

Szegedi Orvostudományi Egyetem Számítástechnikai Központ
és Szegedi Orvostudományi Egyetem I.sz. Belgyógyászati
Klinika

Alakfelismerési módszerek alkalmazása hasnyálmirigy gyul-
ladás diagnosztizálására a STASYS rendszer segítségével

Boda Krisztina, Papp Ákos és Grácin Edit

Bevezetés

A SZOTE-n az orvosi kísérletek, felmérések adatainak biometriai, statisztikai feldolgozása 1976. óta a STASYS programrendszerrel történik R-10-es számítógépen. 1979-80-ban e rendszerbe a legismertebb alakfelismerési eljárásokat programoztuk be.

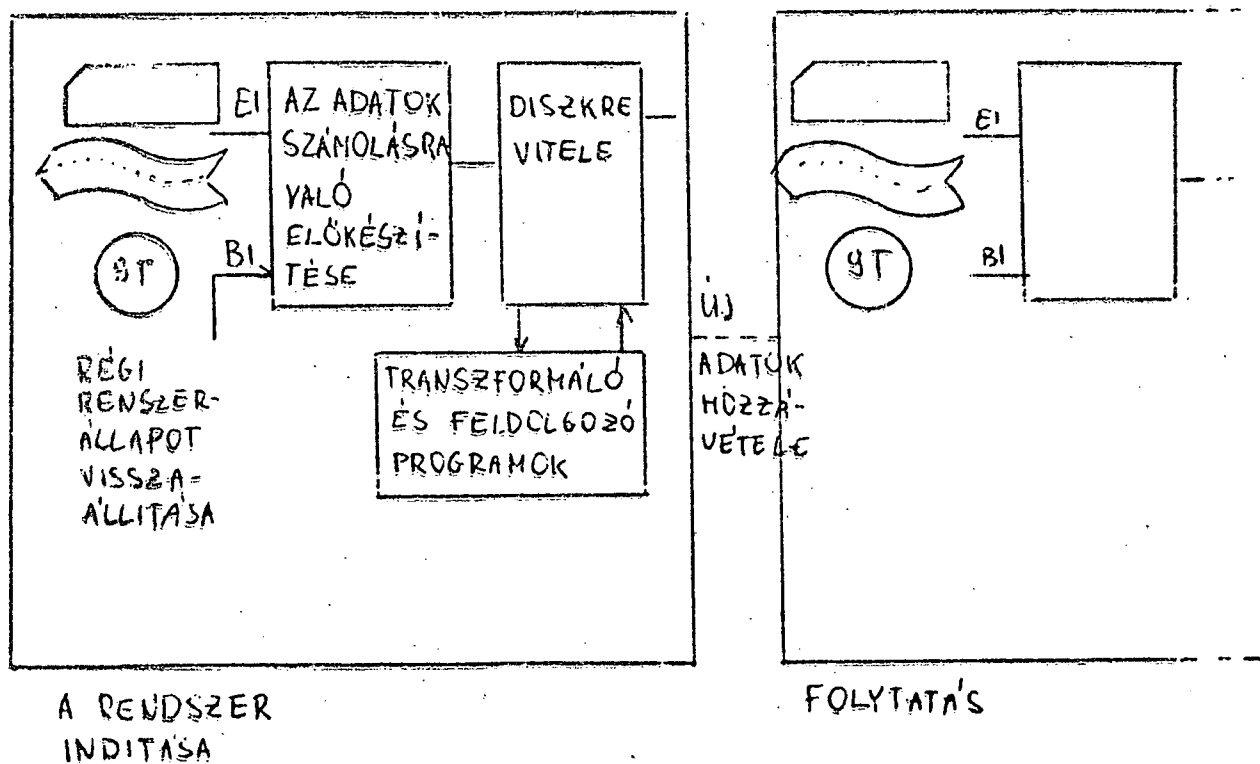
Munkánkban ismertetjük a rendszerbeillesztés problémáit, továbbá a programok diagnosztikai célú alkalmazását hasnyálmirigy gyulladásban szenvedő és kontroll esetek szétválasztására.

A STASYS rendszer főbb jellemzői

STASYS statisztikai programrendszert 1976-ban az orvosi kísérletek kiértékelésére, rutin statisztikai futtatások céljából hoztuk létre. Mivel a rendszert már több helyen (1), (2), (3) ismertettük, most csak a legfontosabb tulajdonságait sorolom fel.

A rendszer diszkorientált, azaz a feldolgozandó adatokat, a számításokhoz szükséges táblázatokat és részeredményeket jól adminisztrált formában a diszken tárolja. Alapegysége a változó hosszúságú vektor, amelyre a feldolgo-

zás előtt adott sorszámmal lehet hivatkozni. A STASYS maga egy futtató rendszer, erre épülnek az adatokat transzformáló programok és a statisztikai eljárások programjai. A futtató rendszer biztosítja az adatok egységes kezelését /azok mennyiségétől és formájától függetlenül/ és a programok automatikus indítását. Az adatok diszkrevelését, egy rendszer-állapot mentését és visszaállítását egy rendszerprogram végzi /1. ábra/.



1. ábra

A rendszerben lehetőség van futó programból rendszerbeli program hívására, így fastrukturájú feladatok megoldására. Ezzel tudjuk biztosítani /- viszonylag kis memória kapacitás mellett -/ a statisztikai módszerek egymásra épülését, azt, hogy egyes eljárások önmagukban és egymással összefüggésben is élni tudjanak. A programok sorrendjét és pa-

ramétereinek megadását egy könnyen kezelhető *célnyelv* teszi egyszerűvé.

Alakfelismerő eljárások rendszerbe illesztése

Az utóbbi 4 év során a rendszer alkalmasnak bizonyult a Számítástechnikai Központba érkező sokrétű és mennyiségű statisztikai feldolgozás elvégzésére. Viszont egyre gyakrabban merültek fel olyan problémák, amelyekre pusztán a szokványos biometriai módszerekkel nem lehetett megnyugtató megoldást adni, és amelyek döntési - osztályozási problémákkal voltak kapcsolatban.

Ismert, hogy az alakfelismerési módszerek egyik fontos alkalmazási területe az orvosi diagnosztika. A diagnosztikai kérdések felmerülésével szükségessé vált az alakfelismerési módszerek programjainak elkészítése. Ezeket a programokat is a rendszerbe illesztettük, ugyanis:

- a rendszerben az adatkezelés, filekezelés jól adminisztrált, megoldott;
- a programból programhívás alakfelismerési eljárásoknál is jól felhasználható;
- az alapstatisztikákat osztályozási problémák esetén is célszerű kiszámítani;
- az alap-probléma általában statisztikai természetűként merül fel.

Az alakfelismerési programokban egy vektort egy egyed, azaz tanulásnál egy tanulópont koordinátáiként fogtuk fel /2. ábra/.

További adatkezelő és adatelőkészítő programok beiktatására is szükség volt, amelyek bizonyos mennyiségeket /pl. kovariancia mátrixokat/ számolnak ki és azt a lehetőséget biztosítják, hogy tetszőleges koordinátákat tartsunk meg az anyagból.

	1. koord.	2. koord.	...,n koord.	
1. egyed	X	X	...,X	1. vektor
2. egyed	X	X	...,X	2. vektor
	.	.	.	
	.	.	.	
	.	.	.	

2. ábra

Az un. tanítóval való tanulás egyes módszereit megvalósító programoknak a következő funkcióik vannak:

1. Döntésfüggvény készítése tanulópontok alapján;
2. Döntésfüggvény tesztelése teszthalmaz alapján;
3. Döntés kategória nélküli pontokról.

Az eddig beprogramozott módszerek a következők:

NN - Legközelebbi szomszéd
Sulypont módszer
Bayes döntésfüggvény
Lineáris diszkriminancia analízis
Committe Machine.

További terveinkben olyan algoritmusok programozása szerepel, amelyek viszonylag kis memória igényűek - általában iteratív vagy adaptív módszerek.

Éppen a kis memóriakapacitás miatt korlátozott a cluster algoritmusok beépítésének lehetősége. Jelenleg két nemhierarchikus cluster algoritmus és egy bináris változókkal dolgozó algoritmus assembly nyelvű programja él a rendszerben.

A módszerek gyakorlati alkalmazása hasnyálmirigy-betegség diagnosztizálására

Az alakfelismerési programokat hasnyálmirigy betegek adatain futtattuk le. A hasnyálmirigy gyulladás /pancreatitis/ diagnosztizálása sok laboratóriumi vizsgálatot igénylő feladat. Ebben a pancreas /hasnyálmirigy/ enzimek vizsgálata az egyik legfontosabb tényező. A többféle sejttípus és a pancreas-enzimek nonparallel szekréciója miatt egyes funkciók korábban károsodhatnak, miközben mások még normális értéket mutatnak. Ezért egyetlen pancreas-funkciós próba alapján a hasnyálmirigy működéséről biztos véleményt nem lehet alkotni, több tényező együttes értékelésére kell törekedni. Kézenfekvőnek látszik tehát az alakfelismerési módszerek alkalmazása.

Beteganyag

A kontroll csoport betegei egyéb, a hasnyálmiriggyel össze nem függő betegséggel ápoltt klinikai esetek voltak. A pancreatitises csoportban a betegség diagnózisát több megelőző orvosi vizsgálat biztosította. A vizsgálat első szakaszában a normálérték meghatározásához 77 biztosan nem pancreatitises kontroll eset adata állt rendelkezésre, a további statisztikai módszerekkel 70 beteg és újabb 67 kontroll eset vizsgálatát hajtottuk végre.

Módszer

A statisztikai analízishez 14 laboratóriumi paraméterből indultunk ki, melyeket a duodenum szondázással nyert anyagban a három fő pancreas enzim: a tripszin, lipáz, amiláz 3 x 10 perces frakciókban mért aktivitásából, továbbá a duodenum nedv térfogatából számoltuk a frakciók maximumát, átlagát és összegét véve alapul /1. táblázat/.

		Átlag	±SEM	P	Nörmál 2.5 %perc	érték 1 képl.
AMILIÁZ	Átlagos aktivitás	IU/ml 99.81	±5.242	0.2272	39.4	38.9
	Maximális aktivitás	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 aktivitás	mIU/ml 521.82	±23.043	0.7131	231.6	219.1
	Maximális aktivitás	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 aktivitás	IU/ml 81.31	± 2.87	0.0169	41.5	41.6
	Maximális aktivitás	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

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

Normálértékek meghatározása

A pancreas funkciós próbák eredményeiből az orvosnak el kell döntenie, hogy a lelet kórosnak vagy normálisnak tekinthető-e. Ezek az értékek mindig pozitív számok, és a kisebb értékek a kórosak.

Normálértéknek nevezzük azt a számot, amelynél kisebb érték kórosnak tekinthető, nagyobb vagy egyenlő pedig normálisnak. A normálértékeket általában valamely egészséges populációból származó minta alapján becsülik. Az irodalom szerint (5) szokásos gyakorlat az, hogy a mintának logaritmikus transzformáció utáni átlagából a kétszeres standard deviációt levonják:

$$N = \exp(\bar{y} - 2S_d) \quad /1/$$

ahol x_i $i = 1, \dots, n$ a mintaelemek, $y_i = \log x_i$, $\bar{y} = \frac{\sum y_i}{n}$

$$S_d = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}$$

Ez az eljárás lognormális eloszlás esetén indokolt.

Esetünkben 77 korábban vizsgált kontroll eset adata állt rendelkezésre a normálértéknek meghatározásához. A 14 paraméter eloszlását χ^2 próbával vizsgáltuk. Egyes esetekben normális, máskor lognormális, vagy még ferdebb volt az eloszlás, ezért a fenti eljárást nem tudtuk egységesen alkalmazni. Ezért minden paraméterre kiszámoltuk a 2.5 %-os percentilis értéket is, és normálérték helyett "normál bizonytalansági intervallumot" használtunk, amelynek két végpontja az /1/ képlet szerint számolt érték és a 2.5 %-os percentilis volt

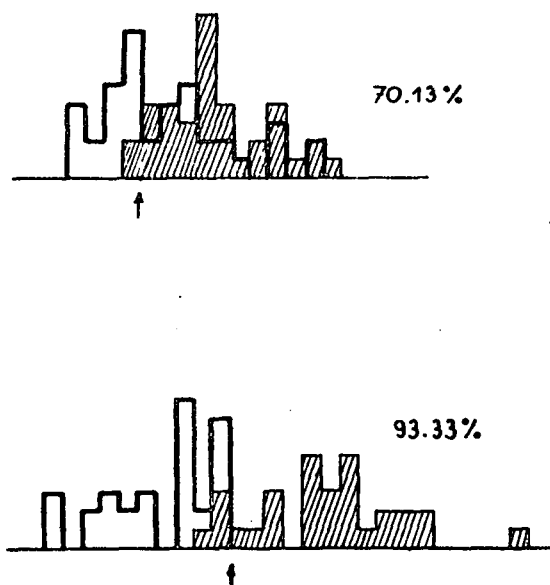
$$I: [\min(2.5 \% \text{ perc}, N), \max(2.5 \% \text{ perc.}, N)]$$

Az eloszlásvizsgálatból származó valószínűségeket és a normálértékeket az 1.sz. táblázat tartalmazza.

A normál és kóros beteganyag összehasonlítása

Korábbi hasonló témájú munkánkban a paraméterek átlagait Student-féle t-próbával, illetve Wilcoxon próbával hasonlítottuk össze (4). A különbség - bár esetenként szignifikáns volt, mégsem bizonyult eléggé meggyőzőnek.

Jelen vizsgálatunkban hisztogramokat készítettünk a kóros és normál esetek eloszlásáról annak elbirálására, hogy melyek azok a vizsgálatok, amelyeknél az eloszlások a legkevésbé fedik egymást. A hisztogramok összehasonlításakor általában 70-75 %-os elválasztás adódott /3a. ábra/.



3. a-b. ábra

Több paraméter együttes figyelembevételének igénye az alakfelismerési módszerekhez vezetett bennünket. Adatainkon a lineáris diszkriminancia analizist, a súlypont módszert és a legközelebbi szomszéd módszert futtattuk le. A tananyag 30-30, a tesztanyag 40-37 beteg, illetve kontroll

eset adataiból állt. A 3b. ábra a lineáris diszkriminancia függvénnel egy dimenzióba transzformált pontok eloszlását mutatja. Látható, hogy az átfedés sokkal kisebb, mint a 3a. ábrán. A három módszer eredményeit, mind a 14 paramétert egyszerre figyelembe véve a 2. táblázat foglalja össze. Az egész számok a táblázatban a helyes döntések számát jelentik, a legalsó sorban az összesetszám van feltüntetve. Kiszámítottuk a helyes döntések arányát az összesetszám %-ában is. A 14 koordinátát orvosi megfontolások alapján csökkentve, valamennyi esetben romlott a döntés biztonsága.

Megbeszélés

A három módszer közül a diszkriminancia analízis adta a legjobb eredményt.

A *súlypont módszer* - mivel az osztályonkénti szórást nem veszi figyelembe, kisebb hatásfoku; egyszerűsége miatt végeztük el, tájékoztató jelleggel.

A *legközelebbi szomszéd módszerrel* kapott kedvezőtlenebb eredmények véleményünk szerint azzal magyarázhatók, hogy programunk jelenleg még csak euklideszi távolsággal tud dolgozni, és elképzelhető, hogy más távolsággal jobb eredményt adna.

A *lineáris diszkriminancia analízis* alkalmazhatóságának feltétele az adatok normális eloszlása és a kovariancia mátrixok osztályonkénti megegyezése, ekkor a legkisebb hibájú döntést adja. Tudomásunk szerint Sarles (5) és munkatársai foglalkoztak a Fisher féle diszkriminancia analízis e témakörben való alkalmazhatóságával, azonban adataik ferde eloszlása miatt nem tartották adekvát módszernek. Kétségtelenül a mi vizsgálatainkban is zavaró volt a ferde eloszlás, mégis kísérletet tettünk a lineáris diszkriminancia analízis segítségével való diagnosztizálásra. Ugyanis

H e l y e s d ö n t é s

M ó d s z e r	Tananyag				Tesztanyag			
	Kontroll	Pancr.	Össz.	%	Kontroll	Pancr.	Össz.	%
Diszkriminancia Analizis	26	30	56	93.33	35	36	71	92.2
Legközelebbi szomszéd	23	25	58	80.00	36	26	62	80.52
Súlypont	24	29	53	88.33	35	31	66	85.71
<u>Összes esetszám:</u>	30	30	60		37	40	77	

2. táblázat. Az osztályozás eredménye három módszerrel 14 paraméter figyelembevételével.

1. van a módszernek egy eloszlásmentes megfogalmazása, amely szerint Fisher féle diszkriminancia függvény egy optimalitási krétériumnak eleget tevő transzformáció (6). /A hibájáról ekkor semmit sem tudunk mondani./

2. Találtunk olyan tanulmányt (7) e témakörben, amely a lineáris diszkriminancia sikeres alkalmazásáról számol be, és az eloszlás kérdésével nem foglalkozik.

3. A tesztalacson is igen jó a helyes döntések aránya.

Külön előnye a lineáris diszkriminancia analízis gyakorlatban való alkalmazásának, hogy a tünetek ismeretében kis kézi számológépen kiszámítható az értéke.

A lineáris diszkriminancia analízissel nyert eredmények lehetőséget adnak arra, hogy az orvosi diagnózist - főleg vitatott esetekben - a döntésfüggvény értékének kiszámításával alátámasszuk.

Irodalomjegyzék

- (1) Szerényi L., Kasza F., Stéhlikné Boda K., Győri I., Forczek E.: Adatfeldolgozás céljait szolgáló programrendszer R-10-es számítógépre, különös tekintettel a biometriai feldolgozásokra. NJSZT 7. Kollokvium, Szeged, 1976. 123-130.
- (2) Győri I., Kasza F., Szerényi L.: Vektorstrukturájú adatok feldolgozó programjainak futtató rendszere. VIDEOTON Software Tájékoztató, 1977. 1-2. 29-31.
- (3) Boda K., Forczek E., Győri I.: R-10-es számítógépre készült statisztikai programrendszer. VIDEOTON Software Tájékoztató, 1980/1. 33-36.

- (4) Pap Á., Berger Z., Varró V.: Összehasonlító vizsgálatok secretin-pancreozymin és Lundh próbával. Orvosi Hetilap, 1980.
- (5) Sarles, H., Salem, E., Cros, R., Figarella, C.: Exploration pancréatique par tubage. Acta gastro-ent. belg. 33, 303-318., 1970.
- (6) Kendall, M., Stuart, A.: The Advanced Theory of Statistics Vol. 3. Charles Griffin and Company Ltd. London, 1976.
- (7) Aparisi, L., Rodrigo, J., Montes, F., Bel, J., Guie, J., Serra, M., Bixquerol, M., Wassel, A.: Application of linear discriminatory analysis to determine the diagnostic value of fecal chymotrypsin. In European Pancreatic Club 12 th Meeting, Copenhagen, Denmark, 11-13., oct. 1979.