Listy Biometryczne
Vol. XXII 1985 No.2
Biometrical Letters

# O PEWNEJ METODZIE PREZENTACJI DANYCH WIELOWYMIAROWYCH

RADOSŁAW KALA, IDZI SIATKOWSKI

Zakład Metod Matematycznych i Statystycznych Akademii Rolniczej w Poznaniu

**Praca wpłynęła 8 października 1984; w wersji ostatecznej 4 marca 1985**

Kala R., Siatkowski I., 1985. On a method of representing multidimentional data. Listy Biometryczne XXII, z.2. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu (Adam Mickiewicz University Press) pp. 19-32, 4 tab., 7 ryc., PL ISSN 1458-0036.

**In the analysis of multivariate data it is very important to represent the data in such a way that the relationships between objects can easily be seen. One of such methods is the method of trees proposed by Kleiner and Hartigan. In the paper the principles of this method are explained and some examples showing its usefulness are described. (On a method of representing multidimentional data).**

## 1. WSTĘP

Przy opracowywaniu wyników eksperymentalnych jednym z ważnych elementów jest zagadnienie przejrzystej ich prezentacji tak, aby ułatwić bezpośrednie wychwycenie istotnych informacji. W przypadku danych jednowymiarowych, a także dwuwymiarowych, wtedy, gdy dążymy do zilustrowania różnic bądź podobieństw między badanymi obiektami, wystarczającym narzędziem jest w zasadzie oś liczbowa bądź płaszczyzna z układem współrzędnych prostokątnych, na których, w zależności od zaobserwowanych wartości cech, zaznacza się punktami kolejne obiekty.

Rozszerzenie eksperymentu poprzez zwiększenie liczby obserwowanych cech prowadzi do danych wielowymiarowych, których graficzna prezentacja napotyka na znaczne trudności, przy czym są one tym większe, im większą liczbę cech uwzględniamy w doświadczeniu. Z drugiej strony ograniczone możliwości bezpośredniego odczytania związków łączących badane obiekty wielowymiarowe wprost z tablicy obserwacji wywołują znacznie większą potrzebę graficznej prezentacji danych wielowymiarowych niż jednowymiarowych. Wychodząc na-

przeciw temu zapotrzebowaniu, w okresie ostatniego ćwierćwiecza zaproponowano w literaturze statystycznej szereg metod, polegających, ogólnie rzecz biorąc, na przyporządkowaniu każdemu p-wymiarowemu obiektowi pewnego umownego symbolu.

Wśród metod elementarnych wymienić można metodę profili, polegającą na wyznaczeniu dla każdego obiektu szeregu prostokątów o stałej szerokości podstawy i o wysokościach proporcjonalnych do wartości kolejnych zmiennych, czy metodę wielokątów, w której profile zastąpione są gwiazdami o odpowiednio dobranych długościach ramion. Metody te, choć skuteczne dla obiektów o małej liczbie wymiarów, stają się nieprzydatne, gdy w prezentacji danych chcemy zmienne zgrupować tak, aby cechy bardziej skorelowane były bardziej do siebie zbliżone.

Spośród metod zaawansowanych na uwagę zasługuje metoda Chernoffa(1973), w której obserwowane zmienne utożsamia się z określonymi rysami twarzy, takimi jak: jej owal, długość nosa, szerokość ust, długość brwi itd. W rezultacie prowadzi to do przyporządkowania badanym obiektom schematycznych twarzy, których obrazy uzyskiwane są przy użyciu komputera. Jak przy tej okazji zauważa Chernoff (1973, s.365-366), w ten sposób powstaje zabawne odwrócenie postępowania znanego w badaniach nad sztuczną inteligencją. Mianowicie, zamiast rozpoznawania za pomocą maszyny cyfrowej twarzy ludzi w oparciu o ich opis liczbowy, w omawianej metodzie przy rozpoznawaniu obiektów opisanych liczbami wykorzystuje się maszynę do pracochłonnego rysowania schematycznych twarzy, pozostawiając sam proces rozpoznawania inteligencji człowieka.

Metoda Chernoffa daje czytelny opis danych wielowymiarowych nawet przy 18 cechach. Stawia jednak przed użytkownikiem szereg problemów wynikających przede wszystkim z odpowiedniego przyporządkowania zmiennych do rysów twarzy, a także z konieczności stosowania skomplikowanych programów rysujących.

Metodami stosunkowo prostymi i, jak się wydaje, pozbawionymi wad metod wyżej przedstawionych są dwie bliźniacze metody, zaproponowane przez Kleinera i Hartigana (1981). W jednej z nich obiekty wielowymiarowe reprezentuje się w formie schematycznych drzew, a w drugiej w postaci zarysów zamków. Metody te można uznać za rozwiązanie kompromisowe. Ich stosowanie nie jest co prawda wolne od wykonania wstępnych obliczeń, ale z drugiej strony uwzględniają one relacje pomiędzy zmiennymi, zachowują możliwość odczytania wartości cech bezpośrednio z symbolu graficznego, a także mogą być użyte nawet dla obiektów o dużej liczbie cech. Oczywiście stwierdzenie o wzroście trudności interpretacyjnych wraz ze wzrostem liczby wymiarów pozostaje w mocy.

Jedną z tych metod, a mianowicie metodę drzew, przedstawimy szczegółowo w dalszych paragrafach. Podamy nie tylko zasady wyznaczania odpowiedniego symbolu graficznego w oparciu o tablicę danych, ale także przedstawimy przykłady obrazujące przydatność tej metody.

# 2. DRZEWA

Niech uzyskane w eksperymencie obserwacje będą uporządkowane w formie tablicy o n wierszach i p kolumnach, przy czym niech jej wiersze - odpowiadają badanym obiektom, a kolumny cechom lub zmiennym. W ten sposób każdy z n obiektów reprezentowany jest przez punkt w p-wymiarowej przestrzeni euklidesowej, wyznaczony za pomocą wektora (wiersza) zaobserwowanych wartości poszczególnych cech lub odwrotnie - każdej z p cech odpowiada punkt w przestrzeni n-wymiarowej, wyznaczony przez wektor (kolumnę) wartości zaobserwowanych dla poszczególnych obiektów.

Podstawą metody Kleinera i Hartigana reprezentacji danych wielowymiarowych jest uporządkowanie cech, co uzyskuje się dokonując hierarchicznego ich grupowania w oparciu o p·(p-1)/2-elementową tablicę odległości,wyznaczonych dla cech jako punktów w przestrzeni n-wymiarowej. Użyta metoda grupowania hierarchicznego nie jest tu istotna, chociaż zalecane są metody, dające w kolejnych etapach grupy cech o zbliżonych liczebnościach. Dobre rezultaty daje tu metoda najdalszego sąsiedztwa (patrz np. Hartigan, 1975 lub Mardia i in. 1979), preferowana przez Kleinera i Hartigana, a także metoda Warda (patrz Wishart, 1969 lub Karoński i Caliński, 1973) wykorzystywana w niniejszej pracy.

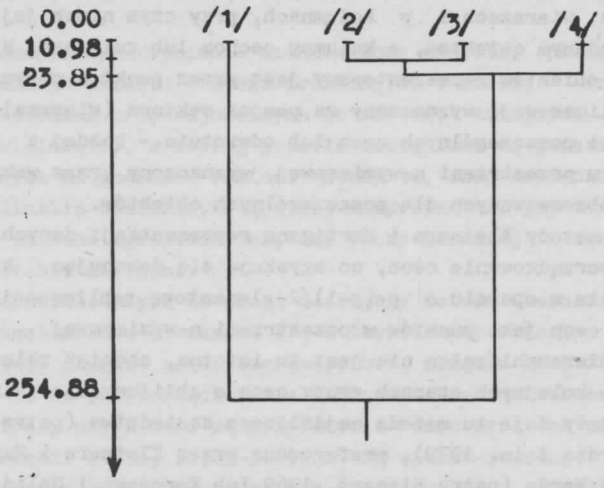
Otrzymany w rezultacie grupowania cech dendrogram, obrazujący związki między zmiennymi, wraz z odległościami między grupami cech oraz z tablicą obserwacji stanowi podstawę do wyznaczenia schematycznych drzew reprezentujących poszczególne obiekty. W konstrukcji tej dendrogram wraz z odległościami między grupami cech wyznaczają jednakową dla wszystkich n obiektów strukturę, podczas gdy tablica obserwacji uśrednionych względem wynikających z dendrogramu grup zmiennych decyduje o długościach poszczególnych segmentów indywidualnych drzew.

W celu zilustrowania postępowania posłużymy się przykładem zaczerpniętym z książki Hartigana (1975). W doświadczeniu badano mleko sześciu (n=6) ssaków, określając procentowy udział czterech (p=4) składników, którymi były: (1) woda, (2) białko, (3) tłuszcz i (4) cukier mlekowy. Uzyskane obserwacje obrazuje tablica 1.

T a b l i c a 1. Procentowy skład mleka dla wybranych ssaków

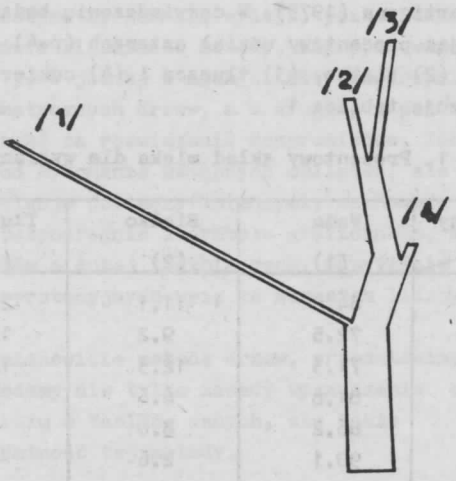
| Cechy / Obiekty | Woda (1) | Białko (2) | Tłuszcz (3) | Cukier mlekowy (4) |
|---|---|---|---|---|
| Wieloryb | 64.8 | 11.1 | 21.2 | 1.6 |
| Szczur | 72.5 | 9.2 | 12.6 | 3.3 |
| Królik | 71.3 | 12.3 | 13.1 | 1.9 |
| Lis | 81.6 | 6.6 | 5.9 | 4.9 |
| Zebra | 86.2 | 3.0 | 4.8 | 5.3 |
| Koń | 90.1 | 2.6 | 1.0 | 6.0 |

Po zastosowaniu metody Warda grupowania hierarchicznego cech otrzymano dendrogram prezentowany na rycinie 1.




**Ryc. 1. Dendrogram grupowania cech dla danych z tablicy 1**

Zgodnie z uzyskanym dendrogramem białko i tłuszcz są cechami sobie najbliższymi (odległość 10.98), wiążącymi się następnie najsilniej z zawartością cukru mlekowego (odległość 23.85) i wreszcie z zawartością wody (odległość 254.88). Otrzymane związki oraz wymienione odległości decydują o strukturze drzewa, które jest grafem złożonym z węzłów odpowiadających poszczególnym cechom oraz z węzłów i krawędzi wynikających z dendrogramu. Schematyczne drzewo charakteryzujące procentowy skład mleka wieloryba przedstawia rycina 2.




Ryc. 2. Drzewo dla procentowego składu mleka wieloryba

Szczegółowe zasady rysowania drzewa są następujące:

(A) Grubość każdej krawędzi drzewa, mierzona w poziomie, jest proporcjonalna do liczby cech znajdujących się na gałęzi tą krawędzią zapoczątkowanej, przy czym grubość podstawy drzewa jest proporcjonalna do liczby wszystkich obserwowanych cech.

(B) Długość każdej krawędzi drzewa jest proporcjonalna do średniej arytmetycznej wartości cech znajdujących się na gałęzi tą krawędzią zapoczątkowanej.

(C) Kąt k pomiędzy dwiema krawędziami wychodzącymi z tego samego węzła jest liniową funkcją logarytmu odległości $d_*$ między grupami zmiennych znajdujących się na gałęziach tymi krawędziami zapoczątkowanych, przy czym funkcja ta jest tak ustalona, aby odległość największa D oraz najmniejsza d odpowiadały ustalonym z góry wartościom kąta maksymalnego M i minimalnego m. Kąt ten wyznacza wzór

$$k = \frac{m\,(\ln D - \ln d_*) + M\,(\ln d_* - \ln d)}{\ln D - \ln d}.$$

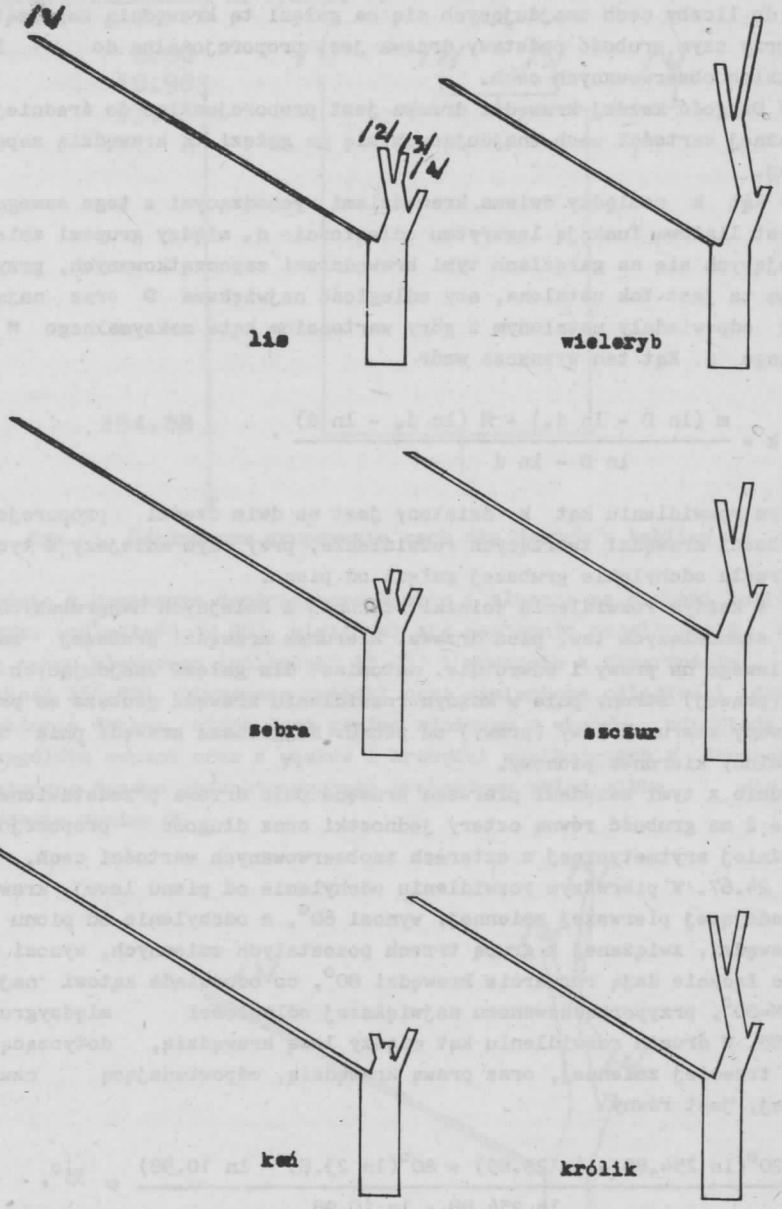
W każdym rozwidleniu kąt k dzielony jest na dwie części proporcjonalne do grubości krawędzi tworzących rozwidlenie, przy czym mniejszy z tych kątów określa odchylenie grubszej gałęzi od pionu.

(D) W każdym rozwidleniu ścieżki złożonej z kolejnych najgrubszych krawędzi, stanowiących tzw. pień drzewa, kierunek krawędzi grubszej zmienia się z lewego na prawy i odwrotnie, natomiast dla gałęzi znajdujących się z lewej (prawej) strony pnia w każdym rozwidleniu krawędź grubsza ma przyporządkowany kierunek lewy (prawy) od pionu. Najgrubsza krawędź pnia drzewa ma ustalony kierunek pionowy.

Zgodnie z tymi zasadami pierwsza krawędź pnia drzewa przedstawionego na rycinie 2 ma grubość równą cztery jednostki oraz długość proporcjonalną do średniej arytmetycznej z czterech zaobserwowanych wartości cech, która wynosi 24.67. W pierwszym rozwidleniu odchylenie od pionu lewej krawędzi, odpowiadającej pierwszej zmiennej, wynosi $60^\circ$, a odchylenie od pionu prawej krawędzi, związanej z grupą trzech pozostałych zmiennych, wynosi $20^\circ$. Kąty te łącznie dają rozwarcie krawędzi $80^\circ$, co odpowiada kątowi największemu $M=80^\circ$, przyporządkowanemu największej odległości międzygrupowej D=254.88. W drugim rozwidleniu kąt między lewą krawędzią, dotyczącą drugiej i trzeciej zmiennej, oraz prawą krawędzią, odpowiadającą czwartej zmiennej, jest równy

$$\frac{20^\circ(\ln 254.88 - \ln 23.85) + 80^\circ(\ln 23.85 - \ln 10.98)}{\ln 254.88 - \ln 10.98} \approx 35^\circ,$$

gdzie kąt najmniejszy wynosi $m=20^\circ$, a $d_* = 23.85$. Ostatnie rozwidlenie związane jest z odległością minimalną d = 10.98, a więc kąt między krawędzią prawą i lewą w tym rozwidleniu wynosi $m=20^\circ$. Zwróćmy jeszcze uwagę na zmianę kierunku w każdym rozwidleniu pnia (prosty, prawy, lewy) oraz na




Ryc. 3. Drzewa dla procentowego składu mleka dla wybranych ssaków

fakt, że krawędzie końcowe mają jednakowe grubości oraz długości proporcjonalne do odpowiednich wartości z pierwszego wiersza tablicy 1.

Komplet drzew dla danych zawartych w tablicy 1 przedstawia rycina 3. Z ich porównania łatwo ustalić podobieństwo procentowego składu mleka wieloryba, szczura i królika. Charakteryzuje się ono niewielką zawartością cukru mlekowego oraz stosunkowo dużą zawartością białka i tłuszczu, przy czym ilość tłuszczu w mleku wieloryba jest największa. Zawartości tych trzech składników w mleku lisa, zebry i konia są bardziej wyrównane, co jest wynikiem zmniejszonej zawartości białka i tłuszczu oraz zwiększonej zawartości wody i cukru mlekowego.

## 3. PRZYKŁADY

W celu zilustrowania przydatności omówionej metody prezentacji danych wielowymiarowych w zagadnieniach klasyfikacyjnych sięgnijmy do klasycznych danych Fishera (1936) dotyczących trzech odmian irysa (iris setosa, iris versi colour, iris virginica). W tablicy 2 przedstawiono po dziesięć pierwszych obserwacji dla każdej odmiany z tego zestawu danych.

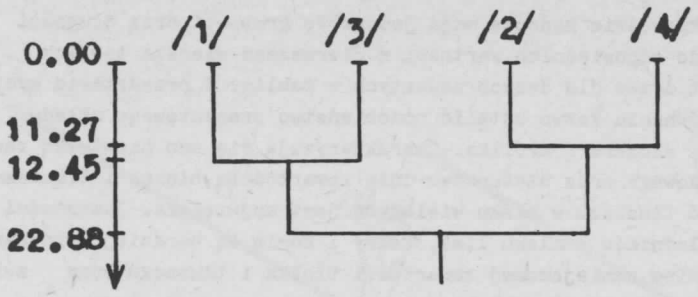
T a b l i c a 2. Pomiary czterech cech dla trzech odmian irysa

| Iris setosa | | | | Iris versicolour | | | | Iris virginica | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (1) | (2) | (3) | (4) | (1) | (2) | (3) | (4) |
| 5.1 | 3.5 | 1.4 | 0.2 | 7.0 | 3.2 | 4.7 | 1.4 | 6.3 | 3.3 | 6.0 | 2.5 |
| 4.9 | 3.0 | 1.4 | 0.2 | 6.4 | 3.2 | 4.5 | 1.5 | 5.8 | 2.7 | 5.1 | 1.9 |
| 4.7 | 3.2 | 1.3 | 0.2 | 6.9 | 3.1 | 4.9 | 1.5 | 7.1 | 3.0 | 5.9 | 2.1 |
| 4.6 | 3.1 | 1.5 | 0.2 | 5.5 | 2.3 | 4.0 | 1.3 | 6.3 | 2.9 | 5.6 | 1.8 |
| 5.0 | 3.6 | 1.4 | 0.2 | 6.5 | 2.8 | 4.6 | 1.5 | 6.5 | 3.0 | 5.8 | 2.2 |
| 5.4 | 3.9 | 1.7 | 0.4 | 5.7 | 2.8 | 4.5 | 1.3 | 7.6 | 3.0 | 6.6 | 2.1 |
| 4.6 | 3.4 | 1.4 | 0.3 | 6.3 | 3.3 | 4.7 | 1.6 | 4.9 | 2.5 | 4.5 | 1.7 |
| 5.0 | 3.4 | 1.5 | 0.2 | 4.9 | 2.4 | 3.3 | 1.0 | 7.3 | 2.9 | 6.3 | 1.8 |
| 4.4 | 2.9 | 1.4 | 0.2 | 6.6 | 2.9 | 4.6 | 1.3 | 6.7 | 2.5 | 5.8 | 1.8 |
| 4.9 | 3.1 | 1.5 | 0.1 | 5.2 | 2.7 | 3.9 | 1.4 | 7.2 | 3.6 | 6.1 | 2.5 |

(1) długość listka, (2) szerokość listka, (3) długość płatka, (4) szerokość płatka.

W rezultacie zastosowania algorytmu Warda hierarchicznego grupowania cech uzyskano dendrogram zobrazowany na rycinie 4.

Dendrogram ten w połączeniu z danymi zawartymi w tablicy 2 prowadzi do 30 schematycznych drzew przedstawionych na rycinie 5, z których każde reprezentuje jeden badany obiekt. Uzyskane rysunki pozwalają łatwo ustalić, że obiekty oznaczone numerami 14, 18, 20 i 27 mogą być mylnie klasyfikowane jako należące do odmiany iris setosa. Ogólnie można stwierdzić, że na podstawie obserwowanych cech rozróżnienie odmian iris versicolour i iris vi-




Ryc.4. Dendrogram grupowania cech dla danych z tablicy 2

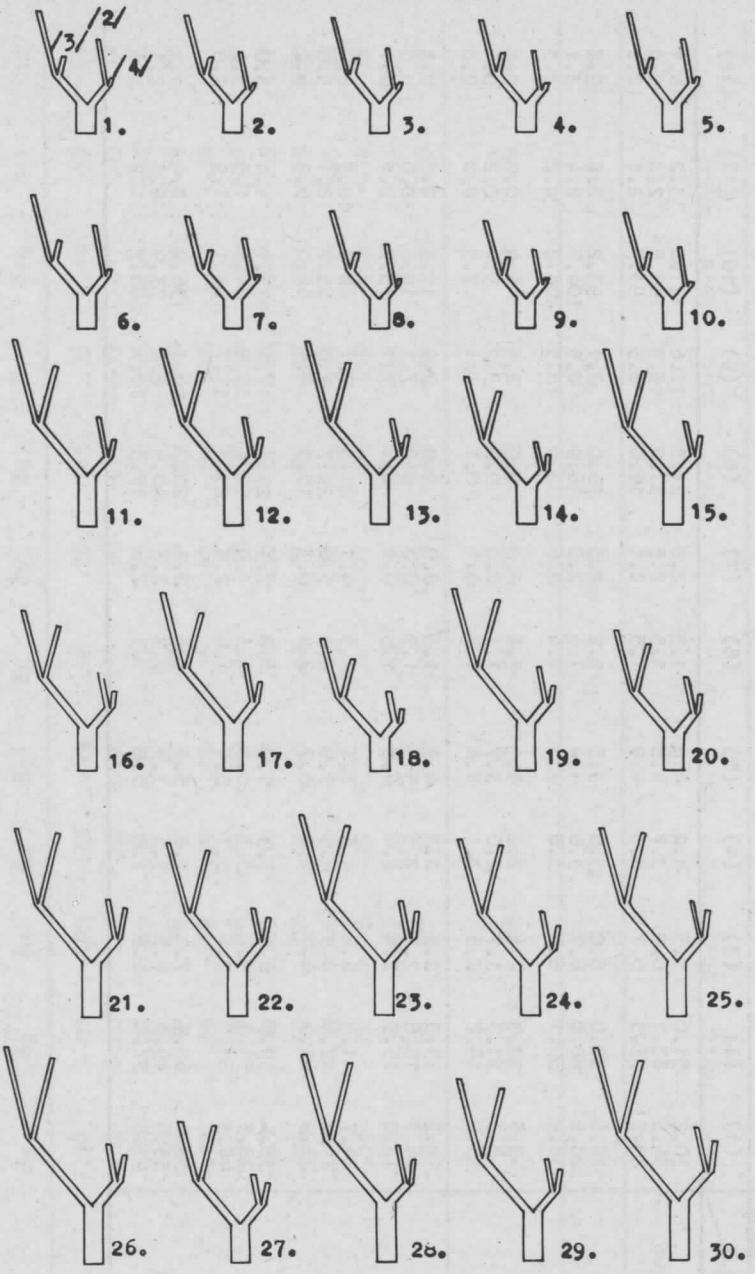
rginica jest zadaniem trudniejszym niż odróżnienie tych odmian od odmiany iris setosa. Ta ostatnia odmiana obejmuje rośliny zdecydowanie mniejsze o bardzo wąskich płatkach i bardzo szerokich listkach.

W omówionych przykładach liczba równocześnie uwzględnianych cech była równa cztery. W przypadku, gdy liczba ta ulega znacznemu powiększeniu, wówczas interpretacja schematycznych drzew poważnie się utrudnia. Wynika to nie tylko z ograniczonych możliwości percepcyjnych oka, ale także stąd, że przy dużej liczbie cech i zróżnicowanych obiektach nie można w zasadzie uniknąć efektu nakładania się gałęzi bądź ich karłowacenia.

Ostatni przykład podyktowany jest chęcią wskazania rozwiązania kompromisowego, w którym prezentacja graficzna dotyczy tylko ograniczonej liczby cech przy zachowaniu jednakże możliwie pełnej informacji o badanych obiektach. Rozwiązanie to polega na zastosowaniu opisanej metody drzew prezentacji danych wielowymiarowych, ale tylko w odniesieniu do cech najistotniejszych, wyboru których dokonuje się przy użyciu analizy składowych głównych.

Proponowane postępowanie zilustrujemy korzystając z obserwacji uzyskanych w doświadczeniu przeprowadzonym w roku 1973 przez Zakład Genetyki Roślin PAN w Poznaniu nad roślinami przelotu (anthyllis vulneraria), pochodzącymi z siedmiu miejscowości usytuowanych w różnych regionach geograficznych Polski. Dla każdej rośliny określono wartości następujących cech: (1) wysokość wzniesienia nad ziemią, (2) długość łodygi głównej,(3) liczba kwiatostanów na łodydze głównej, (4) liczba węzłów liściowych, (5) długość pierwszego liścia łodygi głównej z blaszką, (6) długość ogonka liściowego pierwszego liścia łodygi głównej, (7) liczba odgałęzień pierwszego rzędu łodygi głównej, (8) liczba odgałęzień pierwszego rzędu rośliny, (9) liczba głównych łodyg, (10) liczba kwiatostanów rośliny, (11) liczba kwiatostanów na szypułkach bez węzła, (12) długość szypułki. W tablicy 3 dla każdej miejscowości zestawiono wartości średnie, wyznaczone w oparciu o rośliny stanowiące każdorazowo trzy grupy o zbliżonych liczebnościach (od 18 do 20 roślin w grupie).

W celu wyeliminowania części cech przeprowadzono analizę składowych głównych (patrz Caliński i in., 1975) wyznaczając wektory własne oraz war-




Ryc. 5. Drzewa dla obiektów z tablicy 2

Tablica 3. Średnie obserwacje dwunastu cech dla siedmiu populacji przelotu

| Cechy / Miejscowości | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 30.2 | 51.0 | 5.3 | 4.2 | 4.8 | 1.3 | 2.8 | 21.8 | 11.6 | 121.8 | 3.2 | 4.1 |
| Władysławowo | 34.5 | 63.1 | 6.9 | 3.7 | 4.8 | 1.5 | 2.5 | 31.9 | 13.3 | 168.8 | 2.8 | 2.9 |
| | 23.2 | 57.3 | 6.7 | 3.0 | 4.9 | 1.5 | 2.1 | 16.0 | 8.5 | 93.3 | 2.1 | 2.2 |
| | 29.6 | 39.0 | 6.0 | 3.5 | 3.5 | 1.2 | 2.3 | 15.0 | 9.1 | 93.2 | 2.6 | 2.6 |
| Mielno | 28.6 | 48.0 | 6.7 | 3.8 | 4.4 | 1.4 | 2.8 | 16.9 | 8.5 | 104.7 | 2.4 | 3.1 |
| | 36.4 | 50.1 | 6.7 | 3.8 | 4.5 | 1.4 | 2.6 | 32.6 | 18.5 | 204.2 | 2.2 | 3.3 |
| | 8.9 | 12.9 | 2.4 | 2.0 | 4.1 | 1.4 | 0.0 | 0.0 | 2.5 | 5.4 | 0.0 | 0.0 |
| Kalatówki | 8.3 | 11.3 | 1.7 | 2.0 | 3.7 | 1.1 | 0.0 | 0.0 | 3.2 | 5.4 | 0.2 | 0.4 |
| | 13.2 | 14.7 | 2.5 | 2.1 | 4.4 | 1.3 | 0.1 | 0.1 | 3.4 | 7.5 | 0.0 | 0.0 |
| | 12.2 | 15.3 | 2.5 | 2.3 | 4.2 | 1.3 | 0.0 | 0.0 | 5.8 | 11.3 | 0.2 | 0.1 |
| Skupniów | 14.7 | 18.6 | 2.7 | 2.5 | 5.3 | 1.9 | 0.0 | 0.0 | 7.1 | 14.6 | 0.0 | 0.0 |
| Upłaz | 12.8 | 15.5 | 2.4 | 2.1 | 4.5 | 1.5 | 0.1 | 0.1 | 7.1 | 13.9 | 0.1 | 0.1 |
| | 24.7 | 41.9 | 4.4 | 4.5 | 7.1 | 2.5 | 4.1 | 21.6 | 7.1 | 68.5 | 10.5 | 6.9 |
| Łężyce | 21.8 | 42.0 | 4.9 | 4.5 | 7.0 | 2.5 | 5.0 | 20.7 | 4.8 | 85.7 | 8.1 | 5.6 |
| | 23.9 | 41.9 | 4.7 | 4.8 | 6.7 | 2.4 | 4.3 | 15.5 | 4.1 | 64.0 | 5.2 | 4.2 |
| | 38.4 | 59.5 | 6.6 | 3.3 | 5.9 | 1.8 | 2.5 | 27.1 | 17.1 | 175.1 | 1.4 | 3.0 |
| Rożnowo | 28.5 | 51.5 | 5.7 | 3.3 | 4.6 | 1.3 | 1.9 | 12.7 | 9.8 | 84.5 | 1.5 | 3.1 |
| | 44.3 | 63.7 | 7.1 | 3.7 | 5.4 | 1.6 | 2.3 | 21.3 | 17.7 | 165.1 | 1.9 | 2.8 |
| | 39.8 | 60.9 | 7.1 | 3.8 | 5.1 | 1.5 | 2.3 | 20.2 | 14.1 | 136.9 | 2.4 | 3.5 |
| Międzychód | 33.6 | 49.3 | 6.5 | 3.5 | 3.9 | 0.9 | 2.3 | 10.8 | 8.2 | 73.9 | 0.7 | 2.1 |
| | 41.8 | 57.2 | 6.6 | 3.8 | 5.2 | 1.5 | 2.6 | 17.3 | 10.7 | 108.7 | 1.2 | 2.6 |

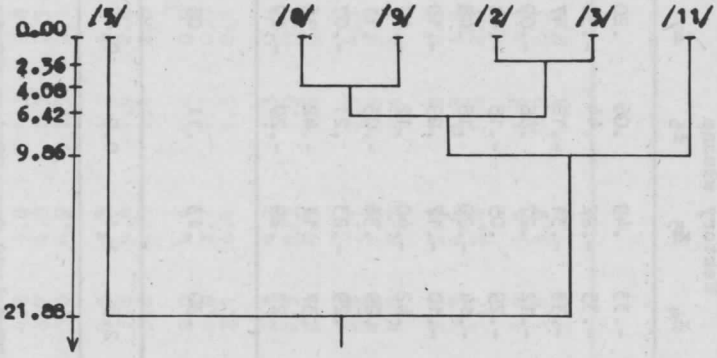
T a b l i c a 4. Wektory własne oraz wartości własne danych z tablicy 3

| Zmienne | Wektory własne | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\underline{x}_1$ | $\underline{x}_2$ | $\underline{x}_3$ | $\underline{x}_4$ | $\underline{x}_5$ | $\underline{x}_6$ | $\underline{x}_7$ | $\underline{x}_8$ | $\underline{x}_9$ | $\underline{x}_{10}$ | $\underline{x}_{11}$ | $\underline{x}_{12}$ |
| (1) | -.30 | -.27 | -.05 | -.33 | .49 | .06 | .20 | .65 | .05 | .12 | .05 | -.01 |
| (2) | -.32 | -.20 | .04 | -.35 | -.22 | -.44 | -.30 | -.14 | -.39 | .17 | .43 | -.14 |
| (3) | -.30 | -.26 | .17 | -.33 | -.31 | -.18 | .47 | -.23 | .16 | -.21 | -.46 | .14 |
| (4) | -.33 | .14 | .28 | -.12 | .27 | .56 | -.09 | -.29 | -.49 | -.18 | -.14 | -.08 |
| (5) | -.21 | .39 | -.49 | -.29 | .05 | -.15 | -.44 | .07 | .16 | -.38 | -.28 | .01 |
| (6) | -.16 | .46 | -.46 | -.14 | -.20 | .18 | .48 | -.07 | -.15 | .44 | .11 | .00 |
| (7) | -.33 | .19 | .29 | -.10 | -.17 | .29 | -.10 | -.02 | .60 | -.07 | .49 | .17 |
| (8) | -.34 | -.10 | -.07 | .42 | -.40 | .12 | -.17 | .39 | -.28 | -.01 | -.12 | .50 |
| (9) | -.24 | -.35 | -.46 | .29 | .38 | -.03 | .11 | -.46 | .09 | -.14 | .25 | .26 |
| (10) | -.31 | -.26 | -.19 | .29 | -.23 | .21 | -.07 | .04 | .19 | .05 | -.12 | -.75 |
| (11) | -.24 | .39 | .19 | .37 | .11 | -.42 | .36 | .16 | -.13 | -.41 | .19 | -.22 |
| (12) | -.32 | .21 | .25 | .23 | .29 | -.29 | -.19 | -.16 | .20 | .59 | -.35 | .07 |
| Wartości własne $\lambda_i$ | 7.61 | 3.15 | .59 | .30 | .13 | .11 | .05 | .03 | .02 | .01 | .01 | .00 |
| $\lambda_i$ (%) | 63.4 | 26.3 | 4.9 | 2.5 | 1.1 | 0.9 | 0.4 | 0.2 | 0.2 | 0.1 | 0.1 | 0.0 |
| Skumulowane $\lambda_i$ (%) | 63.4 | 89.7 | 94.6 | 97.1 | 98.2 | 99.1 | 99.5 | 99.7 | 99.8 | 99.9 | 100.0 | |

tości własne próbkowej macierzy korelacji. Wyniki analizy przedstawia tablica 4.

Biorąc pod uwagę skumulowany procentowy udział kolejnych składowych głównych można ustalić, że zredukowanie problemu do sześciu wymiarów (p=6) pozwala zachować o badanych roślinach ponad 99% informacji wyrażonej sześciu pierwszymi składowymi głównymi. Eliminację cech przeprowadzono kierując się zasadą (patrz Mardia i in., 1979, s.242) usuwania tych cech, których udział mierzony bezwzględną wartością współczynnika w wektorze własnym w kolejnych składowych głównych, poczynając od ostatniej, był największy. W rezultacie ze zbioru cech usunięto kolejne zmienne: (10), (7),(12), (4), (1) oraz (6). Należy w tym miejscu zaznaczyć, że użyta tu procedura eliminacji cech jest jedną z wielu możliwych, przy czym nie wszystkie one muszą koniecznie eliminować te same cechy.

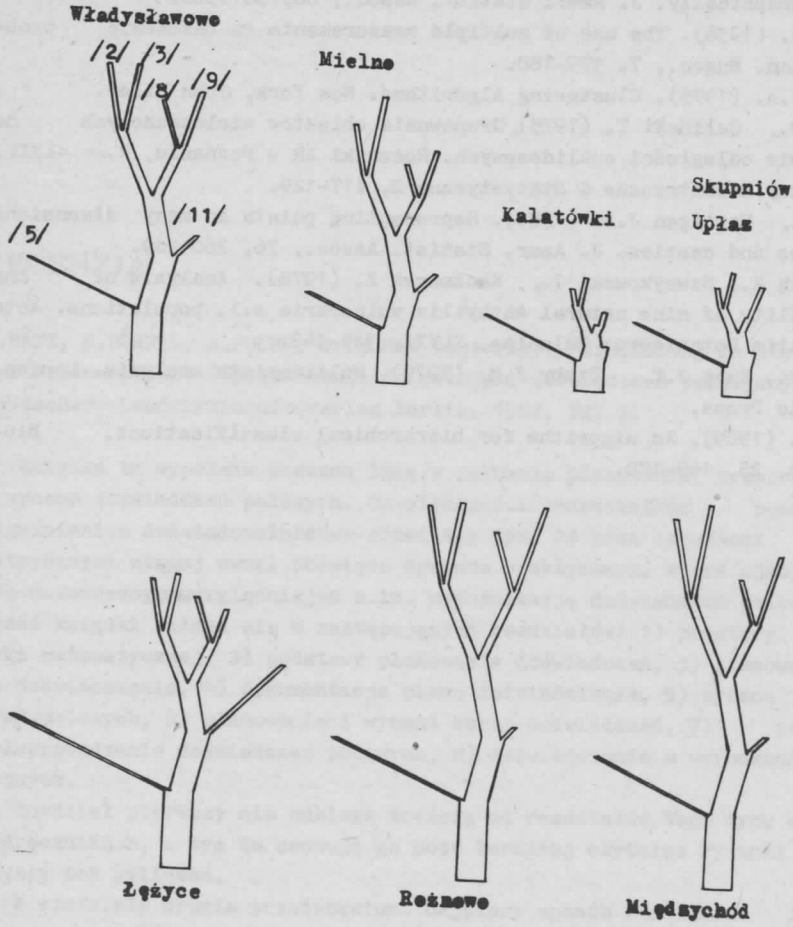
Dla pozostałych zmiennych, poddanych uprzednio standaryzacji w celu zmniejszenia zbytnich dysproporcji, przeprowadzono grupowanie cech metodą Warda, uzyskując dendrogram przedstawiony na rycinie 6.




Ryc. 6. Dendrogram grupowania standaryzowanych cech pozostałych po eliminacji dla danych z tablicy 3

Następnie dla każdej miejscowości wyznaczono po jednym drzewie reprezentującym średni badany obiekt. Komplet siedmiu drzew umieszczono na rycinie 7.

Porównanie schematycznych drzew pozwala na bezpośrednie potwierdzenie wniosków sformułowanych w odniesieniu do anthyllis vulneraria przez Łukaszewską i in., (1978), a sprowadzających się do ustalenia podobieństw populacji pochodzących z tych samych regionów geograficznych Polski, tj. z regionu Wybrzeża (Władysławowo, Mielno), z regionu Tatr (Kalatówki,Skupniów Upłaz) i z regionu Wielkopolski (Rożnowo, Międzychód), oraz na stwierdzeniu znaczących różnic pomiędzy roślinami z różnych regionów geograficznych, w tym również odrębności regionu Sudetów (Łężyce) od pozostałych.




Ryc. 7. Drzewo dla siedmiu populacji przelotu opisanych sześcioma wybranymi cechami

Autorzy pragną podziękować prof. J. Szweykowskiemu oraz mgr K.Łukaszewskiej za udostępnienie danych eksperymentalnych wykorzystanych w rozdziale 3.

LITERATURA

Caliński T., Czajka S., Kaczmarek Z. (1975). Analiza składowych głównych i jej zastosowanie. Roczniki AR w Poznaniu, R. LXXX. Algorytmy Biometryczne i Statystyczne 4, 159-186.

Chernoff H. (1973). The use of faces to represent points in k-dimensional space graphically. J. Amer. Statist. Assoc., 68, 361-368.

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

Hartigan J.A. (1975). Clustering Algorithms. New York, John Wiley.

Karoński M., Caliński T. (1973). Grupowanie obiektów wielocechowych na podstawie odległości euklidesowych. Roczniki AR w Poznaniu, R. LXIV. Algorytmy Biometryczne i Statystyczne 2, 117-129.

Kleiner B., Hartigan J.A. (1981). Representing points in many dimensions by trees and castles. J. Amer. Statist. Assoc., 76, 260-269.

Łukaszewska K., Szweykowski J., Kaczmarek Z. (1978). Analysis of the variability of nine natural Anthyllis vulneraria s.l. populations. Acta Societatis Botanicorum Poloniae, XLVII, 325-342.

Mardia K.V., Kent J.T., Bibby J.M. (1979). Multivariate analysis. London, Academic Press.

Wishart D. (1969). An algorithm for hierarchical classifications. Biometrics. 25, 165-170.