Anna Bartkowiak (Wrocław)
Jerzy Liebhart (Wrocław)
Ewa Liebhart (Wrocław)
Józef Małolepszy (Wrocław)

OCENA TRZECH ALGORYTMÓW DYSKRYMINACYJNYCH DLA CECH ILOŚCIOWYCH NA PRZYKŁADZIE NIEKTÓRYCH SCHORZEŃ UKŁADU ODDECHOWEGO

1. Materiał i metoda

Pomimo istotnych sukcesów w zwalczaniu gruźlicy przewlekłe schorzenia układu oddechowego są coraz poważniejszym problemem społecznym. Postępujące zanieczyszczenie atmosfery pyłami i toksycznymi gazami prowadzi do wzrostu zachorowań na tzw. przewlekłą zaporową chorobę płuc obejmującą: przewlekły nieżyt oskrzeli, dychawicę oskrzelową i rozedmę płuc. Bardzo dobrze ilustrują to zagadnienie dane opublikowane przez Sawickiego i wsp. [7]. Autorzy ci w 1968 r. stwierdzili objawy przewlekłego nieżytu oskrzeli lub dychawicy oskrzelowej aż u 23,2 % zbadanych w tym czasie w Krakowie mężczyzn. W 1973 r. częstość występowania tych schorzeń była jeszcze wyższa i wynosiła 28,6 %. Według Leowskiego [3] w latach 1967-68 choroby układu oddechowego stanowiły 19,6 % ogólnej zachorowalności w Polsce, przy czym częstość ich występowania była około dwukrotnie większa od np. schorzeń układu krążenia (10,5 %) lub układu pokarmowego (9,8 %). Do równie alarmujących wniosków musi prowadzić analiza wskaźników absencji chorobowej, inwalidztwa i umieralności. W świetle przytoczonych danych staje się oczywiste, że problem profilaktyki i zwalczania nieswoistych chorób płuc stawia przed służbą zdrowia nowe, pilne zadania teoretyczne i praktyczne.

Szczególnie ważne jest zagadnienie wczesnego wykrywania tych schorzeń.

W Klinice Chorób Wewnętrznych Akademii Medycznej we Wrocławiu oraz w Instytucie Informatyki Uniwersytetu Wrocławskiego podjęto próbę stworzenia podstaw informatycznego systemu dla zautomatyzowanych poradni pneumologicznych, początkowo o profilu głównie astmologicznym.

W pierwszym etapie pracy objęto badaniami pacjentów hospitalizowanych w Klinice Chorób Wewnętrznych oraz chorych leczonych ambulatoryjnie - razem 303 osoby. Cały materiał podzielono na 5 następujących grup:

Grupa 1 - dychawica oskrzelowa niepowikłana; $n = 103$

Grupa 2 - dychawica oskrzelowa powikłana rozedmą płuc, a w części przypadków także przewlekłym sercem płucnym; $n = 48$

Grupa 3 - przewlekły nieżyt oskrzeli niepowikłany; $n = 23$

Grupa 4 - przewlekły nieżyt oskrzeli powikłany; $n = 24$

Grupa 5 - kontrolna, obejmująca osoby, u których nie stwierdzono objawów wymienionych wyżej schorzeń; $n = 105$

Dla każdego badanego wypełniana była ankieta zawierająca 146 pozycji, w tym: dane personalne, wywiad chorobowy, stan przedmiotowy, badania dodatkowe (między innymi spirometrię, gazometrię, ekg, morfologię krwi, analizę moczu, wartości RR i tętna, wagę i wzrost) oraz rozpoznanie kliniczne.

Celem pracy było zasymulowanie procesu zbierania i składowania informacji, a także automatycznego wspomagania lekarza w jego czynnościach diagnostycznych.

W pierwszej fazie do dalszego opracowania wzięto wyniki tych badań dodatkowych, które miały charakter cech ilościowych ciągłych. Niektóre z nich przetworzono w inne wskaźniki. W rezultacie do opracowania metodami statystyki matematycznej i zbadania własności różnicujących między wyżej wymienionymi jednostkami chorobowymi wzięto następujące zmienne: 1. tętno, 2. RR skurczowe, 3. RR rozkurczowe, 4. OB po 1 godzinie, 5. OB po 2 godzinach, 6. hemoglobina, 7. liczba erytrocytów, 8. liczba leukocytów, 9. $\frac{FEV_1 \text{ akt./ml/}}{FEV_2 \text{ nal./ml/}} * 100$, 10. VC akt. /ml/, 11. VC akt. /%/, 12. FEV_1 akt. /ml/, 13. FEV_1 akt. /%/, czyli wskaźnik Tiffaneau, 14. FEV_1 po salbutamolu lub histaminie, 15. $\frac{\Delta FEV_1 \text{ po salbutamolu lub histaminie}}{FEV_1 \text{ należne}} * 100$, 16. pH, 17. pO_2,

18. pCO_2, 19. SaO_2, 20. HCO_3 st., 21. HCO_3 akt., 22. BE, 23. TCO_2, 24. $\frac{FEV_1 \text{ akt. /ml/}}{VC \text{ należne /ml/}} * 100$, 25. eozynofilia we krwi, 26. ΔFEV_1 po salbutamolu lub histaminie.

Do tak zebranych danych zastosowano następujące metody analizy dyskryminacyjnej:

1. Wybór cech o największej sile dyskryminacji, szeregowanie cech według ich siły dyskryminacyjnej (program *dis5*)
2. Diagnoza, czyli identyfikacja metodą kanonicznych funkcji dyskryminacyjnych (program *dis4*)
3. Diagnoza, czyli identyfikacja metodą liniowych i kwadratowych funkcji dyskryminacyjnych (program *dis6*)

Omówimy teraz bardziej szczegółowo poszczególne metody oraz pokażemy ich działanie na zebranym materiale. Obliczenia były wykonane za pomocą odpowiednich programów napisanych w języku ALGOL 1204. Autorzy dysponują również programami napisanymi w języku ALGOL 1900 z implementacją na m.c. ODRA 1305.

2. Wybór zmiennych o największej sile dyskryminacji

Spośród różnych wskaźników mierzących siłę dyskryminacji można przyjąć wskaźnik używany w wielozmiennej analizie danych, mianowicie tzw. statystykę Λ Wilksa określoną następująco [6]:

$$\Lambda = \frac{|W|}{|T|}, \tag{1}$$

gdzie symbole $|W|$, $|T|$ oznaczają wyznaczniki z macierzy zmienności wewnątrzgrupowych i całkowitych, czyli wyrażenia proporcjonalne do odpowiednich uogólnionych wariancji. Wskaźnik ten jest naturalnym uogólnieniem jednozmiennego kryterium, które zakłada, że siła dyskryminacji danej zmiennej jest tym większa, im obliczony dla tej zmiennej stosunek zmienności wewnątrzgrupowej do całkowitej jest mniejszy.

Niech ciąg $x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_1}^{(1)}, x_1^{(2)}, \ldots, x_{n_1}^{(2)}, \ldots, x_1^{(G)}, \ldots, x_{n_G}^{(G)}$ oznacza wartości zmiennej x zaobserwowane u $n = n_1+n_2+\ldots+n_G$ osobników należących do G różnych populacji. Zmienności wewnątrzgrupowe Z_w i całkowite Z_t obliczamy ze wzorów:

(2) $$Z_w = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_i^{(g)} - \bar{x}_\cdot^{(g)})^2,$$

(3) $$Z_t = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_i^{(g)} - \bar{\bar{x}}_\cdot)^2,$$

gdzie

$$\bar{x}_\cdot^{(g)} = \left(\sum_{i=1}^{n_g} x_i^{(g)} \right) / n_g, \quad g = 1, 2, \ldots, G,$$

$$\bar{\bar{x}}_\cdot = \left(\sum_{g=1}^{G} n_g \bar{x}_\cdot^{(g)} \right) / (n_1 + n_2 + \ldots + n_k).$$

Wprowadzając dodatkowe określenie zmienności międzygrupowej Z_b:

(4) $$Z_b = \sum_{g=1}^{G} n_g (\bar{x}_\cdot^{(g)} - \bar{\bar{x}}_\cdot)^2$$

można łatwo pokazać, że zmienność całkowita Z_t równa się sumie zmienności wewnątrzgrupowej Z_w i zmienności międzygrupowej Z_b:

(5) $$Z_t = Z_w + Z_b.$$

Z równości powyższej wynika natychmiast, że

(6) $$0 \leq \frac{Z_w}{Z_t} \leq 1.$$

Analogicznie niech ciąg

(7) $$x_1^{(1)}, x_2^{(1)}, \ldots, x_{n_G}^{(G)}$$

oznacza ciąg obserwacji wielozmiennych pochodzących z G różnych populacji, tzn.

$$x_i^{(g)} = (x_{i1}^{(g)}, x_{i2}^{(g)}, \ldots, x_{ip}^{(g)}),$$

$$i = 1, 2, \ldots, n_g, \quad g = 1, 2, \ldots, G,$$

gdzie p oznacza liczbę zmiennych obserwowanych u każdego osobnika.

Dla danych obserwacji obliczamy wektory średnich grupowych $\bar{x}_\cdot^{(g)}$, $g = 1, 2, \ldots, G$, oraz wektor średniej generalnej $\bar{\bar{x}}_\cdot$, a następnie określamy

macierze zmienności wewnątrzgrupowych W, międzygrupowych B i całkowitych T:

$$W = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_i^{(g)} - \bar{x}^{(g)})(x_i^{(g)} - \bar{x}^{(g)})', \tag{8}$$

$$B = \sum_{g=1}^{G} n_g (\bar{x}^{(g)} - \bar{\bar{x}})(\bar{x}^{(g)} - \bar{\bar{x}})', \tag{9}$$

$$T = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_i^{(g)} - \bar{\bar{x}})(x_i^{(g)} - \bar{\bar{x}})'. \tag{10}$$

Między macierzami T, B, W zachodzi następujący związek:

$$T = W + B. \tag{11}$$

Ze wzorów (8) - (10) wynika, że macierze W, B, T są macierzami Gramma, a więc są to macierze określone nieujemnie.

Korzystając z faktu, że mamy do czynienia z macierzami Gramma, możemy obliczać wartości wyznaczników z macierzy W i T według wzorów cytowanych m. in. przez McCabe'a [5]. Przypuśćmy, że obliczamy wyznacznik z macierzy $W = \{w_{ij}\}$, $i,j = 1, 2, \ldots, p$. Wyznacznik ten możemy obliczać według wzoru:

$$|W| = w_{11}\, w_{22.1} w_{33.12} \cdots w_{pp.12\ldots(p-1)}, \tag{12}$$

gdzie symbol $w_{ii.12\ldots(i-1)}$, $i = 2, 3, \ldots, p$, oznacza zmienność resztową pozostałą ze zmienności początkowej w_{ii} po odjęciu od zmiennej nr i jej najlepszej (w sensie metody najmniejszych kwadratów) liniowej oceny za pomocą zmiennych o numerach $1, 2, \ldots, i-1$.

Niech symbol $\Lambda_{(p)}$ oznacza wskaźnik Λ zdefiniowany wzorem (1), obliczony dla wektora obserwacji o p składowych odpowiadających p zmiennym, czyli p cechom. Wykażemy teraz, że

$$\Lambda_{(p-1)} \leq \Lambda_{(p)}. \tag{13}$$

Korzystając ze wzoru (13) możemy wskaźnik $\Lambda_{(p)}$ napisać w postaci

$$\Lambda_{(p)} = \frac{w_{11} w_{22.1} \cdots w_{pp.12\ldots(p-1)}}{t_{11} t_{22.1} \cdots t_{pp.12\ldots(p-1)}} = \Lambda_{(p-1)} \frac{w_{pp.12\ldots(p-1)}}{t_{pp.12\ldots(p-1)}}.$$

Wartości $w_{pp.12\ldots(p-1)}$ i $t_{pp.12\ldots(p-1)}$ mogą być interpretowane jako

zmienności wewnątrzgrupowe i całkowite nowej zmiennej, a mianowicie reszty z_i określonej jako różnica między zmienną x_i a jej liniową oceną za pomocą zmiennych $x_1, x_2, \ldots, x_{i-1}$:

$$z_i = x_i - b_0 - b_1 x_1 - \ldots - b_{i-1} x_{i-1},$$

gdzie współczynniki $b_0, b_1, \ldots, b_{i-1}$ są współczynnikami spełniającymi odpowiednie równania normalne wynikające z metody najmniejszych kwadratów. Wobec tej interpretacji jest spełniona nierówność

$$0 \leq \frac{w_{pp.12\ldots(p-1)}}{t_{pp.12\ldots(p-1)}} \leq 1. \tag{14}$$

Powracając do równości (15) otrzymujemy

$$\Lambda_{(p)} = \Lambda_{(p-1)} \frac{w_{pp.12\ldots(p-1)}}{t_{pp.12\ldots(p-1)}} \leq \Lambda_{(p-1)},$$

co kończy dowód prawdziwości wzoru (14).

Analogicznie łatwo wykazać, że jest prawdziwa następująca nierówność:

$$0 \leq \Lambda_{(p)} \leq 1 \quad \forall\, p > 0. \tag{15}$$

Własność statystyki Λ określona wzorem (15) oznacza unormowanie wskaźnika siły dyskryminacyjnej w ten sposób, że jego granicznymi wartościami są wartości 0 i 1. Im mniejsza wartość wskaźnika $\Lambda_{(p)}$ tym większy udział zmienności międzygrupowej w zmienności całkowitej, czyli tym większa siła dyskryminacyjna rozważanego zespołu p zmiennych. Wyznaczniki z macierzy W i T noszą nazwę uogólnionych zmienności wewnątrzgrupowych i całkowitych.

Jeśli $\Lambda_{(p)} = 0$, to wynika stąd, że uogólniona zmienność wewnątrzgrupowa jest równa zeru, a więc uogólniona zmienność międzygrupowa jest równa uogólnionej zmienności całkowitej. W tym przypadku mamy do czynienia z kompletną, tj. stuprocentową dyskryminacją.

Należy jednak zauważyć, że jeśli dyskryminacja jest stuprocentowa, to skaźnik $\Lambda_{(p)}$ określony powyżej niekoniecznie musi być równy zeru.

Własność określona wzorem (13) oznacza, że dodanie nowych zmiennych nie może „popsuć" siły dyskryminacyjnej zmiennych uwzględnionych wcześniej.

W oparciu o wzór (13) napisano procedurę DISSTW [2], która realizuje wybór zbioru zmiennych o największej sile dyskryminacji w sensie przyjętego wyżej kryterium Λ. Procedura ta jest generalnie mówiąc procedurą krokową i może działać w kilku wariantach: Można mianowicie obowiązkowo wprowadzać określone zmienne do zbioru dyskryminacji, następnie dla danych liczb $r2$, $r1$, $0 < r2 \leq p$, $0 < r1 \leq p$, można dołączać do tego zbioru dalsze zmienne, aż ich łączna liczba osiągnie wielkość $r2$, a następnie z danego zbioru dyskryminacji eliminować zmienne, aż liczebność zbioru dyskryminacji osiągnie wielkość $r1$. Procedura ta zastosowana do naszego materiału dała wyniki zestawione w Tablicy 1.

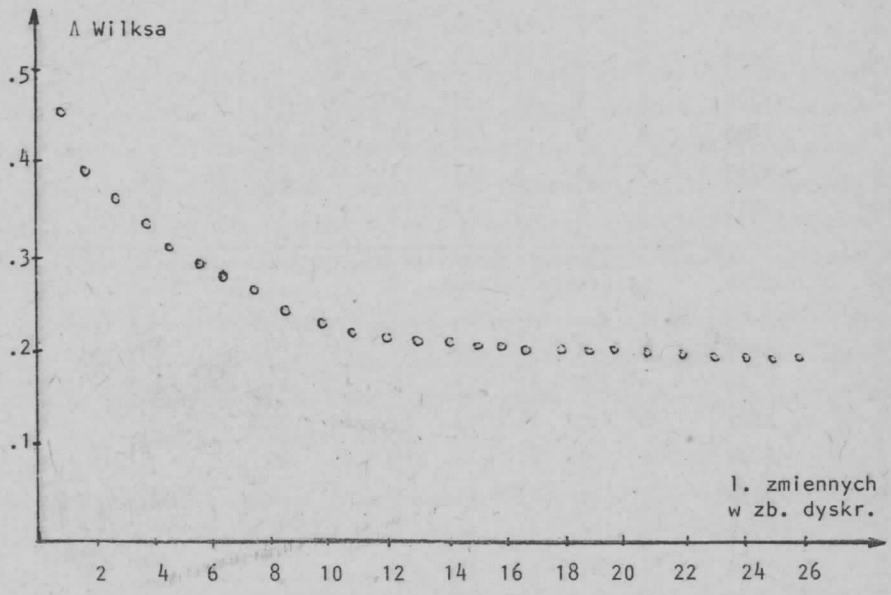
Tablica 1

Proces wyboru i eliminacji zmiennych

Liczba zmiennych w zbiorze dyskr.	Wartość statystyki Λ	Numery zmiennych w zbiorze dyskryminacji									
1	.4582	24									
2	.3993	15	24								
3	.3709	11	15	24							
4	.3456	11	13	15	24						
5	.3238	9	11	13	15	24					
6	.3091	6	9	11	13	15	24				
7	.2950	6	9	11	13	15	24	25			
8	.2838	6	9	11	13	15	24	25	26		
9	.2740	2	6	9	11	13	15	24	25	26	
10	.2651	2	6	9	11	13	15	17	24	25	26
26	.2058	wszystkie zmienne									
10	.2634	6	9	11	13	15	16	18	21	25	26
9	.2732	6	9	11	13	15	18	21	25	26	
8	.2802	6	9	11	13	15	18	25	26		
7	.2893	6	9	11	13	15	25	26			
6	.3036	6	9	11	13	15	26				
5	.3192	6	9	11	13	15					
4	.3530	9	11	13	15						
3	.3691	11	13	15							
2	.4242	11	13								
1	.4882	13									

W tablicy tej przedstawiono proces krokowego dołączania zmiennych do zbioru dyskryminacji, a następnie stopniowej ich eliminacji. Dla kolejnych kroków podano liczbę zmiennych znajdujących się w zbiorze dyskryminacji, wielkość wskaźnika Λ oraz numery zmiennych. Jak widać z wyników umieszczonych w Tablicy 1, najlepsza dziesiątka zmiennych otrzymana przy dołączaniu zmiennych nie jest identyczna z najlepszą dziesiątką otrzymaną przy eliminacji zmiennych. Fakt ten nie powinien budzić zdziwienia, ponieważ metoda krokowa nie gwarantuje rozwiązania optymalnego, a jedynie rozwiązanie suboptymalne, zbliżone do niego.

W rozważanym przykładzie wielkości kryterium Λ są zbliżone: dla $p = 10$ przy dołączaniu zmiennych otrzymujemy $\Lambda_{(10)} = .2651$, natomiast przy eliminacji zmiennych otrzymujemy $\Lambda_{(10)} = .2634$. Najlepszą pojedynczą zmienną wybraną spośród wszystkich 26 zmiennych okazała się zmienna nr 24 (odpowiadająca jej wartość $\Lambda_{(1)}$ wynosi $\Lambda'_{(1)} = .4582$). Eliminując zmienne z pełnego zbioru dyskryminacyjnego otrzymaliśmy na koniec pojedynczą zmienną o numerze 13, dla której odpowiednia wartość $\Lambda'_{(1)}$ wynosi $\Lambda'_{(1)} = .4882$, a więc jest nieco gorsza w sensie przyjętego kryterium od najlepszej pojedynczej zmiennej, jaką jest naprawdę zmienna nr 24.



Rys. 1. Wartość statystyki Λ jako funkcja liczby zmiennych w zbiorze dyskryminacji.

Rys. 1 przedstawia wielkość wskaźnika Λ jako funkcję liczby zmiennych znajdujących się w zbiorze dyskryminacji. Widzimy, że wskaźnik ten maleje najpierw stosunkowo szybko, a następnie w okolicy $p = 10$ stabilizuje się, co oznacza, że dalsze dodawanie zmiennych do zbioru dyskryminacji nie powoduje już w sposób istotny powiększenia jego siły dyskryminacji.

3. Diagnoza czyli identyfikacja metodą kanonicznych funkcji dyskryminacyjnych

Dane są wektory obserwacji p zmiennych u osobników wylosowanych z G rozważanych populacji. Chcemy znaleźć nową zmienną z będącą funkcją liniową zmiennych $x_1, x_2, \ldots, x_p$ taką, żeby jej wartości obliczone dla osobników należących do różnych populacji były możliwie zróżnicowane. Niech $u = (u_1, u_2, \ldots, u_p)'$ oznacza wektor współczynników formy liniowej tworzącej zmienną z:

$$z = u'x = u_1x_1 + u_2x_2 + \ldots + u_px_p.$$

Jako kryterium własności różnicujących zmiennej z przyjmijmy tym razem stosunek zmienności międzygrupowej do zmienności wewnątrzgrupowej (por. [1], [4]):

$$\Phi = \frac{SS_b}{SS_w},$$

gdzie

$$SS_b = \sum_{g=1}^{k} ng(\bar{z}_{\cdot}^{(g)} - \bar{\bar{z}}_{\cdot})^2 = u'Bu,$$

$$SS_w = \sum_{g=1}^{k} \sum_{i=1}^{n_g} (z_i^{(g)} - \bar{z}_{\cdot}^{(g)})^2 = u'Wu,$$

przy macierzach A, W zdefiniowanych wzorami (8) i (9). Wobec powyższego nasze kryterium przyjmuje postać

$$\Phi = \frac{u'Bu}{u'Wu}. \tag{16}$$

Można pokazać, że kryterium Φ osiąga maksimum, gdy jako wektor u przyjmiemy największy wektor własny równania macierzowego:

(17) $$(B-\lambda W)u = 0$$

lub równoważnego mu równania

(17a) $$\left(\frac{1}{n-k} B-\lambda \frac{1}{n-k} W\right)u = 0.$$

Wektor u będzie określony w sposób jednoznaczny, gdy przyjmiemy dodatkowe restrykcje, np. żeby składowe wektora u spełniały warunek sigma-ortogonalności:

(18) $$\frac{1}{n-k} u'Wu = 1.$$

Niech t oznacza rząd macierzy B, przy czym jeśli obserwowane zmienne $x_1, x_2, \ldots, x_p$ są liniowo niezależne, zachodzi równość:

(19) $$t = \operatorname{rg}(B) = \min(k-1, p).$$

Można pokazać, że równanie macierzowe (17) ma t pierwiastków różnych od zera oznaczanych $\lambda_1, \lambda_2, \ldots, \lambda_t$ i t odpowiadających im wektorów własnych $u_1, u_2, \ldots, u_t$. Każdy z nich spełnia warunek simga-ortogonalności (18), który możemy zapisać w postaci $U' \frac{1}{n-k} WU = I$, gdzie $U = (u_1, u_2, \ldots, u_t)$. Każdy z wektorów u_l określa formę liniową tworzącą nową zmienną kanoniczną $z_l = u_l' x$. Zmienne te są ortogonalne, co oznacza, że

$$ZZ' = I,$$

gdzie $Z = (z_1, z_2, \ldots, z_t)'$.

Jednocześnie wartości kryterium Φ wyznaczone dla l-tej zmiennej kanonicznej wynosi

(20) $$\Phi^{(l)} = \frac{u_l' B u_l}{u_l' W u_l} = \lambda_l, \quad \lambda = 1, 2, \ldots, t.$$

Dla naszych danych otrzymaliśmy następujące pierwiastki charakterystyczne (zestawy I i II zmiennych oznaczają zmienne wyszczególnione w Tablicy 1, otrzymane przy dołączaniu i eliminacji zmiennych):

liczba cech	λ_1	λ_2	λ_3	λ_4
$p = 26$	1.739	.400	.150	.102
$p = 10$ (I zestaw)	1.540	.300	.093	.045
$p = 10$ (II zestaw)	1.510	.284	.116	.055

Jednocześnie odpowiedni test (por. [1]) wykazał, że statystycznie istotne są jedynie dwie pierwsze zmienne kanoniczne. Wynika stąd, że proces diagnostyczny może być przeprowadzony za pomocą dwóch pierwszych zmiennych, które dają się odwzorować na płaszczyźnie w układzie współrzędnych ortogonalnych.

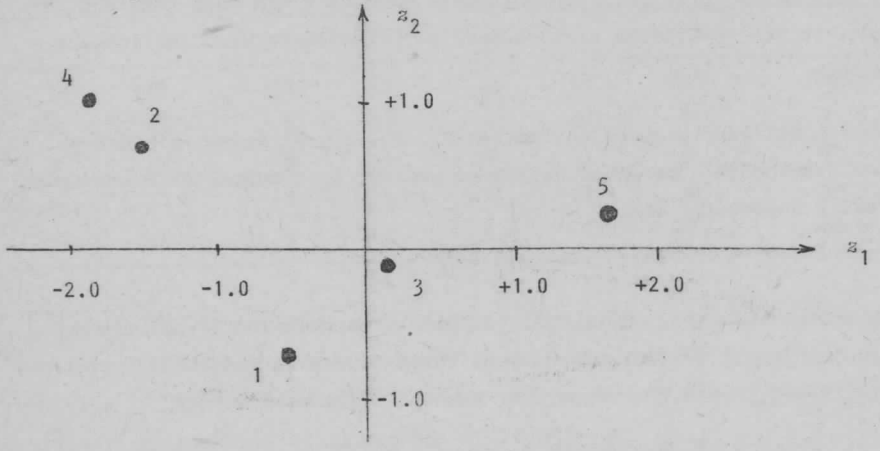
W Tablicy 2 podano współrzędne średnich grupowych w układzie dwóch pierwszych zmiennych kanonicznych.

Tablica 2

Współrzędne średnich grupowych w układzie dwóch pierwszych zmiennych kanonicznych

Numer grupy	$p = 26$		$p = 10$ (I zestaw)		$p = 10$ (II zestaw)	
	$\bar{z}_{1.}$	$\bar{z}_{2.}$	$\bar{z}_{1.}$	$\bar{z}_{2.}$	$\bar{z}_{1.}$	$\bar{z}_{2.}$
1	-.56	-.80	-.50	-.69	-.51	-.69
2	-1.54	+.66	-1.49	+.56	-1.48	+.61
3	+.16	-.11	+.13	-.13	+.09	+.00
4	-1.87	+1.00	-1.75	+.87	-1.70	+.72
5	+1.65	+.28	+1.55	+.25	+1.54	+.23

Wykres średnich dla pełnego zestawu $p = 26$ cech oraz dla I zestawu $p = 10$ cech jest przedstawiony jako rysunek 2.



Rys. 2. Wykres średnich grupowych w układzie dwóch pierwszych zmiennych kanonicznych.

Obliczanie przynależności do grup, czyli diagnoza może odbywać się według jednej z następujących zasad:

a) w przestrzeni kanonicznych zmiennych dyskryminacyjnych należy obliczyć odległości euklidesowe danego punktu indywidualnego od punktów średnich dla poszczególnych grup (chorób); a następnie zaliczyć badany punkt do tej grupy, od której jego odległość jest najmniejsza,

b) dla kolejnych grup $g = 1, 2, \ldots, G$ należy wykonać test statystyczny weryfikujący hipotezę, czy badany punkt może należeć do g-tej populacji; w tym celu obliczamy statystyki $F_{0/1}, F_{0/2}, \ldots, F_{0/G}$ weryfikujące hipotezę, że punkt 0 należy do g-tej populacji, oraz odpowiadające tym statystykom prawdopodobieństwa

$$P_1, P_2, \ldots, P_G$$

przekroczenia obliczonej wartości $F_{0/g}$ pod warunkiem prawdziwości hipotezy H_g orzekającej, że punkt 0 pochodzi z g-tej populacji, $g = 1, 2, \ldots, G$. Badany punkt 0 należy zaliczyć do tej populacji, dla której obliczona wartość P_G jest największa.

Statystyki testowe $F_{0/g}$ obliczamy ze wzoru:

$$(21) \qquad F_{0/g} = k_g = \frac{n-k-t+1}{t(n-k)} \frac{n_g}{n_g+1} \sum_{h=1}^{t} (z_h - \bar{z}_h^{(g)})^2, \quad g = 1, 2, \ldots, G.$$

Testujemy tutaj hipotezę, że badany punkt z ma rozkład normalny $N(\bar{z}_\cdot^{(g)}, I)$. Ponieważ zmienne kanoniczne są kombinacjami liniowymi dużej liczby zmiennych, więc możemy przyjąć, że nawet jeśli zmienne x nie mają rozkładu normalnego, to ich kombinacja liniowa będzie już miała rozkład zbliżony do normalnego.

Zamiast obliczać prawdopodobieństwa $P_1, P_2, \ldots, P_G$ możemy dla danego poziomu istotności α porównać obliczoną wartość k_g z odpowiednim kwantylem rozkładu F Snedecora. Jeśli

$$(22) \qquad k_g \leq F_{t,n-k-t+1;\alpha}$$

to nie możemy wykluczyć możliwości, że badany osobnik został wylosowany z g-tej populacji. Dla naszych danych, to jest przy wartościach $t = 4$, $n = 304$, $k = 5$, $\alpha = 0.05$, odpowiedni kwantyl rozkładu F wynosi

$$F_{4,294;.05} \approx F_{4,300;.05} = 2.40.$$

W naszych obliczeniach przyjęliśmy zasadę, że osobnika zaliczamy do tej grupy, od której jego odległość euklidesowa jest najmniejsza. Stosując tę zasadę obliczono dla posiadanego materiału n = 303 chorych odpowiednie odległości, po czym przeprowadzono diagnozę. Odtworzone liczebności grup są przedstawione w Tablicy 3.

Tablica 3

Liczebności grupowe odtworzone za pomocą kanonicznych zmiennych dyskryminacyjnych

numer grupy	liczebność grupy	odtworzone liczebności grup grupa 1	grupa 2	grupa 3	grupa 4	grupa 5	% dobrych klasyf.
pełny zestaw p = 26 cech							
1	103	63	11	7	9	13	61.2
2	48	8	19	3	15	3	39.6
3	23	5	2	10	1	5	43.5
4	24	1	4	2	17	0	70.8
5	105	2	2	5	1	95	90.5
zestaw I p = 10 cech							
1	103	48	10	17	8	20	46.6
2	48	11	21	3	11	2	43.7
3	23	7	2	6	2	6	26.1
4	24	1	4	2	17	0	70.8
5	105	3	1	16	1	84	80.0
zestaw II p = 10 cech							
1	103	51	12	7	11	22	49.5
2	48	8	17	3	16	4	35.4
3	23	5	1	10	1	6	43.5
4	24	1	5	4	14	0	58.3
5	105	4	1	6	1	93	88.6

Jak wynika z Tablicy 3, siła diagnostyczna zmiennych zestawu I i zestawu II jest zbliżona, jednak zestaw I daje więcej dobrych klasyfikacji w grupie 2 i 4 (powikłania), natomiast zestaw II daje więcej dobrych klasyfikacji w grupie 5, czyli grupie kontrolnej.

Siła diagnostyczna obydwu zestawów $p = 10$ zmiennych jest niewiele mniejsza niż siła diagnostyczna wszystkich $p = 26$ badanych zmiennych, co potwierdza naszą uwagę wypowiedzianą pod koniec rozdziału 2 tej pracy.

Tablica 4

Przykładowe wartości współrzędnych niektórych punktów indywidualnych w układzie zmiennych kanonicznych

Współrzędne obliczone z $p = 26$ cech					
Numer grupy	Numer punktu	z_1	z_2	z_3	z_4
1	1	-1.089	-1.288	-.300	-.494
1	2	+.776	-.018	-.737	+.078
1	3	-1.406	-3.231	1.084	+.717
1	4	-1.049	-2.124	+.738	+.797
2	1	-1.822	+.295	+2.262	+1.440
2	2	-1.637	+.971	+.002	-.403
5	1	+1.673	+.106	-.283	-1.784
5	2	+1.424	+1.616	-.747	-.995
5	3	+2.822	+.115	+.382	-1.082
5	4	+1.848	+.379	+1.324	+1.286
współrzędne obliczone z $p = 10$ cech (zestaw I)					
1	1	-1.098	-1.896	+.551	-.747
1	2	+.890	+.243	+1.004	+1.090
1	3	-1.253	-3.172	-.252	+.713
1	4	-1.520	-1.845	-.628	-.207
2	1	-1.657	+.207	-2.282	+.213
2	2	-1.450	+.942	+.176	+.153
5	1	+1.498	+.157	+1.080	-2.243
5	2	+1.470	+1.713	+1.195	-.697
5	3	+2.743	+.162	+.344	-1.130
5	4	+1.470	+.182	-1.671	-.186

Tablica 5

Przykładowe wartości statystyki k_g umożliwiające diagnozę

Numer grupy	Numer osobnika	k_1	k_2	k_3	k_4	k_5
Wartości statystyki k_g obliczone z p = 26 cech						
1	1	.20+	1.29	1.15	1.42	2.53
1	2	.761	1.77	.20+	.07	.38
1	3	2.04+	3.81	4.02	5.42	5.70
1	4	.79+	2.00	2.11	3.41	3.48
2	1	2.43	1.13+	3.80	3.08	4.67
2	2	1.07	.22	1.58	.10+	2.80
5	1	2.16	3.85	2.10	3.47	.77+
5	2	2.76	3.16	1.75	2.70	.84+
5	3	3.28	5.22	2.91	5.66	.62+
5	4	2.62	3.19	2.22	5.15	.83+
Wartości statystyki k_g obliczone z p = 10 cech, I zestaw						
1	1	.62+	1.93	1.59	2.00	3.05
1	2	1.27	2.18+	.30	2.19	.69
1	3	1.82+	3.49	2.75	4.39	4.96
1	4	.68+	1.42	1.78	2.21	3.48
2	1	1.83	.82+	2.52	2.31	3.78
2	2	.89	.14	.94	.13+	2.34
5	1	2.59	4.12	2.61	3.63	1.48+
5	2	2.81	3.29	1.84	2.74	.99+
5	3	3.06	4.91	2.42	5.15	.67+
5	4	1.83	2.52	1.62	3.95	.67+

W Tablicy 4 podajemy przykładowe współrzędne wybranych na chybił-trafił punktów indywidualnych w układzie wszystkich czterech zmiennych kanonicznych. Współrzędne te są podane zarówno dla zmiennych kanonicznych obliczonych ze wszystkich p = 26 cech jak również dla I-go zestawu zredukowanego obejmującego p = 10 cech. Zauważyliśmy już wcześniej, że tylko dwie pierwsze zmienne kanoniczne są statystycznie istotne dla zagadnienia dyskryminacji, natomiast pozostałe dwie wyrażają inne własności charakteryzujące badanych osobników, ale nie posiadające własności różnicujących. Stosownie

do tej uwagi możemy w Tablicy 4 zobaczyć, że pierwsze dwie współrzędne obliczone dla $p = 26$ i $p = 10$ są dość podobne, pozostałe dwie - różne.

Przypuśćmy, że chcemy postawić diagnozę dla punktu P. Nanosząc współrzędne tego punktu na wykres średnich grupowych przedstawiony przez nas wcześniej jako Rys. 2 możemy „naocznie" zobaczyć podobieństwo danego punktu do rozważanych grup oznaczających różne choroby.

W ten sposób np. punkt oznaczony przez nas symbolem (1,1) o współrzędnych (-1.089,-1.288) jest najbliższy średniej 1-szej grupy, w naszym przypadku astmatyków, natomiast daleki od średnich pozostałych grup. Tak więc w tym przypadku możemy być skłonni do przypuszczenia, że badany punkt (1,1) należy do grupy astmatyków. Analogicznie dla ostatniego punktu Tablicy 4 możemy zobaczyć, że punkt ten jest najbliższy średniej grupy piątej (kontrolnej), a więc że pacjent o tych współrzędnych nie wykazuje objawów chorobowych.

Tablica 5 podaje dla tych samych pacjentów, których uwzględniono w Tablicy 4, odpowiednie wartości statystyki k_g określonej wzorem (22). Przyjmując jako wartość krytyczną tej statystyki wartość $F_{4,300,.05} = 2.40$ dochodzimy np. dla punktów (1,1) i (5,4) do następujących konkluzji:

Punkt (1,1) jest co prawda najbliżej środka grupy I oraz prawdopodobieństwo przynależności tego punktu do grupy I jest największe, jednak na podstawie obliczonych statystyk k_g, $g = 1, 2, \ldots, 5$, możemy wykluczyć jedynie przynależność tego punktu do grupy V.

Jeśli chodzi o punkt (5,4), to najbardziej prawdopodobna jest jego przynależność do grupy V, jednak nie można wykluczyć przynależności tego punktu do grupy II lub IV.

4. Diagnoza, czyli identyfikacja metodą liniowych i kwadratowych funkcji dyskryminacyjnych przy założeniu normalności rozkładu

Niech prawdopodobieństwa $\Pi_1, \Pi_2, \ldots, \Pi_G$ będą prawdopodobieństwami a priori, że badany osobnik pochodzi z populacji o numerze $1, 2, \ldots, G$. Załóżmy dalej, że badamy p cech, czyli że otrzymujemy dla każdego osobnika wektor $x = (x_1, x_2, \ldots, x_p)'$ oraz że wektor ten ma p-wymiarowy rozkład normalny

$N_p(\mu^{(i)},\Sigma_i)$ o gęstości prawdopodobieństwa $f(x;\mu^{(i)},\Sigma_i)$ określonej wzorem

$$(23)\qquad f(x;\mu^{(i)},\Sigma_i) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}\exp\left\{\frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)})\right\},$$

$$i = 1, 2,\ldots, G.$$

Badany osobnik należy do jednej z rozważanych G populacji, które mogą różnić się zarówno wektorem wartości średnich $\mu^{(i)}$, jak i macierzą kowariancji Σ_i. Korzystając z wzoru Bayesa obliczamy prawdopodobieństwo a posteriori, że obserwacja x została wylosowana z rozkładu o numerze i:

$$(24)\qquad P(i/x) = \frac{\Pi_i f(x;\mu^{(i)},\Sigma_i)}{\Sigma\Pi_i f(x;\mu^{(i)},\Sigma_i)}.$$

Badany wektor x zaliczamy do tej populacji, dla której jego prawdopodobieństwo a posteriori jest największe.

Diagnoza na podstawie prawdopodobieństwa $P(i/x)$ określonego wzorem (24) jest równoważna diagnozie wykonywanej na podstawie wartości samego tylko licznika wzoru (24); a również na podstawie logarytmu z wartości tego licznika; czyli na podstawie wartości wyrażenia:

$$(25)\quad S_i = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu^{(i)})'\Sigma_i^{-1}(x-\mu^{(i)}) + \ln\Pi_i, \quad i = 1, 2,\ldots, G.$$

Jest to wyrażenie służące tzw. dyskryminacji kwadratowej, ponieważ identyfikacja odbywa się tu na podstawie formy kwadratowej obliczanej z danego wektora obserwacji x.

Jeśli założymy, że wszystkie macierze kowariancji są równe, czyli

$$\Sigma_1 = \Sigma_2 = \ldots = \Sigma_G = \Sigma$$

to człon $-\frac{1}{2}\ln|\Sigma| - \frac{1}{2}x'\Sigma^{-1}x$ jest wspólny dla wszystkich wartości S_i, $i = 1, 2,\ldots, G$, i możemy go pominąć. Otrzymujemy wtedy wyrażenie

$$(26)\qquad S_i' = \mu^{(i)'}\Sigma^{-1}x - \frac{1}{2}\mu_i\Sigma^{-1}\mu_i + \ln\Pi_i.$$

Jest to wyrażenie służące tzw. dyskryminacji liniowej, ponieważ identyfikacja odbywa się tu na podstawie liniowej funkcji obserwacji. Wektor $\mu^{(i)'}\Sigma^{-1} = \ell$ określa tzw. liniową funkcję dyskryminacyjną. Na ogół nie znamy dokładnych wartości $\mu^{(i)}$ oraz Σ_i i w obliczeniach posługujemy się odpowiednimi estymatorami obliczonymi z próby.

W Tablicy 6 podano liczebności grupowe odtworzone za pomocą liniowych i kwadratowych funkcji dyskryminacyjnych, natomiast w Tablicy 7 podano przykładowe wartości prawdopodobieństw a posteriori obliczonych metodą dyskryminacji liniowej i kwadratowej. Wyniki te dotyczą zestawu I zawierającego $p = 10$ cech.

Nie można było wykonać dyskryminacji kwadratowej dla pełnego zestawu $p = 26$ cech, ponieważ liczebności w grupach 2 i 4 były zbyt małe, żeby obliczyć wyznacznik i macierz odwrotną z próbkowych macierzy kowariancji.

Dla niektórych innych zestawów cech wystąpiła trudność w obliczaniu prawdopodobieństwa a posteriori explicite według wzoru (24); mianowicie okazało się, że wyrażenie znajdujące się w wykładniku funkcji $\exp(\cdot)$ jest zbyt duże lub zbyt małe, i nie można obliczyć wartości tej funkcji. Tak np. używając translatora ALGOLu na m.c. ODRA 1204 (mta3) można obliczyć funkcję $\exp(x)$ jedynie dla $x < 510.99999998 \ln 2$.

Tablica 6

Liczebności grupowe odtworzone za pomocą liniowych i kwadratowych funkcji dyskryminacyjnych obliczanych dla zestawu I, $p = 10$ cech

numer grupy	1	2	3	4	5	liczebność grupy	% dobrych klasyfikacji
			Dyskryminacja liniowa				
1	66	7	1	5	24	103	64.08
2	18	18	1	8	3	48	37.50
3	12	0	3	0	8	23	13.04
4	5	6	1	11	1	24	43.83
5	6	1	0	1	97	105	92.38
			Dyskryminacja kwadratowa				
1	67	5	7	5	19	103	65.04
2	11	23	3	8	3	48	47.97
3	3	1	10	1	8	23	43.48
4	1	1	0	21	1	24	87.50
5	4	1	0	0	100	105	95.24

Tablica 7

Przykładowe wartości prawdopodobieństw a posteriori

Numer grupy	Numer osob.	P(1/x)	P(2/x)	P(3/x)	P(4/x)	P(5/x)
			Dyskryminacja liniowa			
1	1	.92721+	.02893	.02560	.01154	.00673
1	2	.16596	.01160	.26421	.00520	.55304+
1	3	.95751+	.01357	.02642	.00089	.00162
1	4	.88308+	.08736	.01864	.00793	.00299
2	1	.21280	.75744+	.00974	.01597	.00405
2	2	.22295	.47817+	.04240	.24460	.01187
5	1	.09044	.00168	.01609	.00199	.88979+
5	2	.02244	.00370	.03232	.00507	.93648+
5	3	.00735	.00007	.00516	.00002	.98740+
5	4	.08033	.00878	.02455	.00019	.88616+
			Dyskryminacja kwadratowa			
1	1	.92766+	.07219	.00014	.00001	.00000
1	2	.00282	.00288	.00198	.00000	.99232+
1	3	.99992+	.00008	.00000	.00000	.00000
1	4	.34537	.64453+	.00009	.00000	.00000
2	1	.64753+	.35245	.00002	.00000	.00000
2	2	.10712	.28818	.01745	.58725+	.00000
5	1	.00179	.00000	.00000	.00000	.99821+
5	2	.00007	.00000	.00000	.00000	.99993+
5	3	.00014	.00000	.00000	.00000	.99986+
5	4	.00085	.00000	.00026	.00000	.99889+

Jako prawdopodobieństwa a priori przyjęto frakcje liczebności poszczególnych grup liczone w stosunku do sumarycznej liczebności wszystkich grup.

Porównując liczebności grupowe odtworzone za pomocą liniowych i kwadratowych funkcji dyskryminacyjnych widzimy w Tablicy 6, że dyskryminacja kwadratowa daje nieco lepsze wyniki niż dyskryminacja liniowa. W szczególności liczba dobrze zaklasyfikowanych osób przy dyskryminacji kwadratowej równa się 221, natomiast przy dyskryminacji liniowej tylko 203.

5. Wnioski końcowe

W pracy rozważono możliwości stosowania do automatycznej diagnozy następujących metod:

a) klasycznej analizy dyskryminacyjnej poprzez liniowe i kwadratowe funkcje dyskryminacyjne,

b) kanonicznych zmiennych dyskryminacyjnych.

Wydaje się nam, że szczególnie użyteczna przy automatycznej diagnozie może być metoda kanonicznych zmiennych dyskryminacyjnych. Metoda ta w swej zasadniczej części nie wymaga założeń o normalności rozkładów i pozwala znacznie zredukować wymiarowość zagadnienia. Ponadto metoda ta pozwala sporządzić wykres punktów indywidualnych, ukazujący położenie badanego punktu na tle średnich badanych grup, do których ten punkt może należeć.

Niezależnie od tego sądzimy, że również użyteczne może być obliczanie prawdopodobieństw a posteriori za pomocą liniowych funkcji dyskryminacyjnych. Prawdopodobieństwa te są pewną unormowaną miarą możliwości należenia badanego punktu do poszczególnych grup.

Przed rozpoczęciem procesu diagnozowania warto odrzucić zmienne, które powtarzają zawarte w innych zmiennych własności różnicujące badane grupy. Bardzo użyteczna do tego celu okazała się procedura DISSTW opisana w rozdziale 2 tej pracy.

Literatura cytowana

[1] H. Ahrens, J. Lauter, *Mehrdimensionale Varianzanalyse*, Akademie-Verlag Berlin 1974.

[2] A. Bartkowiak, *Stepwise selection of discriminative variables*, Applicationes Mathematicae XVIII.1 (w druku).

[3] J. Leowski i wsp., *Diagnostyka epidemiologiczna chorób narządów klatki piersiowej i jej wykorzystanie w zintegrowanym systemie ochrony zdrowia*, Gruźlica 1974, 42, 1077.

[4] P.A. Lachenbruch, *Discriminant analysis*, Hafner Press, Macmillan Pub. Co, 1975.

[5] Mc Cabe, *Computations for variable selections in discriminant analysis*, Technometrics 17 (1975), 103-109.

[6] C.R. Rao, *Linear statistical inference and its applications*, New York 1965.

[7] F. Sawicki, *Przewlekłe nieswoiste choroby układu oddechowego w Krakowie*, PZH Warszawa 1977.