_1 f_1(\underline{x})}{p_2 f_2(\underline{x})}} \tag{3.14}$$

Ebből a számlálóba a (3.12) helyettesítésével a következő kifejezést kapjuk:

$$\frac{p_1 f_1(\underline{x})}{p_2 f_2(\underline{x})} = \frac{p_1 \alpha_1 \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)^T \underline{A}^{-1}(\underline{x}-\underline{\mu}_1)\}\Phi(\underline{x})}{p_2 \alpha_2 \exp\{-\frac{1}{2}(\underline{x}-\underline{\mu}_2)^T \underline{A}^{-1}(\underline{x}-\underline{\mu}_2)\}\Phi(\underline{x})} =$$

$$= \exp\left[\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_1)^T \underline{A}^{-1}(\underline{x}-\underline{\mu}_1)\right\} - \left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_2)^T \underline{A}^{-1}(\underline{x}-\underline{\mu}_2)\right\} + \log\frac{p_1\alpha_1}{p_2\alpha_2}\right],$$

tehát

$$\hat{p_1}(\underline{x}) = P(C_1 \mid \underline{\xi} = \underline{x}) = \frac{\exp(\underline{x}^T \underline{b} + c)}{1 + \exp(\underline{x}^T \underline{b} + c)}, \tag{3.15}$$

ahol

$$\underline{b} = \underline{A}^{-1}(\underline{\mu}_1 - \underline{\mu}_2),$$

$$c = -\frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)^T \underline{A}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) + \log\frac{p_1\alpha_1}{p_2\alpha_2} \tag{3.16}$$

Ha $\Phi(\underline{x})$pontosan ismert lenne, a maximum likelihood becsléseket alkalmazva $\underline{\mu}_1, \underline{\mu}_2, \underline{A}$ becslésére, az

$$Y = \underline{X}^T \underline{b} + c \qquad (3.17)$$

lineáris függvényt használhatnánk döntésfüggvényként. Ugyanis

$$\underline{X}^T \underline{b} + c > 0 \; ; \quad \text{ha} \quad \hat{p_1}(\underline{x}) > \hat{p_2}(\underline{x})$$

és fordítva.

Ha $\Phi(\underline{x}) \equiv 1$, akkor (3.17) a Fisher féle lineáris diszkriminancia függvényt adja meg a konstans tagtól eltekintve.

Ha $\Phi(\underline{x})$ ismeretlen, meg kell keresnünk a

$$L = \prod_{\underline{x}_i \in B_1} p_1 P(\underline{x}_i | C_1) \cdot \prod_{\underline{x}_i \in B_2} p_2 P(\underline{x}_i | C_2) \qquad (3.18)$$

likelihood függvény maximumát, ahol B_1, B_2 jelöli azon értékek halmazát, amelyre az $\underline{x}_i$ vektor a C_1 illetve C_2 populációból származik.

(3.16) átalakítható:

$$L = \prod_{\underline{x}_i \in B_1} \frac{\exp(\underline{x}_i^T \underline{b} + c)}{1 + \exp(\underline{x}_i^T \underline{b} + c)} \cdot \prod_{x_i \in B_2} \frac{1}{1 + \exp(\underline{x}_i^T \underline{b} + c)} \qquad (3.19)$$

(3.19) logaritmusát véve $\underline{b}$ és c alkalmas iteratív eljárással (pl. Newton-Raphson) meghatározható.

Day és Kerridge eredményeit Anderson (1972) általánosította. Megadta több osztály esetén az a posteriori valószínűségek logisztikus formában történő előállítását, amely az előző jelölésekkel:

$$P(C_j | \underline{X}) = p_{j\underline{x}} = \frac{\exp(\underline{x}^T \underline{b}_j + c_j)}{P(C_R | \underline{X})} \quad , \; j = 1, \ldots, R-1,$$

$$P(C_R | \underline{X}) = p_{R\underline{x}} = \frac{1}{1 + \sum_{j=1}^{R-1} \exp(\underline{x}^T \underline{b}_j + c_j)} \qquad (3.20)$$

Az Anderson által megadott likelihood függvényben az együtthatók becslése annyiban általánosabb, hogy feltesszük, hogy most a minták minden egyes populációtól különkülön állnak rendelkezésre.

3.4 Probit diszkriminancia analízis

Az elnevezést Albert és Anderson (1981) vezették be. A módszer a 3.2 fejezetbeli módszerek további általánosítása, lényege a következő:

Legyen adott két osztály és (N+1) dimenziós vektorok, mint tanulópontok: $(x_0, \underline{x}^T)$, ahol $\underline{x} = (x_1, \ldots, x_N)^T$ és $x_0 \leq a$ (a konstans) esetén a pontok a C_1,

$x_0 > a$ esetén pedig a C_2 osztályhoz tartoznak. Korábbi cikkükre hivatkozva a szerzők állítása szerint, ha x_0 adott $\underline{x}$ melletti feltételes sűrűségfüggvénye $N(\beta_0 + \underline{\beta}^T \underline{x}, \tau^2)$ az $\underline{x}$ eloszlásától függetlenül, akkor az a posteriori valószínűségek

$$P(\underline{x}) = \Phi(\alpha_0 + \underline{\alpha}^T \underline{x})$$

alakban írhatók, ahol $\Phi(.)$ a standard normális eloszlás eloszlásfüggvénye:

$$\Phi(u) = \frac{1}{2\pi} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt,$$

$$\alpha_0 = \frac{\beta_0 - \alpha}{\tau}, \quad \underline{\alpha} = \frac{\beta}{\tau}.$$

Az $\alpha_0 + \underline{\alpha}^T \underline{x}$ függvényt a "probit diszkriminancia függvénynek" nevezték. Az Anderson-féle (3.20) képlet írható

$$P(\underline{x}) = \Lambda(\alpha_0 + \underline{\alpha}^T \underline{x}) \quad \text{alakban,}$$

ahol $\Lambda(t) = \dfrac{\exp(t)}{1+\exp(t)}$

Albert és Chappelle(1981) az $\alpha_0, \underline{\alpha}$ paraméterek becslésére a feltételes likelihood függvényt

$$L_1(\alpha_0, \underline{\alpha}) = \prod_{i=1}^{\hat{n}} \left[\{1 - \phi(\alpha_0 + \underline{\alpha}^T \underline{x}_i)\}^{1-z_i} \{\phi(\alpha_0 + \underline{\alpha}^T \underline{x}_i)\}^{z_i} \right] \quad (3.21)$$

alakban adja meg, ahol $z_i = 0$, ha $x_0 \leq a$, $z_i = 1$, ha $x_0 > a$.

Mivel eloszlására semmiféle kikötés nincs, a módszer tetszőleges valószínűségi változók esetén használható.

4. Az alakfelismerés alkalmazási lehetőségei az orvosi diagnosztikában irodalmi áttekintés és saját tapasztalatok alapján

4.1 A számítógéppel segített orvosi diagnosztika szükségessége és módszerei

Az előző fejezetekben bemutatott módszereket sok más - e tanulmányban nem tárgyalt-módszerrel együtt elterjedten alkalmazzák orvosi diagnosztikai problémákra. Természetesen felmerül a kérdés, mi teszi szükségessé e módszerek, illetve a rájuk alapozott számítógépes programok alkalmazását az orvosi diagnosztikában?

A gépek és az ilyen típusu módszerek szükségszerű felhasználását ezen a területen különféle objektív és szubjektív érvekkel támasztják alá. Igen meggyőzőek például azok az érvek, amelyek az emberi agy korlátozott információfeldolgozó kapacitását a számítógép elvileg korlátlan kapacitásával vetik össze /De Dombal (1972), Grémy és Goldberg (1977)/. Más szerzők arra mutatnak rá, hogy a számítógépes diagnóziskészítés által felvetett problémák megoldása lényeges hatással van a "természetes" diagnóziskészítésre is, mivel együtt jár az orvosi tudás további rendszerezésével és a diagnosztizáló folyamat tudatosabbá válásával.

Az orvos a diagnózis megállapításánál figyelembe veszi a betegről gyűjtött információkat, az orvosi ismeretanya-

got és saját tapasztalatait. Az esetek egy részében nem sikerül a helyes diagnózist megadni. Szentgáli 1976-ben írt könyvében több betegség esetén ad számot irodalmi adatok alapján a hibás diagnózisok gyakoriságáról (csak egy kiragadott példa: a vakbélgyulladást az esetek 60 %-ában sikerült csak helyesen diagnosztizálni!). A hibás diagnózisok okait a világon és hazánkban is többen vizsgálták. Ezek általában a következők:

- a diagnózis megállapításához szükséges információk száma (betegségek, tünetek, a forgalomban lévő gyógyszerek stb.) egyre nő;
- ugyanakkor az emberi agy egyszerre csak viszonylag kisszámú információval tud dolgozni;
- bizonyos ritka betegségek nem ismerése;
- a klinikai leletek, vizsgálatok rossz minősége;
- az orvos személyisége;
- alkalmatlan diagnosztikus eljárások használata.

A számítógéppel segített orvosi diagnosztika módszerei nem korlátozódnak csupán az alakfelismerés alkalmazására, felhasználásuk információelméleti, logikai, algebrai (fuzzy halmazok) módszerek, és nagyon sok számítógépes diagnosztikáló rendszer működik"természetes" diagnóziskészítés utánzásával, intuitív, heurisztikus algoritmusokkal /például Shild és munkatársai (1978)/.

4.2 A statisztikus alakfelismerési módszerek alkalmazása az orvosi diagnosztikában

4.2.1 Világirodalmi áttekintés

Az alakfelismerés alkalmazása triviálisan adódik az orvosi diagnosztikában: a betegről (objektum) a tünetei alapján (jellemzők) el kell dönteni, hogy adott betegségek (kategóriák) valamelyikében szenved-e.

A valószínűségszámítás Bayes tételének alkalmazását Ledley és Lusted (1959) javasolta először. Alapötleteiket és általánosításaikat igen elterjedten alkalmazzák /Bühovszkij (1961), Warner (1972), Fritz (1978), Fryback (1978), Sonnenberg (1982)/.

A statisztikus alakfelismerés módszerei közül igen gyakran szerepel a diszkriminancia analízis. Az írások egy része a "hagyományos" eljárásokat csupán alkalmazzák /Knapp (1977)/, másik része a meglevő eljárások finomítását is célul tűzi ki, vagy a hibabecsléssel foglalkozik, illetve más statisztikai módszereket is használ mellette. A diszkriminancia analízis gyakoriságának több oka is lehet. Indokolhatja a biológiában gyakran előforduló normális eloszlás. De másik oka lehet, hogy az ismert és világszerte elterjedt nagy statisztikai programcsomagokban (BMDP, SPSS stb.) igen jól kidolgozott programjai állnak rendelkezésre. Ezeket a programokat számos betegség differenciál-diagnosztikájára alkalmazták. Például Brooks és munkatársai (1971) a BMDP stepwise diszkriminancia analízis programját alkalmazták érelmeszesedés-

ben bekövetkező stádiumok szétválasztására jó eredménnyel. A diszkriminancia analízis széles körű használatára utal, hogy például az epidemiológia témakörén belül is számos problémára alkalmazták, ezt elemzik Lachenbruch és munkatársai (1980) 36 irodalmat áttekintve.

A statisztikus módszerek alkalmazásának másik gyakori módja több módszer együttes alkalmazása. Ennek két fő célja lehet:

1. Lényegkiemeléssel egybekötött statisztikai alkalmazás (például faktoranalízis, osztályozó algoritmus előtt). Ilyen például Hott és Metton (1980) cikke: 13 folytonos változó alapján 10 betegre nézve 6 féle májbetegség fennállásának lehetőségeit vizsgálták. Faktor és clusteranalízist alkalmaztak a változók csoportosítására, majd az ezek alapján lényegesnek tartott változókon futtatták le a diszkriminancia analízis programját.

2. A módszerek összehasonlítása - azaz ugyanazon a tananyagon több módszert is lefuttattnak. Az adathalmaz lehet valódi vagy szimulált. Igen érdekes ebből a szempontból Rioux és Nakache (1979) cikke, amely a diszkriminancia analízis különböző módszereit, változatait hasonlítja össze. Legkisebb hibájának a kvadratikus, leggyorsabbnak pedig a lineáris diszkirminancia analízist találták. Duewer és munkatársai (1978) májbetegségek differenciáldiagnosztikája során többféle statisztikai módszert alkalmaztak. (normalizálás, Fisher-súlyok, Karhunen-Loéve sorfejtés, k-NN módszer, SIMCA) A módszerek

összehasonlítása adott klinikai adatbázisra vonatkozóan történt.

A statisztikai módszerek diagnosztikai alkalmazásai között nem túl gyakori, de nem elhanyagolható rész a cluster analízis felhasználása. Egyrészt lényegkiemelési módszerként használják - változók clusterezésére - például Abott és Metton (1980) - másrészt önállóan - például Wong, Liu (1976), akik új iteratív algoritmusokat szimulációs és klinikai adatokon teszteltek.

A fejezethez megadott irodalomjegyzék nem teljes, de meg kell említenem, hogy számos bibliográfia létezik, például Wagner és munkatársai (1978) témakörök szerint csoportosítva több mint 800, Wilke (1978) 7000 irodalmi hivatkozást sorol fel, a Cacoullos által szerkesztett könyben S. M. Gupta, illetve a szerkesztő egy 547 cikkből álló bibliográfiát ad meg.

4.2.2 Hazai alkalmazások

Magyarországon a számítógéppel segített orvos-diagnosztikai kisérletek az 1960-as években kezdődtek. Kalmár László már 1969-ben felhívta a figyelmet erre a lehetőségre. /Kalmár (1969)/. A számítógépes orvosi diagnosztika elterjedése azonban csak az 1970-es években indult meg a számítógépek számának növekedésével. Az elterjedés sem volt gyors, és nehezítette, hogy sokan kételkedtek benne, még azok is, akik egyébként magának a számítógépnek a szerepét az egészségügyben elismerték. A kisszámú alkal-

mazás nem volt elég átütő ahhoz, hogy ez a terület általánosan ismert legyen. Ezt szemlélteti az a tény, hogy például az Orvos és Technika folyóiratban 1972-74-ben megjelent "A számítógépek és orvosbiológiai alkalmazásuk" című sorozat 9 cikkéből csak egy említi a diagnosztikai lehetőségeket /Fedina (1974)/, az is speciális esetben.

A hazai irodalomban fellelhető közlemények alapján egy tanulmányban /Boda (1980)/ foglaltuk össze a számítógépes diagnosztika hazai helyzetét. Az információelméleti, logikai módszerek mellett főként a valószínűségszámítás és a statisztikus alakfelismerés módszereinek az alkalmazása a jellemző. A valószínűségszámítás Bayes tételét alkalmazta Győri (1979) a tünet információmértékének kidolgozásakor, Prónay és Hunya (1970) bélamoebiasis differenciáldiagnosztikájára. Kanyár és munkatársai (1972) négy pajzsmirigy betegség diagnosztizálásához alkalmazták a Bayes tételt. A feltételes valószínűségek becsléséhez a tünetek függetlenségét tételezték fel, az eredmény 92, 65, 51 és 9 %-ban egyezett az orvosi véleménnyel.

Alakfelismerési módszerek legfontosabb alkalmazási területei:

a/ <u>EKG görbék automatikus kiértékelése</u> terén Győrfi és Csibi (1972) a tanuló algoritmusokat, elsősorban a potenciálfüggvényes algoritmust használták fel, és 1971-ben egy kisszámítógépes EKG diagnosztikai rendszert dolgoztak ki /Battisting és munkatársai (1971), Bak és munkatársai (1971), Kobzos és munkatársai (1978)/. Később a KFKI-ben

TPA i 16 K kisszámítógépre terveztek egy görbekiértékelő és osztályozó programrendszert, ahol az alkalmazott módszerek a NN, és egy cluster algoritmus /Szlávik és munkatársai (1976), Bolyki és munkatársai (1978)/.

b/ A koraszülés okai, következményei, a koraszülöttek súlyát befolyásoló tényezők és mindezek előrejelzésének szükségessége többek részéről és egymástól függetlenül is felmerült. A SZOTÉ-n lineáris diszkriminancia analízissel választottuk szét a vércserére kerülő és kontroll eseteket. /Boda és munkatársai (1970, 1971)/. Paksy és Győrfi (1973) a KSH adatokból véletlenszerűen kiválasztott újszülöttek adatain futtatták le a Bayes döntésfüggvény közelítésére írt iteratív algoritmust annak érdekében, hogy eredményeiket szűrésre is felhasználják. Sebők és munkatársai (1974, 1976) a koraszülöttek koponyaűri vérzését befolyásoló tényezőket vizsgálták. A Kashyap-Ho féle gradiens algoritmus egy általuk kidolgozott gyorsított változatát és egy lineáris programozási módszert alkalmaztak.

c/ Légzésfunkciós vizsgálatokban főleg szűrővizsgálatok kidolgozása volt a cél. Például Csukás és munkatársai (1975) faktoranalízis után futtatták le a módosított McQueen cluster algoritmust, amellyel jól elkülönülő betegség-osztályokat kaptak.

d/ EEG görbék alapján alvásfázisok elkülönülését vizsgálják Pál és munkatársai (1978) lineáris szeparálás, Committe Machine, NN és lineáris diszkriminancia analízis segítségével.

e/ A kérdőívek alkalmazásával főleg tömeges szűrővizsgálatokhoz nyerünk adatokat. A tömeges szűrővizsgálatok célja esetlegesen beteg elválasztása a biztosan egészségesektől. A kiszűrt lehetséges betegeket ezután tüzetesebb orvosi ellenőrzésnek vetik alá. A valószínűleg beteg személyek kiválasztásával az alakfelismerési módszerek újabb alkalmazási területe a szűrés. B. Nagy és munkatársai (1976, 1977) kérdőívek alapján tüdő- illetve szívbetegek szűrésére, illetve pszichológiai vizsgálatokban a neurotikusok felfedezésére alkalmazott tanuló algoritmusokat. Szilárd és munkatársai (1979) keringésrendszeri kérdőíves előszűrés modelljének kialakításáról számolnak be. A beteg és egészséges populáció szétválasztására a diszkriminancia analízist találták a legalkalmasabbnak. /Csébfalvi és munkatársai (1977) Hunya és munkatársai (1980) neurotikus betegekről kérdőívvel gyűjtött adatokon faktor- és clusteranalízist végeztek./

f/ Az eddig felsoroltakon kívül egyedi jellegű alkalmazások történtek: például a leukoplákiás betegek osztályozása cluster analízissel /Fenyő és munkatársai (1975)/, májbetegségek differenciáldiagnosztikája /Győri és munkatársai (1970)/. Számos képfelismeréssel kapcsolatos kutatás folyik: Hajnal (1977), Naszlady és munkatársai (1978), Csernay és munkatársai (1970, 1976).

Összefoglalva a magyarországi helyzetet, a leggyakrabban a differenciáldiagnózis felállítására és tömeges

szűrés megvalósítására alkalmaztak főként alakfelismerési módszereket. Ezek széles körű orvosi felhasználása egyes területeken már megindult.

4.3 Saját tapasztalataink az alakfelismerési módszerek orvosi diagnosztikai illetve szűrővizsgálati célokra történő alkalmazásával

A Szegedi Orvostudományi Egyetemen több esetben merült fel annak szükségessége, hogy tünetek, laboreredmények birtokában a diagnózis meghatározását matematikai módszerekkel segítsük. Három orvosi probléma kapcsán szerzett tapasztalataink a következők:

1./ 1970-ben az újszülöttkori sárgaságot befolyásoló tényezőket vizsgáltuk 2000, a SZOTE Gyermekklinikáján vizsgált csecsemő adatai alapján. A súlyos esetekben az újszülötteken vércserét kellett végrehajtani, ami nem veszélytelen beavatkozás. Felmerült az a kérdés, vajon lehet-e osztályozni az újszülötteket a sárgaságot befolyásoló tényezők alapján úgy, hogy a vércsere szükségessége a lehető legnagyobb biztonsággal eldönthető legyen. Orvosi és előzetes statisztikai (például korrelációszámítás) meggondolások alapján az eredeti 13-ból végül 5 tényező (jellemző) maradt. Ezek közül kettő, az Incompatitivitás és a Coombs próba három, illetve két értékű változó volt. Ezeket úgy vettük figyelembe, hogy a kombinációknak megfelelően 6 csoportba osztottuk a beteganyagot és csoportonként a megmaradt három folytonos paraméter alapján diszkriminancia analízissel kíséreltük meg a vércserés és nem vércserés esetek szétválasztását

/Boda és munkatársai (1971, 1972)/. A helyes döntések aránya 77,3 % - 89,8 % volt. A modellt újabb eseteken is leteszteltük, itt a helyes döntések aránya közel azonos volt a tananyagon kapott százalékkal. A tesztanyagban szerepeltek olyan esetek is, akiket vércsere helyett az újabban bevezetett "kékfénybesugárzással" kezeltek. A modell a besugárzás után az eseteket a "nem kell vércsere" osztályba sorolta. A biztató kisérletek végül is nem kerülhettek gyakorlati alkalmazásra, mivel a kékfénybesugárzás veszélytelenebb, és csaknem helyettesítette a vércserét.

2. A hasnyálmirigy (-pancreas) különböző megbetegedéseinek diagnosztizálása hosszadalmas, sok laboratóriumi vizsgálatot igénylő feladat. Ebben a pancreas-enzimek vizsgálata az egyik legfontosabb tényező. A többféle sejttípus és a pancreas-enzimek nem párhuzamos szekréciója miatt egyes funkciók korábban károsodhatnak, miközben mások még normális értékűek maradnak. Egyetlen pancreas-funkciós próba alapján a hasnyálmirigy működéséről nem lehet biztos véleményt alkotni; ezért minél több tényező együttes értékelésére kell törekedni.

A pancreas különböző funkcióit jelző adatokat többen próbálták már matematikai-statisztikai úton értékelni és azt a hasnyálmirigy-betegségek diagnosztikájában hasznosítani, Schmidt és munkatársai (1977) szekvenciális statisztikai eljárásokat alkalmaztak. Durbec és munkatársai

(1978) a kontingencia táblából nyert log kereszt-szorzat arányokra és a kölcsönös információs tulajdonságokra alapított adatszűrést írtak le. Sarles (1970) illetve Capitaine (1971) egy- és többváltozós módszereket alkalmaztak a különböző hasnyálmirigy betegek és egészségesek diagnosztizálására. Mivel az enzimaktivítások eloszlása méréseik alapján lognormális, esetenként pedig normális volt, a Fisher féle diszkriminancia analízist nem tartották megfelelőnek, így helyette új analitikai módszert dolgoztak ki. Aparisi és munkatársai (1979) a széklet chymotripszin diagnosztikus értékének meghatározására sikeresen alkalmazták a lineáris diszkriminancia analízist (a hibás osztályozás valószínűsége 0,39 % volt); Farini és munkatársai (1982) pedig biokémiai és klinikai adatok alapján próbálkoztak négy hasnyálmirigy-betegség és egy egészséges csoport elkülönítésével a lineáris diszkriminancia analízis módszerét alkalmazva. Összesen 97 vizsgált személy adata állt rendelkezésükre, melyből 19 volt a kontroll, csak a biokémiai jegyek alapján 57,1 % volt a helyesen "diagnosztizált" esetek aránya. Klinikai adatok hozzávételével ez az arány 75 %-ra nőtt. Konklúziójuk éppen ez, a klinikai adatok hozzávétele megerősítette a korrekt diagnózist.

Az általunk vizsgált paraméterek és összehasonlításuk

A mi vizsgálatainkat a SZOTE I. sz. Belgyógyászati

Klinikáján fekvő hasnyálmirigy-betegeken illetve kontroll eseteken végzett mérések alapján végeztük. Az adatok matematikai-statisztikai módszerekkel való feldolgozása mintegy 6 éve kezdődött, Pap Á. és munkatársai (1977), legújabb eredményeinkről 1981-ben és 1983-ban számoltunk be /Boda és munkatársai (1981, 1983)/. 14 laboratóriumi paramétert vizsgáltunk, amelyeket a duodenum szondázással nyert három fő pancreas-enzim (amitáz, lipáz, tripszin) 3 x 10 perces frakciókban mért aktivitásából, továbbá a duodenum nedv térfogatából számoltunk - a frakciók maximumát, átlagát és összegét véve alapul. A paramétereket, valamint 77 kontroll eset átlagát és szórását a 4.1 táblázat első két oszlopa szemlélteti.

Kezdetben /Pap Á., (1977)/ a fenti paraméterek kóros és egészséges eseteken mért átlagait és szórásait t-illetve Wilcoxon próbával hasonlítottuk össze. A különbség - bár szignifikáns volt - nem adott az orvosnak újat a diagnózis felállításához - éppen a nempárhuzamos működés miatt. A mért paraméterek pozitív számok, és minél kisebbek, annál inkább utalnak a kóros elváltozásra. A non-parallel szekréció azt jelenti, hogy még egy pancreas -beteg embernek is lehetnek bizonyos paraméterértékei normálisak. Az orvos szerint még az a meglepő tény is előfordul, hogy ugyanannak az embernek egy másik alkalommal megmérve esetleg egy másik enzim értéke

mutat kóros elváltozást, mint a korábban mért. Ez a tény igen megnehezíti a döntést a betegségről, és ezzel magyarázható az is, hogy hisztogramot készítve a normál és a kóros esetek eloszlásáról, az átfedés viszonylag nagy. (4.1 ábra)

A döntéshez először meg kell határozni azt a határt (küszöbértéket), amelynél kisebb érték kórosnak tekinthető. Ezt a számot normálértéknek nevezik.

A normálértékek meghatározása

Az irodalom szerint a szokásos gyakorlat /Sarles (1970) az, hogy a normálértéket valamely egészséges populációból származó minta alapján határozzák meg. Általában a mintának a logaritmikus transzformáció utáni átlagából a kétszeres standard deviációt levonjuk:

$$N = \exp(\bar{y} - 2s_d), \tag{4.1}$$

ahol x_i, $i=1,\dots,n$ a mintaelemek, $y_i = \log x_i$,

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_d^2 = \left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)/(n-1)$

Ez az eljárás a paraméterek ferde - leggyakrabban lognormális - eloszlása miatt terjedt el.

Megvizsgáltuk a 77 kontroll eset eloszlását, χ^2 próbával normalitásvizsgálatot végeztünk. A 4.1 táblázat 3. oszlopa a kapott valószínűségeket mutatja. Látható, hogy a paraméterek nagy részére a próba szignifikáns. A logaritmikus transzformáció után újabb χ^2 próbával még mindig volt a szignifikáns eset. Tehát vizsgált paramétereink között volt normális, lognormális, sőt még ferdébb el-

	Paraméter neve		Átlag	± SEM	P	2.5 % perc	Normál érték (4.1) formula
AMILÁZ	Átlagos aktivítás	IU/ml	99.81	+ 5.242	0.2272	39.4	38.9
	Maximális aktivítás	IU/ml	141.60	± 7.584	0.4952	58.5	53.4
	Összegezett output	IU/30min	24835.0	±1724.939	0.021	8913.6	8660.9
	Maximális output	IU/10min	15234.1	±1307.298	0.0006	5174.2	4611.6
TRIPSZIN	Átlagos aktivítás	mIU/ml	521.82	± 23.042	0.7131	231.6	219.1
	Maximális aktivítás	mIU/ml	759.92	± 37.197	0.5574	300.7	289.1
	Összegezett output	IU/30min	120.59	± 6.355	0.0084	49.4	45.4
	Maximális output	IU/10min	75.01	± 5.242	0.0001	29.1	23.7
LIPÁZ	Átlagos aktivítás	IU/ml	81.31	+ 2.811	0.0169	41.5	41.6
	Maximális aktivítás	IU/ml	107.81	± 3.329	0.1356	66.0	61.0
	Összegezett output	IU/30min	17868.19	± 753.807	0.9409	8792.6	7880.0
	Maximális output	IU/10min	10601.04	± 548.262	0.0057	3728.0	3586.0
	Összegezett térfogat	ml/30min	229.22	± 6.689	0.9409	141.0	131.0
	Maximális térfogat	ml/10min	127.90	± 6.053	0.0001	68.0	57.0

4.1 táblázat Az alkalmazott paraméterek és 77 kontroll eset adataiból számított jellemzők.

oszlásból származó is. Ezért nem használhattuk a fenti képletet. Mivel az említett módszerrel kapott érték az eloszlás 2.5 %-át hagyja figyelmen kívül, kiszámoltuk a 2.5 % percentilis becsléseit is. A két módon kapott normálértékeket az 4.1 táblázat utolsó két oszlopa mutatja. Nyilvánvalóan nem egyenlő mindig a két szám, konkrét beteg esetén az orvos azonban mindkét értéket figyelembe vette.

Alakfelismerési módszerek alkalmazása

Miután a paraméterenkénti vizsgálat nem vezetett eredményre, egy adott betegről való döntés esetén, természetes módon vetődött fel a több paramétert egyszerre figyelembe vevő alakfelismerési módszerek alkalmazása.

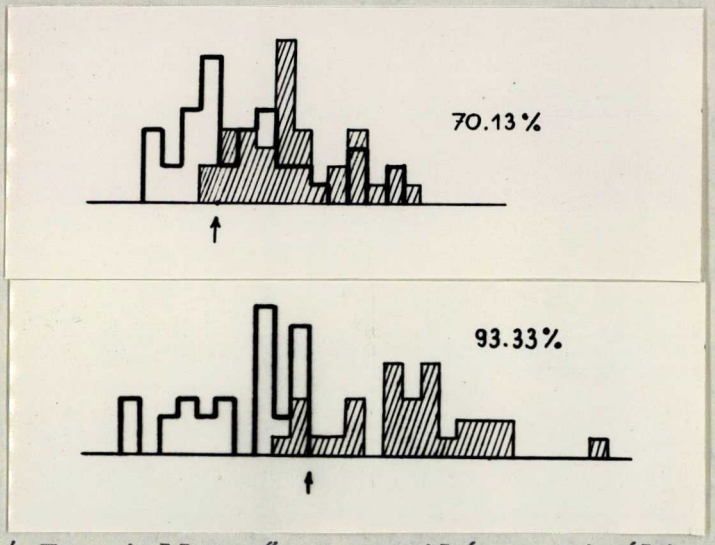
Három módszerrel is próbálkoztunk: a súlypont, a legközelebbi szomszéd és a lineáris diszkriminancia analízis módszerrel.

Az adatokat újabb 137 eset laboreredményei képezték, ebből 67 volt a kontroll és 70 a beteg. Az adathalmaz egy része (30 + 30 eset) szolgált tananyagul, a többin pedig a tesztet végeztük. Az eredményeket a 4.2 ábra mutatja 14 jellemző alapján. Az egész számok a táblázatban a helyes döntések számát jelentik, a legalsó sorban pedig az össz-esetszám van feltüntetve. A 4. és 8. oszlopban álló törtszámok a helyes döntést fejezik ki az össz-esetszám százalékában. Például a

	Tananyag				Tesztanyag			
M ó d s z e r	Kontroll	Pancr.	Össz.	%	Kontroll	Pancr.	Össz.	%
Diszkriminancia analízis	26	30	56	93.33	35	36	71	92.2
Legközelebbi szomszéd	23	25	48	80.00	36	26	62	80.52
Súlypont	24	29	53	88.33	35	31	66	85.71
Összes esetszám:	30	30	60		37	40	77	

4.2 táblázat Az osztályozás eredménye három módszerrel 14 jellemző figyelembevételével.

lineáris diszkriminancia analízissel a tananyagon 30 kontroll közül 26 esetben, 30 beteg közül 30 esetben kaptunk helyes döntést. Ez összesen 56 eset, amely a 60-nak 93.33 %-a. Látható a táblázatból, hogy ez egyúttal a legjobb eredmény a három módszer közül, melyet a teszthalmazon nyert helyes döntések aránya is megerősít (92.2 %). A 4.1 ábrával szemléltetjük a módszert: a felső rész az egyik jellemző eloszlását mutatja kóros és normál esetben; az alsó rész pedig a lineáris diszkriminancia függvénnyel egy dimenzióba transzformált pontok eloszlását mutatja.



4.1 ábra. a/ Egy jellemző, az amiláz maximális koncentráció eloszlása kóros és normál (vonalkázott) esetben.

b/ A lineáris diszkriminancia analízissel egy dimenzióba transzformált pontok eloszlása kóros (sima) és normál (vonalkázott) esetben.

A 14 jellemzőből 7-7-et elhagytunk annak vizsgálatára, vajon mennyire romlik a döntés. Amint a 4.3 illetve 4.4 táblázatból látható, hogy valamennyi esetben, ha jelentéktelenül is, de romlott a döntés biztonsága.

Tudomásunk szerint enzimértékek alapján lineáris diszkriminancia analízissel eddig nem értek el ilyen megbizható döntést.

Sarles és munkatársai (1970) pedig nem is tartották adekvát módszernek az adataik ferde eloszlása miatt. A mi vizsgálatunkban is zavaró volt a ferde eloszlás, azonban a módszer 3.1 pontbeli megfogalmazása - mely szerint a Fisher féle diszkriminancia függvény egy optimalítási kritériumnak eleget tevő transzformáció - alapján indokoltnak tartottuk az alkalmazást. Eredményeink is megerősítik ezt, és további előnye a lineáris döntésfüggvények alkalmazásának, hogy folyamatos tesztelésükhöz, a döntésfüggvény adott esetben történő helyettesítési értékének a kiszámításához elegendő egy kis kézi számológép is.

A súlypont módszert egyszerűsége miatt végeztük el, tájékozódó jelleggel. A legközelebbi szomszéd módszerrel kapott kedvezőtlenebb eredményeink - véleményünk szerint - azzal magyarázhatók, hogy a programunk csak euklideszi

távolsággal tud dolgozni (jelenleg).

Az eredményeinkről szóló cikk a Computers in Biology and Medicine c. lapban került elfogadásra és várhatóan ez évben megjelenik.

Azóta még egy módszer programját futtattuk le adatainkban: a probit diszkriminancia analízissel az eredeti teszthalmazon teljes szétválasztást sikerült elérni. Ha pedig a teszt- és tananyagot egyesítettük tananyagnak, összesen 5-öt tévesztett.

Eredményeink tesztelését folyamatosan végezzük; más vizsgálatok és orvosi paraméterek alapján tervezzük módszereink további alkalmazását.

Helyes döntés

	Tananyag				Tesztanyag			
Módszer	Kontroll	Pancr.	Össz.	%	Kontroll	Pancr.	Össz.	%
Diszkriminancia analízis	27	28	55	91.67	35	34	69	89.61
Legközelebbi szomszéd	26	24	50	83.33	35	29	64	83.12
Súlypont	25	29	54	90.00	35	31	66	85.71
Összes esetszám:	30	30	60		37	40	77	

4.3 táblázat Az osztályozás háromféle módszerrel kapott eredménye 7 jellemző alapján (a maximumokat kihagyva).

	Helyes döntés							
	Tananyag				Tesztanyag			
Módszer	Kontroll	Pancr.	Össz.	%	Kontroll	Pancr.	Össz.	%
Diszkriminancia analízis	29	26	55	91.67	37	31	68	88.31
Legközelebbi szomszéd	23	26	49	81.67	36	26	62	80.52
Súlypont	24	29	53	88.33	35	31	66	85.71
Összes esetszám:	30	30	60		37	40	77	

4.4 táblázat Az osztályozás eredménye háromféle módszerrel 7 jellemző alapján (a koncentrációk kihagyva).

3. A légúti megbetegedések kimutatásának egyik fontos módszere az ugynevezett flow-volumen görbe vizsgálata. Ezt a görbét egy speciális műszer adja belégzés utáni erőltetett kilégzéskor. A görbére jellemző, hogy eleinte gyorsan emelkedik, majd többé-kevésbé meredeken csökken.

A görbéről - a csúcs helyéről, laposságáról stb. - a gyakorlott orvos sokszor ránézésre is felismeri az asztmát vagy a felső légutak megbetegedését. Vannak azonban olyan esetek, amikor ez nem egyértelmű.

A SZOTE Gyermekklinikán levő légzésfunkciós laboratóriumban egy kisszámítógéppel egybekötött készülék működik, amely e görbét vizsgálja. Felmerült e kisszámítógépbe szűrővizsgálati céllal olyan döntőkészülék beépítése, amely a görbéből nyert paraméterek alapján jelzi a megbetegedésre gyanús eseteket.

Az első lépéseket tettük meg ezirányban, probléma ugyanis a figyelembe veendő jellemzők megválasztása.

Az irodalom nem egységes e téren, ma is sokan vizsgálják a görbe egyes jellemzőit illetve ezeknek a testmagasságtól való függését. Probléma a paraméterek "normálértékeinek" meghatározása /Sahakian (1981)/.

Előzetes vizsgálatainkat /Gyurkovits, Boda (1981)/ 24 kontroll és 39 asztmás beteg gyerek fújási görbéin végeztük. Az asztmások közül 20 a vizsgálat időpontjában tünetmentes volt, ezeket külön csoportnak tekintettük. A vizsgált paramétereket a 4.3 táblázat mutatja.

1. testsúly
2. testmagasság
3. FEV1 - 1 sec. alatti térfogat
4. FEV összes térfogat
5. TIFF = FEV1/FEV x 100
6. FMAX áramlási csúcssebesség
7. F75 sebesség a vitálkapacitás 75 %-ánál
8. F50 sebesség a vitálkapacitás 50 %-ánál
9. F25 sebesség a vitálkapacitás 25 %-ánál
10. T(FV) a görbe alatti terület
11. DMAX = F25/FMAX
12. D75 = F25/F75
13. D50 = F25/F50
14. M 50 % és 75 % közötti meredekség
15. TA F50 alatti terület aránya az egész területhez
16. MEFR középáramlási max. sebesség

A paraméterek fele megegyezik a Sahakian (1981) által vizsgáltakkal, a többi is irodalmi adat. Az általunk bevezetett új paraméter a meredekség.

E paraméterek alapján lefuttattuk a lineáris diszkriminancia analízis programot, először 16, majd pedig az egymással korreláló tüneteket elhagyva, 9 tünet alapján.

Eredményeinket a 4.5 táblázat szemlélteti. A táblázatban az egyes csoportok elnevezéseit rövidítettük: K-kontroll, Atm-asztmás, tünetmentes, Ast-asztmás, "státusban". A csoportokat páronként hasonlítottuk össze. Az eredményeket 2x2-es táblázatokban szemléltetjük, a sorokban az eredeti, az oszlopokban a döntött osztályok gyakoriságai állnak. Például az első táblázatban a 24 kontroll esetből 23 esetet döntött a kontroll csoportba, és 1 esetet döntött a tünetmentes asztmásnak. A táblázatok alján a helyes döntések aránya található %-ban.

Eredeti \ Döntött	K	Atm	Össz.
K	23	1	24
Atm	1	19	20

95.45 %

Eredeti \ Döntött	Atm	Ast	Össz.
Atm	19	1	20
Ast	2	18	20

92.5 %

Eredeti \ Döntött	K	Ast	Össz
K	24	0	24
Ast	1	19	20

97.72 %

4.5 táblázat 16 koordináta alapján lineáris diszkriminancia analízissel kapott döntési táblázata és a helyes döntések aránya %-ban kontroll (K), tünetmentes asztmás (Atm) és asztmás (Ast) esetek vizsgálatakor

Döntött

Eredeti		K	Atm	Össz.
	K	23	1	24
	Atm	1	19	20

95.45 %

Döntött

Eredeti		Atm	Ast	Össz.
	Atm	19	1	20
	Ast	4	16	20

87.5 %

Döntött

Eredeti		K	Ast	Össz.
	K	24	0	24
	Ast	2	18	20

95.45 %

4.6 táblázat 9 koordináta alapján lineáris diszkriminancia analízissel kapott döntési táblázatok és a helyes döntések aránya %-ban kontroll (K), tünetmentes asztmás (Atm) és asztmás (Ast) esetek vizsgálatakor.

A vizsgált jellemzők közül elhagytuk azokat, amelyek között szoros összefüggés volt, így 9 jellemző maradt. A diszkriminancia analízis eredményét a 4.6 táblázat mutatja.

Örvendetes az a tény, hogy a tünetmentes asztmásokat is nagy biztonsággal sikerült a kontroll esetektől

elkülöníteni. Természetesen további, nagyszámú eset szükséges a finomabb vizsgálatokhoz és a teszteléshez, hogy a módszer szűrővizsgálatra is alkalmazható legyen.

E három példával kívántam szemléltetni azt, hogy az általunk beprogramozott - kisgépen is alkalmazható - alakfelismerési módszereket hogyan használtuk fel az orvosi diagnosztika különböző területein.

5. Összefoglalás

E dolgozatban egyes statisztikus alakfelismerési módszereket tekintettem át, valamint ezek orvos-diagnosztikai felhasználási lehetőségeit mutattam be irodalmi adatok és saját esettanulmányok alapján.

Az alakfelismerés alapproblémájának általános ismertetése után - mely szerint az alapvető feladat objektumoknak bizonyos jellemzőik alapján adott kategóriák valamelyikéhez való hozzárendelése - leírom azt a döntést, amelynél a hozzárendelés hibájának valószínűsége a legkisebb (Bayes döntés). Ezután a dolgozat az úgynevezett tanítóval való tanulás módszerei közül mutat be néhányat. Megmutatom, hogy a Bayes döntésfüggvény normális eloszlás és egyenlő kovariancia mátrixok esetén lineáris, és e függvény egy optimalizálási transzformáció útján is megkapható. Ez a Fisher féle diszkriminancia analízis. /Fisher, R. A. (1936)/. Ennek általánosításait is ismertetem, melyek nem csak normális eloszlású valószínűségi változókra alkalmazhatók. /Anderson, S. A. (1972), Albert, A. (1981)/.

A dolgozat irodalmi adatok alapján áttekintést ad az alakfelismerés orvos-diagnosztikai felhasználásának lehetőségeiről, kitérve a számítógéppel segített orvosi diagnosztika helyzetére külföldön és hazánkban.

Végül leírom azokat a diagnosztikai célú vizsgálatokat, amelyeket a Szegedi Orvostudományi Egyetem orvosaival együttmüködve végeztem. Egyik esettanulmányom a gyermek-

József Attila Tudományegyetem Központi Könyvtára SZEGED

kori légúti betegségek vizsgálata, ahol bizonyos légzésfunkciós paraméterek alapján kellett következtetni az asztmás megbetegedés fennállására. A másik probléma, a hasnyálmirigy gyulladás diagnosztizálásakor laborparaméterek (enzim értékek) alapján vizsgáltuk a betegség meglétét. A lineáris diszkriminancia analízissel kapott 93 %-os helyes döntést az utólagos tesztek is megerősítették, így a döntésfüggvény segítségével egy paciensről nagy valószínűséggel megmondható, hogy egészséges-e vagy beteg. Eredményeink - melyekről több közlemény ad számot,-orvosi szempontból is értékesek, és az orvosi kutatást és gyakorlatot is segíthetik.

Köszönetet mondok Dr. Győri Istvánnak, a SZOTE Számítástechnikai Központ vezetőjének az értekezés elkészítésében nyújtott támogatásért, hasznos elméleti és gyakorlati útmutatásért.

Ezúton mondok köszönetet Dr. Csirik Jánosnak, a JATE Számítástudományi Tanszék tudományos főmunkatársának az értekezés elolvasása kapcsán tett számos igen hasznos észrevételéért.

Köszönettel tartozom Dr. Pap Ákosnak és Dr. Gyurkovits Kálmánnak, a SZOTE adjunktusainak a felmerült orvosi vonatkozású problémák megoldásában nyújtott segítségéért.

Ezúton mondok köszönetet a SZOTE Számítástechnikai Központban a matematikai statisztikával foglalkozó munkacsoport minden egyes tagjának, akikkel számos közös publikáció mutatja az együttmüködést.

6. Irodalomjegyzék

6.1. Irodalomjegyzék az 1-3. fejezetekhez

1. Albert, A., Anderson, J.A.(1981): Probit logistic discriminant functions.Commun.Statist. 10, 641.
2. Albert, A., Chappelle, J.P., Smeets, J.P.(1981): Stepwise Probit Discrimination with Specific Application to Short-Term Prognosis in Acute Myocardial Infarction. Comput. Biomed, Res. 14, 391-398.
3. Anderberg, M.R,(1973): Cluster Analysis for Applications. Academic Press New York and London.
4. Anderson, J.A.(1972): Separate sample logistic discrimination. Biometrika 59, 19-35.
5. Anderson, J.A.(1973): Logistic Discrimination with Medical Application. In "Discriminant Analysis and Applications", Ed. by T. Cacoullos, Academic Press New York and London. 1-15.
6. Anderson, T.W.(1974): An Introduction to Multivariate Analysis. John Wiley and Sons.
7. Cochran, W.G., Hopkins, C.E.(1961): Some classification problems with multivariate qualitative data, Biometrics 17, 10-32.
8. Cover, T., Hart, P.(1967): The Nearest Neighbour Classification, IEEE-IT 13, 21-27.
9. Cox, D.R.(1970): The Analysis of Binary Data. Methuen: London.

10. Day, N.E., Kerridge, D.F.(1967): A general maximum likelihood discriminant. Biometrics, 23, 313-323.

11. Duda, R. O., Hart, P.E.(1973): Pattern Classification and Scene Analysis. John Wiley and Sons.

12. Duran, B.S., Odell, P.L.(1974): Cluster Analysis. A Survey. Lecture Notes in Economics and Mathematical Systems, 100. Springer Verlag Berlin,Heidelberg, New York.

13. Fisher, R.A.(1936): The use of multiple measurements in taxonomic problems. Ann. Eugen 7, 179-188.

14. Freeman, H.(1962): On the digital computer classification of geometric line patterns. Proc. Nat. Elect. Conf. 18, 312-324.

15. Fritz, J., Révész, P.(1974): Az alakfelismerés statisztikus módszerei. MTA tanfolyam jegyzet, Budapest.

16. Harman, H.H.(1976): Modern Factor Analysis. Chicago.

17. Hartigan, J.A.(1975): Clustering Algorithms. John Wiley and Sons.

18. Hughes, G.I.(1968): On the mean accuracy of statistical pattern recognizers. IEEE IT-14. 55-63.

19. Kendall, M., Stuart, A.(1977): The Advanced Theory of Statistics. Charles Griffin Co.

20. Lachenbruch, P.A., Goldstein, M.(1979): Discriminant Analysis. Biometrics 35, 69-85.

21. Linhart, H.(1959): Techniques for discriminant analysis with discrete variables. Metrika 2, 138-149.

22. Martin, D.C., Bradley, R.A.(1972): Probability models, estimation and classification for multivariate

dichotomous populations. Biometrics, 28, 203-221.

23. Meisel, W.S.(1972): Computer Oriented Approaches to Pattern Recognition. Academic Press New York, London.
24. Meyer-Brötz, G., Schürmann(1972): Methoden der automatischen Zeichenerkennung. Academie Verlag Berlin.
25. Nilsson, N.J.(1965): Learning Machines. Mc Graw Hill, New York.
26. Nollau, V.(1975): Statistische Analysen. VEB Fachbuchverlag. Leipzig.
27. Parzen, E.(1962): On Estimation of a Probability Density Function and Mode. Annals of Math. Statist. 33, 1065-1076.
28. Patrick, E.A., Fisher, F.P.(1970): Generalized k -Nearest Neighbour Decision Rule. J. Information and Control 16, 2, 128-152.
29. Rao, C.R.(1973): Lineare statistische Methoden und ihre Anwendungen. Akademie Velag Berlin.
30. Raudys, S.J.(1979): Determination of optimal dimensionality in statistical pattern classification. Pattern Recognition 11,2, 263-270.
31. Rosenblatt, F.(1961): Principles of neurodinamics: Perceptrons and the theory of brain mechanism. Spartan Books, Washigton.
32. Specht, D.F.(1967): Genaration of polinomial discriminant functions for pattern recognition. IEEE Trans. on Electronic Computers, 16, 3, 308-319.

33. Tou, Y.T., Heydorn, R.P.(1967): Some approaches to optimum feature extraction. Comp. and Inf. Sciences II. Academic Press New York.

34. Wagner, T.(1971) : Convergence of the nearest neighbour rule. IEEE IT 17, 566-571.

35. Wald, A.(1950): Statistical Decision Functions. Wiley, New York.

36. Watanabe, S., Lampert, P.F., Kulikowski, C.A., Buxton, J.L., Walker, R.(1967) : Evaluation and selection of variables in pattern recognition.Computer and Information Sciences II, Academic Press New York. 9o-122.

37. Weber, E.(1969) : Grundriss der biologischen Statistik. Thieme Verlag 4. 428-448.

38. Wolverton, C.T., Wagner, T.I.(1967): Asymptotically optimal discriminant functions for pattern classification. IEEE IT 13.

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

6.2. Irodalomjegyzék a 4. fejezethez

Világirodalom

1. Abbott, L.A., Metton, J.B. (1980): Diagnoses generated by numerical taxonomic methods, applied to standard blood variables. Meth. Infirm. Med. 19, 4, 205.
2. Aparisi, L., Rodrigo, J., Montes, F., Bel, J., Guix, J., Serra, M., Bixquert, M., Wassel, A. (1979): Application of linear discriminatory analysis to determine the diagnostic value of fecal chymothrypsin. "European Pancreatic Club 12th Meeting" Copenhagen, 11-13. oct.
3. Bühovszkij, M.L., Visnyevszkij, A.A., Harnasz, S.S. (1969): Ekszper. Hirurg i Anaszt. 4.3.
4. Brooks, S.H., Crawford, D.W., Sebar, R.H., Blankenhorn, D.H., Brandt, R. (1978): Discrimination of human arterial pathology by computer processing of angiograms for serial assessment of atherosclerosis change. Comput. Biomed, Res. 11, 469-480.
5. Cacoullos, T., Styan, G.T.H. (1973): A bibliography of discriminant analysis. In" Discriminant Analysis and Applications", Ed by Cacoullos, Academic Press New York and London.
6. Capitaine, Y., Cros, R., Gullo, L., de Barros. Mott, C., Pastor, J., Pauli, A.M., Sarles, H. (1971): Étude multivariée des données de l'exploration pancréatique par tubages. Biologie et Gastro-Entérologie IV, 301-308.

7. De Dombal, F.T. (1972) : Computer-assisted diagnosis. In "Principles and Practice of Med. Comp " 179-199.

8. Duewer, P.L., Kowalski, B.R., Clayson, K.J., Roly, R. J. (1978) : Elucidating the structure of some clinical data. Comput. Biomed. Res. 11, 567-580.

9. Durbec, J.P., Cornée, J., Berthezene, P. (1978) : Data Screening Methods- Application to Differential Diagnosis in Pathology from Radiological Signs. Meth. Inform. Med. 17, 1. 33-40.

10. Das Gupta, S. (1973) : Theories and Methods in Classification: A Review . In "Discriminant Analysis and Applications". Ed. by Cacoullos. Academic Press New York and London.

11. Farini. R., Piccoli, A., Fabrio, G., Del Favero, G., Grassi, F., De Beri, T., Redravocali, S., Baccaglini, U., Brosolo, P., Naccarato, R. (1982) : Discriminant Analysis in the Biochemical and Clinical Diagnosis of Chronic Pancreatitis and Pancreatic Cancer. Digestion 25, 1, 28.

12. Fritz, K.J., Polasick, M.A., Potts, A.M. (1978) : Computer Assisted Diagnosis for Opthalmology. Comput. Biol. Med. 8, 3, 223-228.

13. Fryback, D.G. (1978) : Bayes'theorem and conditional non-independence of data in medical diagnosis. Comput. Biomed. Res. 11, 423-434.

14. Grémy, F., Goldberg, M. (1977) : Decision Making Methods in Medicine. In "Medizinische Informatik and Statistic 3". Springer Verlag Berlin, Heidelberg, New York.

15. Knapp, B.G., Levi, S., Lurie, D., Westphal, M.(1977): A computer generated diagnostic decision guide: A comparison of statistical diagnosis and clinical diagnosis. Comput. Biol.Med. 7.223-230.

16. Knocke, J.D.(1982): Discriminant Analysis with Discrete and Continuous Variables. Biometrics 38, 191-200.

17. Lachenbruch, P.H., Clarke, W.R.(1980): Discriminant analysis and its applications in epidemiologie. Meth. Inform. Med. 19, 4, 221.

18. Ledley, R.S., Lusted, L.B.(1959): Reasoning Foundations of Medical Diagnosis.

19. Rioux, P., Nakache, J.P.(1979): Discriminant analysis: methods and program. Computer Programs in Biomedicine 1o, 43-47.

20. Rogers, W., Ryack, B., Hoeller, G.(1979): Computer-aided Medical Diagnosis: Literature Rewiew.

21. Sahakian, J.(1981) : A Pocket Computer Program for Normal Values of Selected Lung Function Tests in Children and Young Adults. Comput. Biol. Med. 11, 4, 231-235.

22. Sarles, H., Salem, E., Cros. R., Figarella, C.(1970): Exploration pancréatique par tubages. Acta gastro-ent. belg. 33, 303-318.

23. Schmidt, B., Cornée, J., Delachaume-Salem, D., Mercier, J 1977 : Application des procedures statistiques séquentielles à l'étude des concentrations enzymatiques du suc pancréatique humain normal. Comtes rendus des séances de la Société Biologie, 164, 8-9, 1813.

24. Shild, W., Lunenfeld, B., Gavish, B.(1978): Computer aided diagnosis with application to endocrinology, IBM Journal of Research and Development 22, 518-532.

25. Sonnenberg, A.(1982): Sequential iteration of Bayesian formula by pocket calculator and its use in clinical routine. Comput. biol. Med.12, 4, 357-360.

26. Wagner, G., Tautu, P., Wolber, U.(1978): Problems of Medical Diagnosis - A Bibliography. Meth. Inform. Med. 17,1, 55-82.

27. Warner, H.R., Rutherford, B.D., Hontcheres, B.(1972): A sequential Bayesian approach to history taking and diagnosis. Comput. Biomed. Res. 5,256-262.

28. Wilke, J. 1978 : Bibliographie multivariate Statistik und mehrdimensionale Klassification. Akademie Verlag Berlin.

Hazai alkalmazások

1. Bak J., Kobzos L., Gulyás O.(1971): EKG görbék automatikus kiértékelése és szeparálása tanuló algoritmusok felhasználásával. Neumann János Számítógéptudományi Társaság /NJSZT/ 2.kollokviumának kiadványa. Szeged. 167-175.
2. Battisting, Gy., B.Nagy A., Rét A., Ungvári L.(1971): Elektrokardiogramok számítógépes feldolgozásának rendszertechnikai és hardware feltételei. NJSZT 2. kollokvium kiadványa, 153-166.
3. B.Nagy A., Németh J.(1976): Ishaemiás szívbetegségek meghatározása kardiohistogram és klinikai kérdőív együttes vizsgálatával. NJSZT 7. kollokviumának kiadványa, Szeged. 431-438.
4. B.Nagy A., Ghyczy K.(1977): Kérdőív felhasználása betegség és vele kapcsolatos terápia meghatározására. Orvos és Technika 15, 7.121-124.
5. Boda K., Győri I., Kovács Z.(1971): Sárgaság miatt felvett újszülöttek adatainak értékelése diszkriminancia analízissel. NJSZT 2. kollokviumának kiadványa, 71-79.
6. Boda K, Kovács Z., Győri I., Veres I.(1972): Újabb tapasztalataink az újszülöttkori sárgaságnál alkalmazott diszkriminancia analízissel. NJSZT 3. koll. 203-208.
7. Boda K., Forczek E., Győri I.(1980): R-10 számítógépre készült statisztikai programrendszer. Videoton Software Tájékoztató 80/1.33-36.
8. BodaK., Pap Á., Grácin E.(1980): Alakfelismerési módszerek alkalmazása hasnyálmirigy gyulladás diagnosztizálására. NJSZT 10. kollokvium kiadványa, Szeged.

9. Boda K.(1980): Számítógéppel segített orvos-diagnosztikai törekvések Magyarországon. Tanulmány, készült az "Alakfelismerés alkalmazása a biológiában, az orvostudományban és a kémiában" c. pályázat keretében.

10. Boda K., Forczek E.,Pap Á., Mayer J., Eller J.(1980): Statisztikai és alakfelismerési eljárások alkalmazása a STASYS programrendszerrel biometriai és orvosi diagnosztikai feladatokra. Kórház és Orvostechnika 19, 3, 89-92.

11. Boda K., Pap Á.(1983): Diagnostics of pancreatic insufficiency using multivariate statistical and pattern recognition methods. Computers in Biology and Medicine /megjelenés alatt/.

12. Bolyki J., Kozmann Gy., Szlávik F., Wolf T., Antalóczy Z., Regős L., Bukosza I., Bíró S.(1978): A KFKI_OTKI kisszámítógépes EKG diagnosztikai rendszerfejlesztés tapasztalatainak áttekintése. NJSZT 9.koll.446-458.

13. Csernay L., Csirik J.(1970): Számítógép használata a szcintigráfiás eljárás diagnosztikus problémáinak megoldásában. NJSZT 1.koll. 111-116.

14. Csébfalvi Gy., Szilárd I., Tényi J.(1977): Nemparaméteres statisztikai módszerek a kérdőíves cardiovascularis előszűrés adatainak számítógépes kiértékelésében,NJSZT 8.koll.109-118.

15. Csirik J., Scherer F., Kuba A., Nemessányi Z., Csernay L. (1976): A SEGAMS képfeldolgozó programjaival szerzett tapasztalatok. NJSZT 7.kollokviuma, Szeged. 9-17.

16. Csukás A., Galgóczy G., H.Gaudi I., Mándi A.(1975): A légzésfunkciós elváltozások vizsgálata faktor és cluster analízis segítségével. NJSZT 6.kollokvium kiadványa Szeged. 339.o.

17. Fedina L.(1974): A számítógépek és orvos-biológiai alkalmazásuk. Orvos és Technika 11, 6, 161-168.

18. Fenyő I., Bánóczy J., Sima D., Siminszki M.(1975): A cluster analízis alkalmazása leukoplákiás betegek carcinoma veszélyeztetettségének megállapítására. NJSZT 6.kollokvium kiadv. Szeged. 353.o.

19. Győrfi L., Csibi S.(1972): Az orvosdiagnosztika segítése tanuló algoritmusokkal. NJSZT 3. koll. 181-183.

20. Győrfi L.(1972): Estimation of probability density and optimal decision functions in RKHS. Submitted to the Proceedings of European Meeting of Statisticians, Budapest.

21. Győri I.(1970): Megjegyzések a diagnosztikai eljárások matematikai modellezéséhez. NJSZT 1.koll.91-97.

22. Győri I., Nagy I., Benedek Sz., Boda K.(1972): Különböző statisztikai és alakfelismerési módszerek tapasztalatai májbetegségek differenciáldiagnosztikájában. NJSZT 3.koll. 145-148.

23. Gyurkovits K., Boda K.(1981): Flow-volume curve analysis by different biometrical methods. SEPCF Congressus Working Group for Ped. Resp. Physiol. Göteborg.1981.jún.

24. Hajnal M., Pályi A., Székely Gy.(1977): Daganatsejtek kromoszómáinak számítógépes osztályozása. NJSZT 8. kollokvium kiadványa Szeged. 19-27.

25. Hunya P., Pethő Z., Vas Z., Eller J.(1980): Clusteranalízis alkalmazása az önértékelő kérdőívvel történő pszichometriai vizsgálatokra. MBT Biometriai Szakoszt. szakülése Budapest, 1980. febr. 21.

26. Kalmár L.(1969): Digitális számológépek és célgépek alkalmazása az orvosi diagnosztikában. Orvos és Technika 1, 14-18.

27. Kanyár B., Garas Zs.(1972): Számítógép alkalmazása a pajzsmirigybetegségek diagnózisához. Orv. Hetilap 113, 25-27.

28. Kobzos L., Kerekes I., Wolf T.(1978): Az R-10 alapú kisérleti diagnosztikai mintaállomás üzembe helyezésének tapasztalatai. Orvos és Technika 15, 3.

29. Kovács Z., Dudás B., Boda K., Győri I., Veres I.(1973): Computer method of discriminancy analysis in the diagnosis of newborn jaundice. XVIII.Congr. Ped. Checosl. Bratislava, 1973. 09.4-7.

30. Naszlady A., Hajnal M., K.Tihanyi K., Loványi I., Vajta L.(1978): EKG görbék osztályozása nehezen detektálható görbeszakaszok mint képek számítógépes feldolgozása alapján. NJSZT 9.koll.404-425.

31. Paksy A., Győrfi L.(1973): A születési súlyt befolyásoló néhány tényező vizsgálata számítógépes tanuló algoritmus segítségével. NJSZT 4.koll.329-335.

32. Pap Á., Varró V.(1977): Lundh-teszt, a pancreas exocrin funkciójának értékes klinikai vizsgáló módszere. Orv. Hetil. 118, 2078-2080.

33. Prónay G., Hunya P.(1970): Entamoeba histolicita okozta bélamoebasis differenciáldiagnosztikájának gépi segítséggel történt feldolgozása. NJSZT 1. kollokviumának kiadványa Szeged. 135-141.

34. Pál I., Bak J., Halász P., Kundra O., Rajzo P., Simon G., Vargha M.(1978): Az alvásfolyamat dinamikájának vizsgálata osztályozó algoritmusokkal. NJSZT 9.koll. 329-340.

35. Sebők J., Fritz J., Kiszel J., Paksy A., Srajber B. (1974): Pathológiás terhességi és szülési tényezők jelentőségének vizsgálata koraszülöttek koponyaűri vérzésének kialakulására, lineáris diszkriminációs módszerekkel. NJSZT 5.koll. 443-452.

36. Sebők J., Srajber B., Paksy A., Kiszel J. (1976): Koraszülöttek koponyaűri vérzése okainak vizsgálata lineáris diszkriminációs módszerekkel. Orv.Hetil. 117, 19, 1143-1146.

37. Szentgáli Gy.(1976): A számítógép kórházi alkalmazása Közg. és Jogi Könyvkiadó, Budapest. 13-24.

38. Szerényi L., Kasza F., Stéhlik J-né, Győri I., Forczek E. (1976): Adatfeldolgozás céljait szolgáló programrendszer R-10-es számítógépre, különös tekintettel a biometriai feldolgozásokra. NJSZT 7.kollokviumának kiadványa.

39. Szilárd I., Csébfalvy Gy., Tényi J., Bihari L., Veszprémi B.(1979): Számítógépes kiértékeléssel egybekötött keringésrendszeri előszűrés modelljének kialakítása I-II.

Népegészségügy, 60,5, 257-271.

40. Szlávik F., Antalóczy Z., Kozmann Gy., Bolyki J., Maros I. (1976): Kisszámítógépes interaktív EKG diagnosztikai rendszerfejlesztés célkitűzései. NJSZT 7. kollokviumának kiadványa Szeged. 387-399.

Függelék

A STASYS programrendszer és a benne levő statisztikus alakfelismerési módszerek programjainak ismertetése

A STASYS programrendszer R-10-es számítógépre készült statisztikai programrendszer. 1976-ban hoztuk létre TIM munkában a SZOTE Számítástechnikai Központ munkatársaival. /Boda és munkatársai (1980)/. A rendszer célja az egyetemen folyó orvos-biológiai kisérletek adatainak kiértékelése matematikai-statisztikai módszerekkel.

A STASYS rendszer vektorstruktúrájú adatok kiértékelésére alkalmas, off-line rendszer. Alapegysége egy vektor, amely változó hosszúságú lehet, és amelynek egy sorszám az azonosítója. Egy vektor általában egy statisztikai mintának felel meg, de a feladattól függően lehetnek gyakoriságok is az elemei.

A rendszer diszk-orientált, azaz a feldolgozandó vektorokat, valamint a programokat és egyes statisztikai táblázatokat is diszken tárolja. A vezérlő program a rendszergenerálás után a diszken különböző zónákat hoz létre, majd biztosítja az adatok input-outputját és a programok batch-jellegű futtatását.

Az adatok ki- és bevitelét egy rendszerprogram végzi. Fő jellemzője, hogy egységes adatkezelést biztosít az adatok mennyiségétől függetlenül. Az adatokat for-

mátum nélkül, tetszőleges adathordozóról be lehet olvasni, és a futás bármely fázisában a teljes vagy részleges rendszermentés megvalósítható.

A programok automatikus, batch jellegű futtatása egy egyszerű célnyelven megadott utasítás-sorozat alapján valósítható meg. Egy másik rendszerprogram végzi e célnyelv feldolgozását és ellenőrzését.

A STASYS rendszerhez az eddig említett rendszerprogramokon kívül az úgynevezett feldolgozó programok tartoznak. Ezek vagy valamilyen transzformációt hajtanak végre egyes vektorok elemein - ezáltal új rendszerbeli vektorokat hoznak létre - vagy egy adott statisztikai módszert futtatnak le adott sorszámú vektorokon. A célnyelven tulajdonképpen e programok sorrendje és paraméterei adhatók meg.

A rendszer egyik fontos tulajdonsága a programból programhívás lehetősége. Ez azt jelenti, hogy futó feldolgozó program futását megszakítva egy másik rendszerbeli programot indíthat el. Ezzel a statisztikai módszerek egymásra épülése miatt fa-struktúraszerűen elágazó feladatok is megoldhatók - viszonylag kis memória mellett.

A programok nyelve tetszőleges, az R-10-en használható nyelv lehet. A rendszerprogramok és egyes szubrutinok assembey nyelven készültek, a feldolgozó programok nyelve fortran.

A feldolgozó programok a diszket egy speciális

József Attila Tudományegyetem Számítóközpont SZEGED

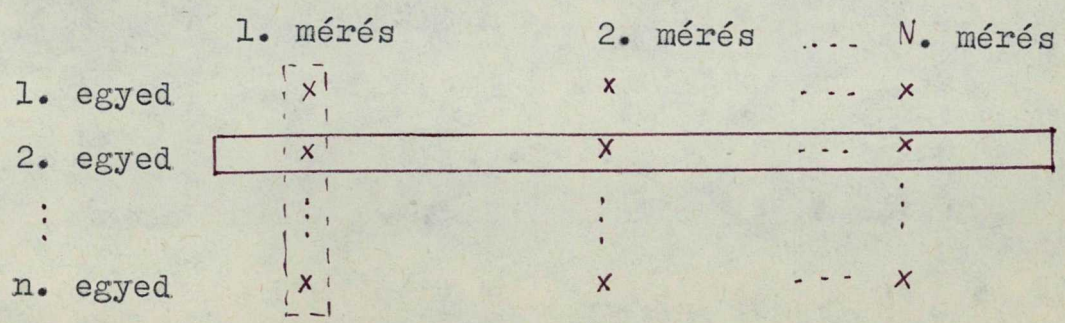
szubrutin-rendszer segítségével érhetik el. Ezek a szubrutinok valósítják meg a file-kezelést a diszken.

A STASYS rendszerről az évek folyamán bebizonyosodott, igen jól alkalmazható rutinszerű feldolgozásoknál. Ugyanakkor nagyobb méretű, nem rutin feladatok is megoldhatók benne, mivel új program rendszerbe illesztése igen egyszerű.

Az alakfelismerési módszereket igénylő feladatok nem gyakoriak és nem is rutin jellegűek. E módszerek programjait mégis a STASYS rendszerbe illesztettük, mivel itt

- az adatkezelés, file-kezelés megoldott,
- a programból programhívás lehetősége jól felhasználható,
- a rendszerbeli egyéb statisztikai programokra, transzformáló programokra alakfelismerési problémák esetén is szükség van.

Az alakfelismerési programok rendszerbe illesztésénél egyetlen problémát az jelentett, hogy itt egy egyedről készített méréseket volt célszerű vektorként felfogni. Ez a statisztikai programok felfogásával szemben - a vektorstruktúra elve szerint - transzpolálást jelent. Ezt szemlélteti az F. 1. ábra



F. 1. ábra

Az alakfelismerési programok input adatai tehát egy egyedről készült mérések vektorai. Ha a vektor a tananyagba vagy a tesztanyagba tartozik, akkor a "tanítás" koordináta a vektor első eleme kell, hogy legyen, 1,2,..., R egész számokkal kódolva.

Az alakfelismerési programoknak három fő tevékenysége:

1./ A tanuló anyag alapján a döntésfüggvény előállítása,

2./ adott döntésfüggvény és tesztanyag alapján tesztelés,

3./ adott döntésfüggvény alapján ismeretlen kategóriájú pontról döntés.

E leírásnak nem célja a programok működésének részletezése, csupán a dolgozatban említett és az esettanulmányok futtatása során felhasznált módszerek programjainak rövid leírását adjuk meg.

Súlypont módszer

Feladata: adott sorszámú oszlopok által meghatározott tananyagon döntésfüggvény készítés a súlypont módszerrel; a döntésfüggvény tesztelése és adott sorszámú pontokról döntés.

Forrásnyelve: FORTRAN

Periféria igény: a szokásos rendszer-hozzárendeléseken kívül még M:EI-ről esetenként olvas

Input:

1. oszlopsorszámok. Többparaméteres program, az oszlopok egy-egy pontot határoznak meg (tanuló vagy tesztpont)

2. konstansok:

= 1: tanulás

= 2: adott tanulópontok tesztelése output altáblában lévő súlyvektor segítségével

= 3: döntés ismeretlen kategóriáju pontokról output altáblában lévő súlyvektor segítségével

= 4: adott tanulópontok tesztelése M:EI-ről beolvasott súlyvektor segítségével

= 5: döntés M:EI-ről beolvasott súlyvektor segítségével

3. Diszk input: KESZIT program által kitöltött altáblák

4. Az M:EI-hez rendelt perifériáról 4-es és 5-ös konstans esetén a következőket olvassa be:

(2II∅) formátummal koordináták és osztályok száma

(4F2∅5) formátummal a súlyvektor koordinátái osztályonként. Minden osztály új kártyán (rekordban) kell, hogy kezdődjön!

Megkötések a méretre: koordinátaszám ≤ 50

koordinátaszám x osztályok száma ≤ 500

Output

Sornyomtatón: két osztályt elválasztó hipersík egyenlete,

a súlypontok koordinátái osztályonként,

tesztelés eredménye (sorszám, eredeti és döntött osztály, döntési mátrix)

döntés eredménye (sorszám, döntött osztály)

A program működése:

A konstansok alapján diszkről (subroutine diszk) vagy kártyáról (subroutine kártya) beolvassa a koordináták összegét illetve az átlagokat, kiszámítja a hipersíkok egyenleteinek együtthatóit (subroutine hiper), majd tesztel vagy dönt (subroutine teszt). Lásd az F.2. ábrát!

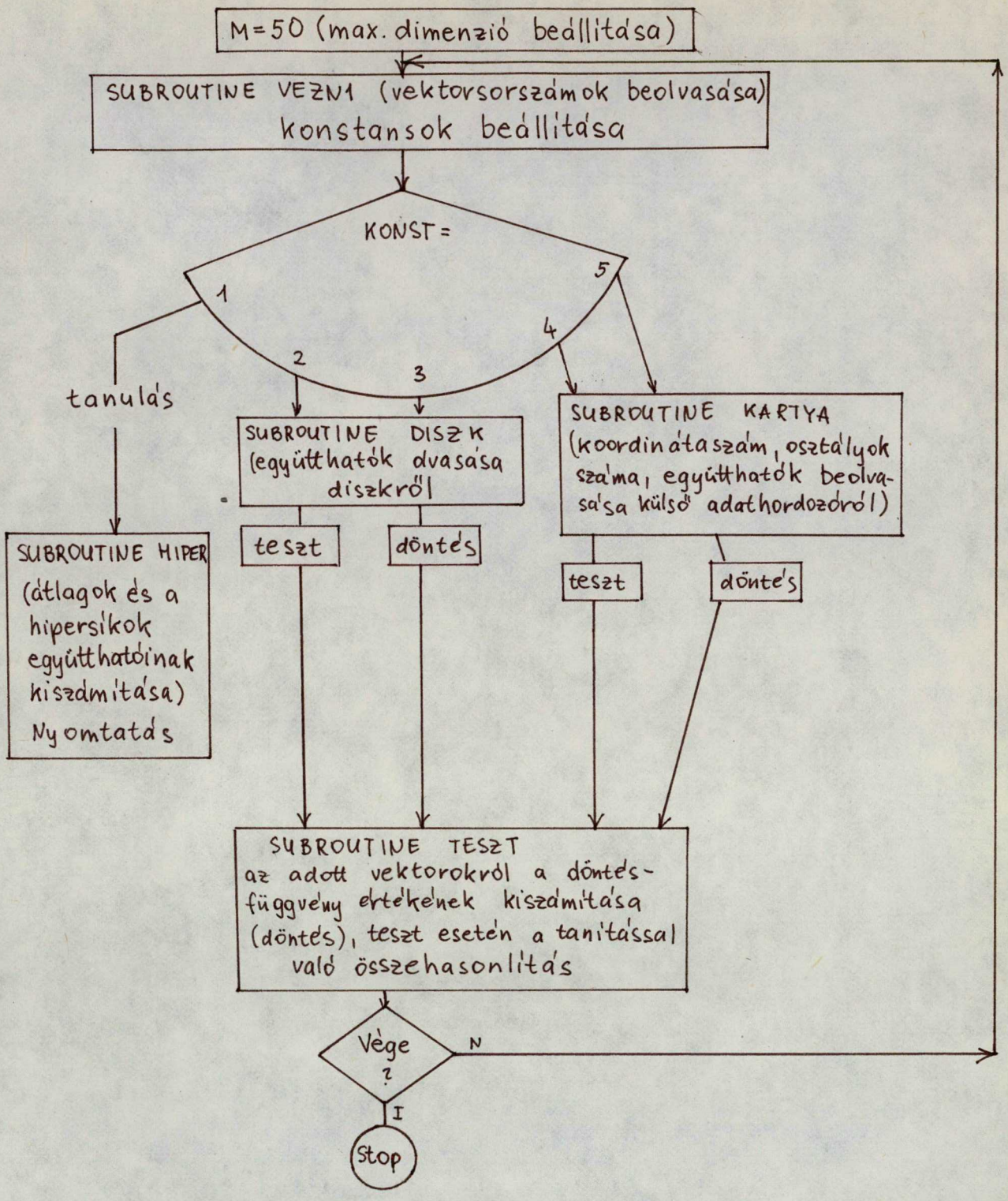
Szubrutinok: VEZNI

DISZK

KARTYA

HIPER

TESZT



F.2. ábra. A súlypont módszer programjának blokkdiagramja

Diszkriminancia analízis

Feladata: adott sokszámú vektorokon mint tananyagon a Fisher-féle lineáris diszkriminancia analízis elvégzése; azaz két osztály között bizonyos értelemben optimális lineáris elválasztó függvény előállítása.

Forrásnyelve: FORTRAN

Periféria igény: a szokásos statisztikai rendszerbeli periféria hozzárendeléseken kívül esetleg még kártyaolvasó inputra kell számítani.

Program input:

1. oszlopsorszámok. Többparaméteres program. Az oszlopsorszámok által megadott vektorok jelenleg maximum 20 koordinátásak lehetnek.

2. Konstansok.

1. konstans: tevékenységszám. Értékei:

= 1: tanulás adott tananyagon

= 2: tesztanyag tesztelése előző eredmény alapján (azaz a döntésfüggvény e futás során készült vagy a 25. output altáblában van) /alapfeltételezés/.

= 3: teszt kártyáról beolvasott paraméterek segítségével

= 4: előző eredmény alapján ismeretlen kategóriájú egyedekről döntés

= 5: kártyáról beolvasott együtthatók alapján ismeretlen kategóriájú pontokról döntés

2. konstans: az első osztály tanítása, mely természetes számmal lett kódolva (alapfeltétel: = 1)

3. konstans: a második osztály tanítása, mely természetes számmal lett kódolva (alapfeltételezés: = 2)

3. Diszk input: a KESZIT program által elkészített 4,6 output altáblák

4. Kártya input: (1. konstans = 3 v 5 esetén)
I3 formátummal a dimenziószám (M)
5F16.8 formátummal M1=M+1 számú együtthatható. Az utolsó a konstans tag.

A program működése:

A program feltételezi, hogy a 4. és 6. output altábla ki van töltve, tehát előbb a KESZIT programot fel kell hívni!

Beállítja M-ben a dimenziószámot (M=20). A forráslistában x-gal meg vannak jelölve azok a sorok, amelyeken a dimenziószám növelésekor változtatni kell.

A TKONST (7) tömbben tárolja a konstansokat.

TKONST (1) = 1 esetén a 4. és 6 output altáblákból előállítja a 2. és 3. konstanssal megegyező tanítású osztályokat figyelembe véve előállítja a döntésfüggvény együtthatóit, oszlopokat generál a döntésfüggvény segítségével. A generált két oszlop a két osztály 1 dimenzióra transzformált pontjait tartalmazza. Ezek között az optimális elválasztó pont (az elválasztó függvény konstansa) megkereséséhez felhívja visszatéréssel a HISZT és a HISHAS programokat.

TKONST (1) = 2 esetén a felsorolt oszlopsorszámok tesztpontokat jelentenek. Ezért megnézi, hogy a 25. altáblában a döntésfüggvény együtthatói ott vannak-e. Ha nem, akkor a TKONST (1) = 1 esetnek megfelelően előállítja azokat és kiteszi a táblába. Ha igen, akkor előállítja minden egyes pontra a döntésfüggvény értékét, és összehasonlítja a döntést a tanítással. Mindezt ki is nyomtatja a pont sorszámával együtt, valamint a helyes döntést %-ban.

TKONST (1) = 3 esetén a TKONST (1) = 2-nek megfelelő tevékenységet kártyáról beolvasott együtthatók alapján végzi.

TKONST (1) = 4 esetén a felsorolt oszlopsorszámok ismeretlen kategóriájú pontokat jelentenek, tehát a hozzájuk tartozó vektorok tanítás koordinátát nem tartalmazhatnak. A 25. output altáblából vett döntésfüggvény segítségével kiszámolja annak értékét minden pontra, és azt a pont sorszámával együtt kinyomtatja. Ha a döntésfüggvény nincs az altáblában, akkor vége.

TKONST (1) = 5 eset megfelel a TKONST (1) = 4-nek, csak a döntésfüggvény kártyáról olvasandó be.

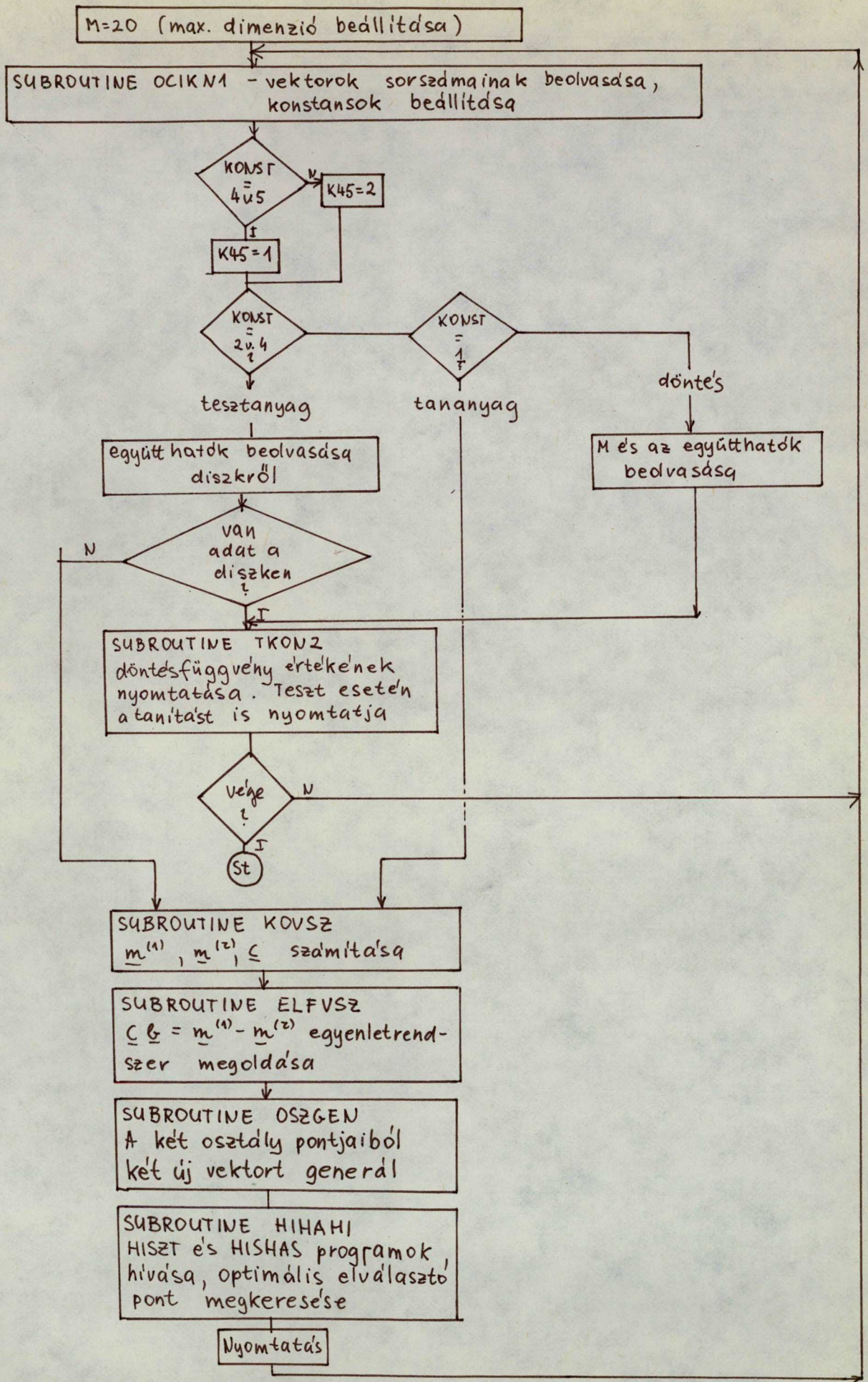
Program output:

1. Eredmények listázása sornyomtatón
2. Hibaüzenetek M:EO-hoz rendelt perifériára
3. Két új generált oszlop a generálás szabályainak megfelelő sorszámmal ellátva

4. A döntésfüggvény együtthatói a 25. output altáblában.

Hívott szubrutinok: STDISK
STEXIT
OCIKN1
TKON2
ELFVSZ
OSZGEN
HIHAHI

Hívott programok: HISZT
HISHAS



F.3. ábra.
A lineáris diszkriminancia analízis programjának blokkdiagramja

Legközelebbi szomszéd

Feladata: adott sorszámú vektorokon a legközelebbi szomszéd módszer végrehajtása; adott tananyag tesztelése illetve döntés ismeretlen kategóriájú pontokról. Euklideszi távolsággal dolgozik. (Fritz: Az alakfelismerés statisztikus módszerei. MTA 1974.)

Forrásnyelv: FORTRAN

Program input:

1. Oszlopsorszámok. Kétparaméteres program, a két oszlopsorszámmal az első és utolsó feldolgozandó vektort kell megadni. Ezért az egy menetben feldolgozandó vektoroknak egymás utáni sorszámúaknak kell lennie, üres vagy nem létező oszlop ne legyen közben!
2. Konstansok.
   1. konstans: a tanítás hanyadik koordináta (ez a program megengedi, hogy ez ne csak az első legyen)
   2. konstans: az értékes jellemzők honnan kezdődnek (azaz a vektor koordinátáiból melyeket akarunk megtartani)
   3. konstans: osztályok száma
   4. konstans: dimenziószám (tanítás nélkül)
   5. konstans: tevékenységszám.
      = 1: alaphalmazon teszt
      = 2: teszthalmazon teszt

= 3: új halmazon döntés

6. konstans: a maszkvektor sorszáma

Alapfeltételezés a konstansokra: 1,2,2. összelemszám-1, 1,Ø

Program output: az eredmények listázása a sornyomtatón. Alaphalmazon és teszthalmazon teszt esetén a döntési mátrixot és döntés esetén a sorszámot és a döntést nyomtatja ki.

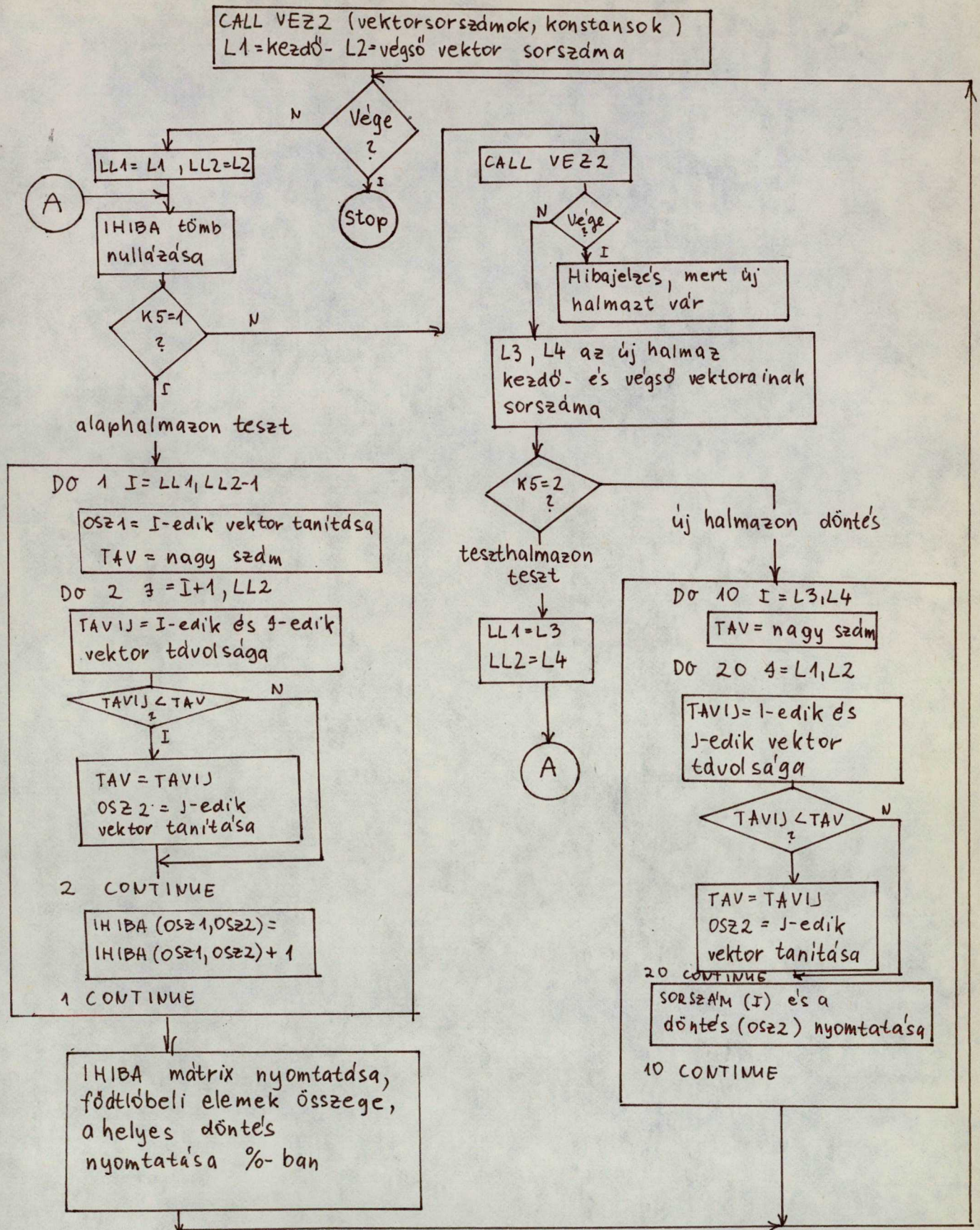
A program működése:

- Ha az 5. konstans = 2, akkor a következő két oszlopsorszám egy tesztanyagot határoz meg. Ekkor tehát a program még egy számpárt vár.
- Ha az 5. konstans = 3, akkor is még egy számpárt vár a program, amely most az ismeretlen osztályú pontok sorszámát adja meg.

A döntési halmaz pontjainak dimenziója 1-gyel kevesebb, mint az alaphalmazbelieké (a tanítás miatt).

Szubrutinok: STDISK

STEXIT



F.4. ábra.

A legközelebbi szomszéd módszer programjának blokkdiagramja

LEGKOZELEBBI SZOMSZED MODSZER

ALAPHALMAZON TESZT

HALMAZ KEZDOPONTJA: 1
HALMAZ VEGPONTJA: 60
OSZTALYHOVATARTOZAS 1 -DIK KOORDINATA
A VEKTORI A 2 -DIK POZICION KEZDODIK
KOORDINATAK SZAMA: 14
OSZTALYOK SZAMA: 2
TEVEKENYSEGSZAM: 1

DONTESI TABLAZAT

23 7

5 25

HELYES DONTES 48

ROSSZ DONTES 12

URES JELET TARTALMAZO OSZLOPOK SZAMA 0

HELYES DONTES SZAZALEKBAN 80.00
****

LEGKOZELEBBI SZOMSZED MODSZER

TESZTHALMAZON TESZT

HALMAZ KEZDOPONTJA: 71
HALMAZ VEGPONTJA: 147
OSZTALYHOVATARTOZAS 1 -DIK KOORDINATA
A VEKTOR A 2 -DIK POZICION KEZDODIK
KOORDINATAK SZAMA: 14
OSZTALYOK SZAMA: 2
TEVEKENYSEGSZAM: 2

DONTESI TABLAZAT

36 1

14 26

HELYES DONTES 62

ROSSZ DONTES 15

URES JELET TARTALMAZO OSZLOPOK SZAMA 0

HELYES DONTES SZAZALEKBAN 80.52
****

OSZLOP SORSZAM	SZAMOLASI ERTEK	EREDETI OSZTALY BESOROLAS	OSZT.BESOROLAS SZAMOLASI ERTEK SEGITSEGEVEL
132	-0.9177	1	2
133	0.8062	1	1
134	11.6554	1	1
135	18.8582	1	1
136	11.5583	1	1
137	10.0325	1	1
138	1.7141	1	1
139	3.6326	1	1
140	3.6800	1	1
141	1.1039	1	1
142	2.0810	1	1
143	12.0953	1	1
144	4.8539	1	1
145	0.4250	1	1
146	5.0985	1	1
147	6.0994	1	1

AZ OSZLOPOK SZAMA: 77

A HELYES DONTESEK ARANYA SZAZALEKBAN 92.00

OSZLOP SORSZAM	SZAMOLASI ERTEK	EREDETI OSZTALY BESOROLAS	OSZT.BESOROLAS SZAMOLASI ERTEK SEGITSEGEVEL
71	-6.8890	2	2
72	-2.6068	2	2
73	-3.6843	2	2
74	-4.1070	2	2
75	-3.2120	2	2
76	-2.8484	2	2
77	-0.3424	2	2
78	0.8489	2	1
79	-1.0044	2	2
80	-3.4659	2	2
81	-7.7606	2	2
82	-4.8957	2	2
83	9.3425	2	1
84	-3.3061	2	2
85	-2.7140	2	2
86	-3.0816	2	2
87	0.5559	2	1
88	1.2953	2	1
89	-8.6060	2	2
90	-3.5065	2	2
91	-5.1077	2	2
92	-5.9518	2	2
93	-7.4065	2	2
94	-1.6961	2	2
95	-5.3439	2	2
96	-5.2539	2	2
97	-5.4673	2	2
98	-0.4684	2	2
99	-2.0945	2	2
100	-3.9398	2	2
101	-1.0593	2	2
102	-0.7452	2	2
103	-3.0670	2	2
104	-1.3613	2	2
105	-0.7175	2	2
106	-6.3404	2	2
107	-0.0339	2	2
108	-3.7980	2	2
109	-4.4128	2	2
110	-5.5960	2	2
111	10.1269	1	1
112	-2.4799	1	2
113	3.4063	1	1
114	1.3028	1	1
115	5.3354	1	1
116	9.8439	1	1
117	10.3920	1	1
118	9.4467	1	1
119	3.4234	1	1
120	24.8470	1	1
121	3.3539	1	1
122	6.3438	1	1
123	3.2266	1	1
124	7.5311	1	1
125	3.7583	1	1
126	4.7561	1	1
127	4.7729	1	1
128	4.0669	1	1
129	1.8343	1	1
130	3.9598	1	1
131	0.7063	1	1

KET OSZTALY SZETVALASZTASA LINEARIS DISZKRIMINANCIA FUGGVENNYEL

KONSTANSOK: 2 1 2

ADOTT LINEARIS DISZKRIMINANCIA FUGGVENY TESZTELESE

AZ EGYUTTHATOK SZ MA: 14

A LINEARIS DISZKRIMINANCIA FUGGVENY EGYUTTHATOI:

0.0076	0.0000	-0.0083	0.0102	0.0517
-0.0004	0.0066	-0.0579	0.0004	0.0158
-0.0243	0.0228	0.0002	-0.0034	

A DISZKRIMINANCIA FGV. KONSTANSA: -8.2466

HA A DISZKRIMINANCIA FGV. EGY OSZLOPON VETT ERTEKE >=0.0 AKKOR 1 OSZTALYHOZ,
HA < 0.0 AKKOR 2 OSZTALYHOZ SOROLJUK

A VALTOZOK HATASA AZ ALTALANOSITOTT TAVOLSAG KIALAKITASARA:
OSSZES HATAS:

0.0572184570	0.0426581140	-0.2459811600	0.0500319000	0.2833957600
-0.2800223900	0.0185662440	-0.3804976300	0.8241764400	0.4389217500
-0.1926526500	0.1022253600	0.3052577900	-0.0232979120	

DIREKT HATAS:

0.0155396640	0.0046430790	1.2591398000	0.0194281460	0.2728180300
0.2912953500	0.0136141360	0.5528486300	1.3708151000	1.7873823000
0.2112469100	0.0375583880	0.2204253600	0.0058013946	

KET HISZTOGRAM OSSZEHASONLITASA. OSZLOPSORSZAMOK: 201 202

1.OSZTALY GYAKORISAGA:

0 3 0 2 3 2 3 0 8 2 7 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0

2.OSZTALY GYAKORISAGA:

0 0 0 0 0 0 0 0 0 1 3 1 1 3 0 5 3 5 1 2 2 2 0 0 0
0 1

AZ OPTIMALIS ELVALASZTO PONT 8.247
EBBEN AZ ESETBEN A HELYES DONTES VALOSZINUSEGE 93.33%

OSSZ ELEMSZAM 1.OSZTALY: 30 2.OSZTALY: 30 EGYUTT: 60

HIBAS DONTES 1.OSZTALY: 0 2.OSZTALY: 4 EGYUTT: 4

A LINEARIS DISZKRIMINANCIA FUGGVENY EGYUTTHATOI:

0.0076010638	0.0000362779	-0.0083386785	0.0101563990	0.0516631370
-0.0004235876	0.0065884221	-0.0579397260	0.0004056420	0.0158190490
-0.0242641220	0.0227745050	0.0002246369	-0.0034462345	

A DISZKRIMINANCIA FGV. KONSTANSA: -8.2466472000

HA A DISZKRIMINANCIA FGV. EGY OSZLOPON VETT ERTEKE >=0.0 AKKOR 1 OSZTALYHOZ,
HA < 0.0 AKKOR 2 OSZTALYHOZ SOROLJUK

A SZORASOK BECSLESEI:

43.6543890000	4999.6876000000	358.1965600000	36.5306370000	26.9114670000
3391.6032000000	47.1405920000	34.1592430000	7682.9493000000	224.9624100000
50.4210830000	22.6509270000	5563.2890000000	58.8305350000	

A MAHALANOBIS TAVOLSAG = 7.085373 FELBONTASA:

OSSZES HATAS:

0.4054141000	0.3022486600	-1.7428684000	0.3544947000	2.0079649000
-1.9840634000	0.1315487600	-2.6959676000	5.8395975000	3.1099242000
-1.3650161000	0.7243050900	2.1628654000	-0.1650744000	

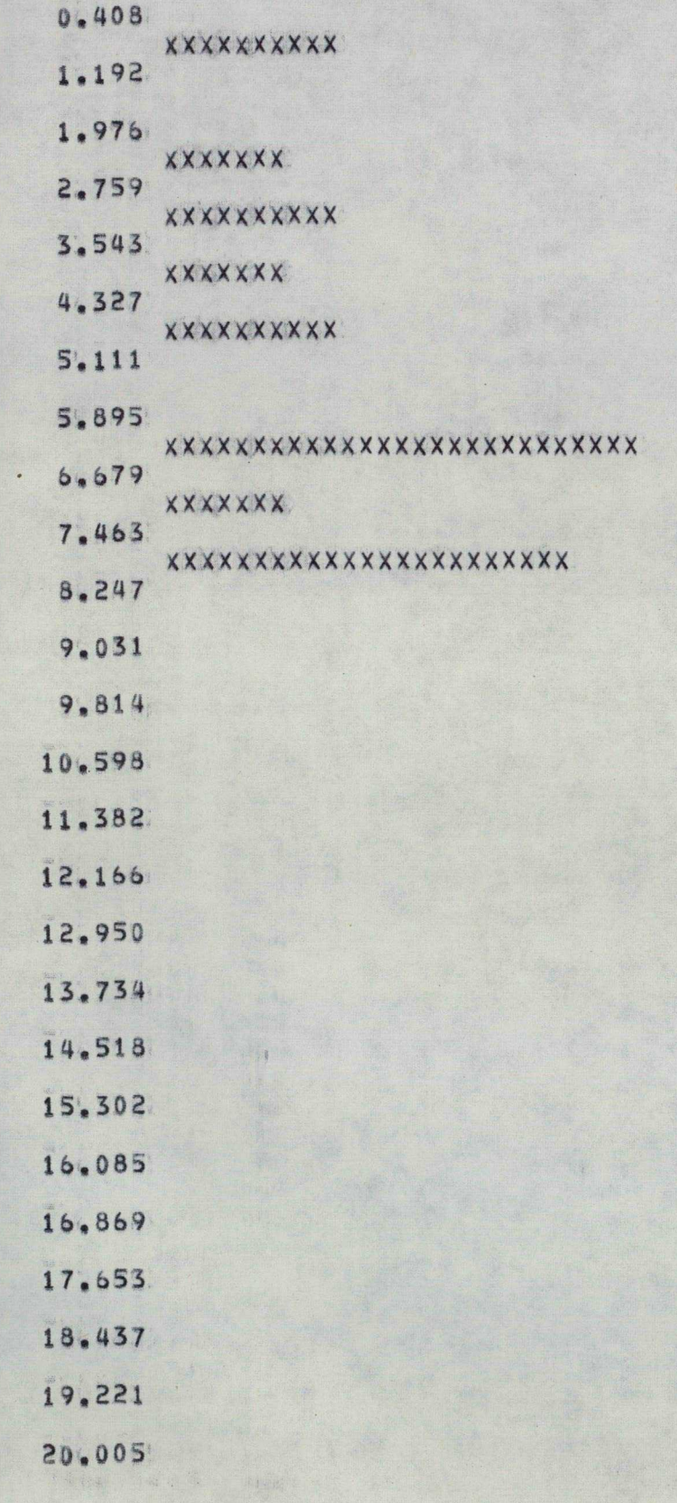
DIREKT HATAS:

0.1101043200	0.0328979560	8.9214820000	0.1376556700	1.9330180000
2.0639360000	0.0964612370	3.9171391000	9.7127378000	12.6642770000
1.4967633000	0.2661151900	1.5617961000	0.0411050550	

HISZTOGRAM KESZITES A 201-DIK OSZLOPROL

		GYAK	REL(%)
0.408		0	0.00
1.192		0	0.00
1.976		0	0.00
2.759		0	0.00
3.543		0	0.00
4.327		0	0.00
5.111		0	0.00
5.895		0	0.00
6.679		0	0.00
	XXX	1	3.33
7.463			
	XXXXXXXXXX	3	10.00
8.247			
	XXX	1	3.33
9.031			
	XXX	1	3.33
9.814			
	XXXXXXXXXX	3	10.00
10.598		0	0.00
11.382			
	XXXXXXXXXXXXXXXX	5	16.67
12.166			
	XXXXXXXXXX	3	10.00
12.950			
	XXXXXXXXXXXXXXXX	5	16.67
13.734			
	XXX	1	3.33
14.518			
	XXXXXXX	2	6.67
15.302			
	XXXXXXX	2	6.67
16.085			
	XXXXXXX	2	6.67
16.869		0	0.00
17.653		0	0.00
18.437		0	0.00
19.221		0	0.00
20.005		0	0.00
	XXX	1	3.33

HISZTOGRAM KESZITES A 202-DIK OSZLOPROL



GYAK	REL(%)
0	0.00
3	10.00
0	0.00
2	6.67
3	10.00
2	6.67
3	10.00
0	0.00
8	26.67
2	6.67
7	23.33
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00
0	0.00

1983. 05. 05.

PANCREATITIS

KET OSZTALY SZETVALASZTASA LINEARIS DISZKRIMINANCIA FUGGVENNYEL

KONSTANSOK: 1 1 2

F-PROBA

SZABADSAGI FOKOK: 14, 45 F-ERTEK: 5.8899

P-ERTEK: 0.000

A HIBAS STDISK TEV. SZAMA220 A HIBASZAM: 6

AZ EGYUTTHATOK SZAMA: 14

1983.05.05. PANCREATITIS

SULYPONT MODSZER

1-DIK ES 2-DIK OSZTALYT ELVALASZTO HIPERSIK EGYENLETE: -435294740.00000 +

53.337*X(1) 8331.508*X(2) 209.011*X(3) 34.904*X(4) 38.867*X(5) 4683.961*X(6)
19.967*X(7) 46.531*X(8) 14395.958*X(9) 196.594*X(10) 56.257*X(11) 31.803*X(12)
9628.301*X(13) 47.900*X(14)

1-DIK OSZTALY SULYPONTJA

121.92661	12920.53400	632.23314	65.23924	106.66656	9281.38670
123.83330	95.10556	21967.54800	465.31737	102.83256	79.65328
16967.44500	219.10000				

2-DIK OSZTALY SULYPONTJA

68.58992	4589.02720	423.22242	30.33559	67.79994	4597.42610
103.86664	48.57491	7571.58570	268.72337	46.57593	47.84992
7339.14440	171.19998				

A TESZTELES 60-DB EMBEREN TORTENT, HELYES DONTESEK SZAMA: 53 (88.00 %)

24 6
1 29

1-DIK OSZTALY SULYPONTJA

121.92661	12920.53400	632.23314	65.23924	106.66656	9281.38670
123.83330	95.10556	21967.54800	465.31737	102.83256	79.65328
16967.44500	219.10000				

2-DIK OSZTALY SULYPONTJA

68.58992	4589.02720	423.22242	30.33559	67.79994	4597.42610
103.86664	48.57491	7571.58570	268.72337	46.57593	47.84992
7339.14440	171.19998				

A TESZTELES 77-DB EMBEREN TORTENT, HELYES DONTESEK SZAMA: 66 (85.00 %)

35 2
9 31

301187. 70212066. 1386090. 314151. 228306. 49108690. 629673.
**********326805000. 76934701. 52250594.**********149949130.
7489525. 1680230. 1107235.231959280. 3053933.
401825. 244303. 53995597. 703644.
201072. 42737091. 518320.
**********115174420.
1505999.
30.
205989. 11546016. 1123694. 66711. 158975. 9377956. 211007. 147668. 19991970. 723075.
104422. 115058. 15637690. 344966.
902463730. 66377521. 5319998. 9507205.773731580. 16293091. 8368847.********** 43703538. 8074288.
6932063.********** 26329583.
10729766. 658919. 998178. 62262237. 1318033. 770021.109433060. 6558865. 1017084. 723794.
102015970. 2195877.
52582. 62376. 5222460. 106434. 46442. 8425757. 410341. 80934. 45585. 8079528.
176912.
169500. 10661393. 202518. 111572. 15691888. 625427. 95725. 120149. 17040110. 336068.
893492460. 15695500. 6660245.********** 39550608. 7791034. 7704633.********** 25693667.
411446. 147322. 26921326. 840827. 168188. 139221. 24789888. 636517.
108628. 14723849. 503070. 72998. 81420. 11117077. 243405.
********** 72751617. 13153708. 11359286.********** 43654782.
4107725. 642814. 457941. 65108049. 1404252.
127943. 69813. 12318820. 277325.
87714. 12339002. 232897.
********** 40697104.
1014168.

```
%DISGEN/BI

            LUNDH

%DISGEN/PT



        P R O G R A M N E V   T A B L A Z A T



    1    DISKRI: G(1,1) K3 P1
    2    KESZIT:  S  K2  PT
    3    SULYA:   S  K1  PT
    4    NN:  S  K6  P2
    5    %EOD

%EOD


%HIVGEN/
        1   ::830505PANCREATITIS,200;
        2   S KESZIT(1,FOR2,60C2,15)"
        3   S SULYA(1,FOR2,60C1; 71,FOR72,147)"
        4   G DISKRI(1; FOR1,60C1,1,2; 1; FOR71,147C2,1,2)"
        5   S NN(1,60C1,2,2,7,1,0; 1,60C1,2,2,7,2,0; 71,147)
        6   %EOD


      OSSZEGEK
        30.     3658.   387616.    18967.     1957.     3200.   278442.     3715.     2853.   659026.
     13960.     3085.     2390.   509023.     6573.
        30.     2058.   137671.    12697.      910.     2034.   137923.     3116.     1457.   227148.
      8062.     1397.     1435.   220174.     5136.


      NEGYZETEK
        30.   491662.********** 14077002.   180111.   351743.**********   501133.   301187.**********
   7489525.   401825.   201072.**********  1505999.
        30.   205989.902463730. 10729766.    52582.   169500.893492460.   411446.   108628.**********
   4107725.   127943.    87714.**********  1014168.


      SZORZATOK
        30.
    491662. 52106702.  2386564.   249188.   391891. 34722704.   448268.   379928. 88525903.  1732969.
    390252.   292921. 62775183.   804263.
**********254866640. 30812192. 41687965.********** 51041502. 41251522.**********191201210. 46112913.
 31105775.********** 88211328.
 14077002.  1438132.  2047926.172666900.  2308165.  1878281.443378440. 10125339.  2236372.  1505509.
31125325@.  4108467.
   180111.   212205. 19738609.   260170.   201624. 49268705.  1086990.   264180.   155017. 34250879.
   444622.
   351743. 30788338.   397551.   303653. 70240992.  1503250.   333565.   262300. 55837214.   702973.
********** 36398816. 27107632.**********128881630. 30243742. 23203534.********** 63030177.
   501133.   353695. 83463871.  1726349.   397989.   291934. 64302850.   847985.
```

```
%JOB
 / 0 / 0 / 0 /41

%ME 184-ES MSZ-OT AZ A-RA

%C/MTAS/BI/BN/A
MTAS  0001
  6CB6 6CBA 6E1A
 A=D:3

%AS/M:EO,T:TY,AN

%AS/M:EI,T:CR,AN

%ME IT,KV,AB

%ME
```