Marek Walesiak

Katedra Ekonometrii i Informatyki
Akademia Ekonomiczna im. Oskara Langego
we Wrocławiu

REKOMENDACJE W ZAKRESIE STRATEGII POSTĘPOWANIA W PROCESIE KLASYFIKACJI ZBIORU OBIEKTÓW*

1. Podstawowe problemy zagadnienia klasyfikacji

Według najogólniejszej koncepcji klasyfikacja jest zbiorem klas odpowiednio wyróżnionym z klasyfikowanego zbioru obiektów. Zawężone sformułowanie zagadnienia klasyfikacji zbioru A o elementach A_i ($i = 1,..., n$) na klasy $P_1,..., P_u$ spełnia warunki: zupełności, rozłączności i niepustości. Można wyróżnić cztery podstawowe problemy decydujące o skali trudności w zakresie klasyfikacji (por. [Milligan 1996, s. 343–344], [Gordon 1999, s. 3–4]).

Pierwszy problem dotyczy liczby klasyfikowanych obiektów. Liczbę wszystkich podziałów zbioru n obiektów na u niepustych klas wyznacza się ze wzoru (zob. [Everitt, Landau, Leese 2001, s. 99], [Gordon 1999, s. 40]]:

$$L(n,u) = \frac{1}{u!}\sum_{s=1}^{u}(-1)^{u-s}\binom{u}{s}s^n. \tag{1}$$

Przykładowe liczby wszystkich możliwych podziałów zbioru n obiektów na u niepustych klas są następujące:

$$L(5,3) = 25;\ L(10,4) = 34\ 105;\ L(50,5) > 7{,}4 \cdot 10^{32};\ L(100,6) > 9 \cdot 10^{74}.$$

Zatem liczbę wszystkich podziałów zbioru obiektów wyznacza się ze wzoru:

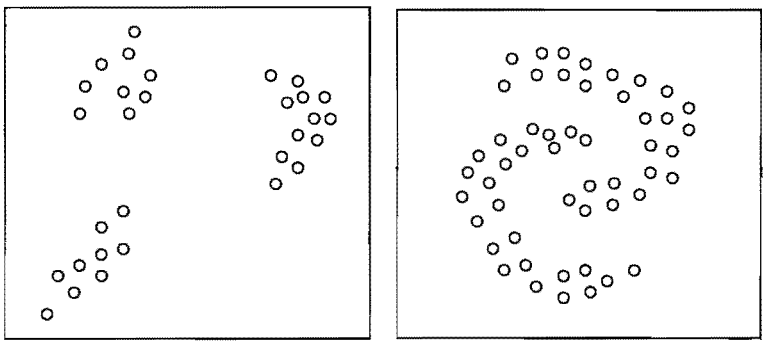
$$L(n) = L(n,1) + L(n,2) + ... + L(n,n). \tag{2}$$

---

* Artykuł stanowi zasadniczo zmodyfikowaną (tzn. poprawioną i uzupełnioną) wersję opracowania [Walesiak 2004].

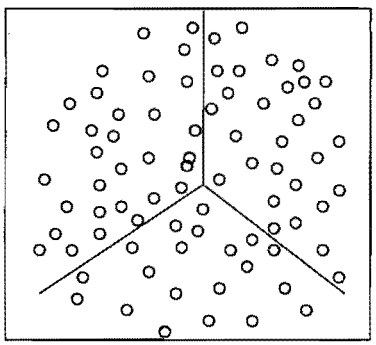
Przykładowe liczby wszystkich możliwych podziałów zbioru n obiektów są następujące:

$$L(5) = 52;\ L(10) = 115\,975;\ L(50) > 1{,}8 \cdot 10^{47};\ L(100) > 4{,}7 \cdot 10^{115}. \quad (2)$$

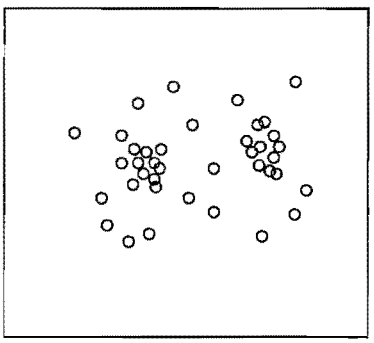


a) klasy spójne i izolowane

b) klasy izolowane, ale niespójne



c) klasy spójne, ale nie izolowane



d) obszary o większej zwartości przestrzennej obiektów oddzielone obszarami o mniejszej zwartości przestrzennej

Rys. 1. Ilustracja koncepcji wewnętrznej spójności i zewnętrznej izolacji klas

Źródło: opracowanie własne na podstawie pracy: [Gordon 1999, s. 4].

Wraz ze wzrostem liczby klasyfikowanych obiektów liczba możliwych podziałów zbioru n obiektów staje się ogromna. Zatem rozpatrzenie wszystkich możliwych podziałów zbioru n obiektów, z punktu widzenia pewnego kryterium, i wybór na tej podstawie podziału najlepszego nie jest wykonalne dla większych liczebności zbioru obiektów. W tej sytuacji niezbędne są efektywne metody i algorytmy klasyfikacji.

Drugim elementem decydującym o skali trudności problemu klasyfikacji jest liczba zmiennych opisujących badane obiekty. Dla jednej zmiennej uzy-

skujemy uporządkowanie obiektów na prostej, a dla dwóch zmiennych rozmieszczenie obiektów na płaszczyźnie. W zasadzie dla tych dwóch przypadków istnieje możliwość wizualizacji rozmieszczenia obiektów. Uwzględnienie więcej niż trzech zmiennych powoduje, że do rozwiązania problemu klasyfikacji niezbędne stają się odpowiednie metody i algorytmy klasyfikacji.

Trzeci problem dotyczy rozmieszczenia obiektów w przestrzeni klasyfikacji i braku powszechnie akceptowanej definicji klasy. W literaturze przedmiotu wypracowano wiele definicji klas (por. m.in. [Everitt 1974, s. 43-48], [Dąbrowski, Laus-Mączyńska 1978, s. 62–66]), które znajdują zastosowanie w specyficznych przypadkach. Głównym celem klasyfikacji jest badanie podobieństwa lub odrębności obiektów i ich zbiorów. Celem tym jest więc podział zbioru obiektów na klasy zawierające obiekty podobne ze względu na obserwacje na zmiennych (tzw. klasy względnie jednorodne). Ponadto obiekty znajdujące się w różnych klasach powinny być jak najmniej podobne. Postuluje się, aby wyodrębnione klasy spełniały dwa kryteria: wewnętrznej spójności i zewnętrznej izolacji (por. [Gordon 1999, s. 3]). Analiza rys. 1 pokazuje, że wyodrębnienie w zbiorze obiektów właściwej struktury klas nie jest zadaniem trywialnym (dodatkowo komplikuje się w wypadku większej liczby zmiennych).

Czwarty problem decydujący o skali trudności problemu klasyfikacji dotyczy braku szeroko akceptowanej ujednoliconej teorii klasyfikacji.

2. Etapy występujące w typowym studium klasyfikacyjnym

W literaturze przedmiotu wyodrębnia się osiem etapów procesu klasyfikacji (por. np. [Milligan 1996, s. 342–343], [Walesiak 2004]): wybór obiektów do klasyfikacji, wybór zmiennych charakteryzujących obiekty, wybór formuły normalizacji wartości zmiennych, wybór miary odległości, wybór metody klasyfikacji, ustalenie liczby klas, walidacja wyników klasyfikacji, opis (interpretacja) i profilowanie klas. W kolejnych częściach, na podstawie światowej literatury klasyfikacyjnej oraz własnych doświadczeń, przedstawiono rekomendowane strategie postępowania.

3. Wybór obiektów do klasyfikacji

Należy odpowiedzieć na pytanie, czy badaniem objąć całą populację, czy tylko jej próbkę? Jeśli zdecydowano się na badanie próbkowe (z takimi badaniami mamy zazwyczaj w analizach marketingowych), to należy określić elementarną jednostkę badania, wybrać metodę doboru próby i określić jej liczebność. W każdym badaniu statystycznym, w tym również w niewyczerpującym badaniu wielowymiarowym, można przyjąć jedno z dwóch podejść: stocha-

styczne lub opisowe. W podejściu stochastycznym zakłada się, że zbiór obserwacji (obiektów) stanowi próbę losową pochodzącą z populacji (o nieskończonej lub skończonej liczebności). Podejście stochastyczne, w którym rozpatrywane zmienne są losowe, wolno przyjąć przede wszystkim w wypadku badań eksperymentalnych, tzn. gdy istnieje możliwość powtórzenia badania w takich samych warunkach. Wtedy zbiór obserwacji może być traktowany jako próba losowa. W podejściu opisowym zmienne nie są losowe, lecz są zmiennymi w zwykłym sensie. Badaniu nie podlegają wtedy właściwości stochastyczne zbioru obserwacji. Podejście opisowe przyjmuje się z reguły wtedy, gdy dane pochodzą ze sprawozdawczości statystycznej.

Dobór próby powinno się tak przeprowadzić, aby klasy wyodrębnione na jej podstawie odpowiadały strukturze klas populacji.

4. Wybór zmiennych charakteryzujących obiekty

Dobór zmiennych jest jednym z najważniejszych, a zarazem najtrudniejszych zagadnień. Od jakości zestawu zmiennych zależy bowiem wiarygodność ostatecznych wyników klasyfikacji i trafność podejmowanych na ich podstawie decyzji.

W procedurze klasyfikacji należy uwzględnić tylko te zmienne, które mają zdolność dyskryminacji zbioru obiektów. Podejście polegające na uwzględnieniu jak największej liczby zmiennych jest nieuzasadnione. Dodanie do zbioru jednej lub kilku nieistotnych zmiennych nie pozwala na odkrycie w zbiorze obiektów właściwej struktury klas (zob. [Milligan 1994]).

Do rozwiązania zagadnienia doboru zmiennych służą zasadniczo dwa ujęcia: dobór merytoryczny w ścisłym tego słowa znaczeniu oraz dobór merytoryczno-formalny. Obydwa ujęcia obejmują dwie fazy. Faza I jest taka sama w obydwu ujęciach, różnice zaś występują w fazie II. Punktem wyjścia obydwu ujęć (faza I) jest skonstruowanie wstępnej listy zmiennych na podstawie własnej hipotezy roboczej badacza (wynikającej z jego znajomości przedmiotu badania oraz wiedzy płynącej z szeroko pojętej teorii ekonomii) oraz współpracy z przedstawicielami odpowiednich dyscyplin naukowych (ekspertami).

Redukcja wstępnej listy zmiennych z wykorzystaniem analizy merytorycznej (faza II) jest działaniem w głównej mierze subiektywnym. Dokonuje się jej na podstawie własnej znajomości przedmiotu badania, wykorzystując współpracę przedstawicieli odpowiednich dyscyplin naukowych (ekspertów) oraz opierając się na szeroko pojętej teorii ekonomii.

Redukcja wstępnej listy zmiennych z wykorzystaniem metod doboru zmiennych (faza II) polega na tym, że najpierw usuwa się zmienne, które charakteryzują się małą zmiennością. Następnie do tak zredukowanej liczby zmiennych stosuje się formalny algorytm wyboru zmiennych. W zagadnieniu klasyfikacji

zbioru obiektów celem zastosowania tych algorytmów jest wybór takiego zestawu zmiennych, w którym zmienne są wzajemnie niezależne oraz są zależne od zmiennych nie wchodzących do wybranego zestawu (postulat niepowielania informacji). Podejście oparte na niepowielaniu informacji niekoniecznie prowadzi do właściwych rezultatów. P.H.A. Sneath (zob. [Milligan 1996, s. 348]) pokazał, że redukcja przestrzeni klasyfikacji (w wyniku zastosowania np. analizy głównych składowych) może spowodować utratę struktury klas z pierwotnej przestrzeni. Z tego względu w literaturze proponowane są inne podejścia pozwalające określić zdolność zmiennych do dyskryminacji zbioru obiektów:

– E.B. Fowlkes, R. Gnanadesikan i J.R. Kettenring [1988] zaproponowali procedurę doboru zmiennych, znaną w analizie regresji pod nazwą procedury selekcji „w przód",

– A. Sokołowski [1992, s. 12–13, 50–51] zaproponował miarę zdolności grupowania dla indywidualnych zmiennych i zestawu zmiennych,

– opracowano algorytmy łączące problem doboru zmiennych z doborem wag dla zmiennych (por. np. [Gnanadesikan, Kettenring, Ksao 1995], [Milligan 1989]).

Szerzej o problemach selekcji i ważenia zmiennych w zagadnieniu klasyfikacji traktuje praca M. Walesiaka [2005].

5. Wybór formuły normalizacji wartości zmiennych

Celem normalizacji wartości zmiennych jest pozbawienie mian wyników pomiaru oraz ujednolicenie ich rzędów wielkości. Normalizację przeprowadza się, gdy zmienne opisujące obiekty badania mierzone są na skali przedziałowej i (lub) ilorazowej. Z uwagi na to, że jedynymi dopuszczalnymi przekształceniami na skali przedziałowej i ilorazowej są przekształcenia liniowe, formuły normalizacyjne można wyrazić ogólnym wzorem:

$$z_{ij} = bx_{ij} + a(b > 0), \tag{3}$$

gdzie:

x_{ij} (z_{ij}) – wartość (znormalizowana wartość) j-tej zmiennej w i-tym obiekcie.

Formuły normalizacyjne oraz charakterystyki rozkładu wartości zmiennych po normalizacji zawiera tabela 1. Przy wyborze formuły normalizacyjnej należy brać pod uwagę:

a) skale pomiaru zmiennych:

– przekształcenia ilorazowe (formuły 6–11 z tabeli 1) można stosować tylko wtedy, gdy zmienne są mierzone na skali ilorazowej (istnieje dla niej absolutny punkt zerowy);

– pozostałe formuły normalizacyjne (1–5 z tabeli 1) stosuje się, gdy zbiór zawiera zmienne mierzone na skali przedziałowej lub ilorazowej. Formuły te

Tabela 1

Formuły normalizacyjne oraz charakterystyki rozkładu wartości zmiennych po normalizacji

Lp.	Formuła	Średnia arytmetyczna[a]	Odchylenie standardowe[a]	Rozstęp
1	$z_{ij}=\frac{(x_{ij}-\bar{x}_j)}{s_j}$	0	1	$\frac{r_j}{s_j}$
2	$z_{ij}=\frac{(x_{ij}-Me_j)}{1{,}4826\cdot MAD_j}$	0	1	$\frac{r_j}{1{,}4826\cdot MAD_j}$
3	$z_{ij}=\frac{(x_{ij}-\bar{x}_j)}{r_j}$	0	$\frac{s_j}{r_j}$	1
4	$z_{ij}=\frac{\left[x_{ij}-\min\limits_i\{x_{ij}\}\right]}{r_j}$	$\frac{\left[\bar{x}_j-\min\limits_i\{x_{ij}\}\right]}{r_j}$	$\frac{s_j}{r_j}$	1
5	$z_{ij}=\frac{(x_{ij}-\bar{x}_j)}{\max\limits_i\lvert x_{ij}-\bar{x}_j\rvert}$	0	$\frac{s_j}{\max\limits_i\lvert x_{ij}-\bar{x}_j\rvert}$	$\frac{r_j}{\max\limits_i\lvert x_{ij}-\bar{x}_j\rvert}$
6	$z_{ij}=\frac{x_{ij}}{s_j}$	$\frac{\bar{x}_j}{s_j}$	1	$\frac{r_j}{s_j}$
7	$z_{ij}=\frac{x_{ij}}{r_j}$	$\frac{\bar{x}_j}{r_j}$	$\frac{s_j}{r_j}$	1
8	$z_{ij}=\frac{x_{ij}}{\max\limits_i\{x_{ij}\}}$	$\frac{\bar{x}_j}{\max\limits_i\{x_{ij}\}}$	$\frac{s_j}{\max\limits_i\{x_{ij}\}}$	$\frac{r_j}{\max\limits_i\{x_{ij}\}}$
9	$z_{ij}=\frac{x_{ij}}{\bar{x}_j}$	1	$\frac{s_j}{\bar{x}_j}$	$\frac{r_j}{\bar{x}_j}$
10	$z_{ij}=\frac{x_{ij}}{\sum_{i=1}^{n}x_{ij}}$	$\frac{1}{n}$	$\frac{s_j}{\sum_{i=1}^{n}x_{ij}}$	$\frac{r_j}{\sum_{i=1}^{n}x_{ij}}$
11	$z_{ij}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}}$	$\frac{\bar{x}_j}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}}$	$\frac{s_j}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}}$	$\frac{r_j}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}}$

[a] dla standaryzacji Webera: mediana i medianowe odchylenie bezwzględne, 1 – standaryzacja klasyczna, 2 – standaryzacja Webera [Lira, Wagner, Wysocki 2002, s. 91], 3 – unitaryzacja, 4 – unitaryzacja zerowana, 5 – normalizacja [Rybaczuk 2002, s. 147] w przedziale [–1; 1], 6–11 – przekształcenia ilorazowe; $\bar{x}_j, s_j, r_j$ – średnia arytmetyczna, odchylenie standardowe, rozstęp dla j-tej zmiennej, Me_j (MAD_j) – mediana (medianowe odchylenie bezwzględne) dla j-tej zmiennej

Źródło: opracowanie własne na podstawie: [Jajuga, Walesiak 2000, s. 109], [Walesiak 2004].

wprowadzają jednolicie określoną wartość zerową (umowną) dla wszystkich zmiennych. Standaryzacja klasyczna (standaryzacja Webera), unitaryzacja, normalizacja w przedziale [–1; 1] określają umowną wartość zerową na poziomie średniej wartości zmiennej (mediany), a unitaryzacja zerowana – na poziomie wartości minimalnej;

– zastosowanie formuł normalizacyjnych 1–5 z tabeli 1 do zmiennych mierzonych na skali ilorazowej, formalnie poprawne, spowoduje stratę informacji wskutek „przejścia" wszystkich zmiennych na skalę przedziałową. Strata informacji przejawia się m.in. ograniczeniem zastosowania różnych technik statystycznych i ekonometrycznych;

b) charakterystyki rozkładu zmiennych z uwagi na to, że normalizacja wprowadza pewien sposób ważenia zmiennych (por. [Milligan, Cooper 1988, s. 182]). Analiza tabeli 1 pozwala sformułować następujące wnioski (zob. [Jajuga, Walesiak 2000, s. 110–111], [Walesiak 2002, s. 20], [Walesiak 2004]):

– formuły normalizacyjne 3, 4 i 7 są cenne, ponieważ zapewniają znormalizowanym wartościom zmiennych zróżnicowaną zmienność (mierzoną odchyleniem standardowym) i jednocześnie stały rozstęp dla wszystkich zmiennych,

– formuły normalizacyjne 1, 2 i 6 powodują ujednolicenie wartości wszystkich zmiennych pod względem zmienności. Oznacza to wyeliminowanie zmienności jako podstawy różnicowania obiektów. Standaryzację Webera należy stosować, gdy rozkład empiryczny badanych zmiennych jest silnie asymetryczny (zob. [Lira, Wagner, Wysocki 2002, s. 91]),

– przekształcenia ilorazowe o postaci 8 i 11 zapewniają znormalizowanym wartościom zmiennych zróżnicowaną zmienność, średnią arytmetyczną i rozstęp,

– przekształcenia ilorazowe o postaci 5, 9 i 10 zapewniają znormalizowanym wartościom zmiennych zróżnicowaną zmienność i rozstęp oraz stałą dla wszystkich zmiennych średnią arytmetyczną. Formuła 10 stanowi podstawę normalizacji w badaniach strukturalnych,

– wszystkie formuły normalizacyjne, będące przekształceniami liniowymi obserwacji na każdej zmiennej, zachowują skośność i kurtozę rozkładu zmiennych. Ponadto dla każdej pary zmiennych nie zmieniają wartości współczynnika korelacji liniowej Pearsona.

6. Wybór miary odległości

Wybór miary odległości zależy od:

– skali pomiaru zmiennych, gdy zmienne są mierzone na tej samej skali pomiaru. W literaturze wypracowano wiele propozycji miar odległości znajdujących zastosowanie do zmiennych mierzonych na skali: ilorazowej, przedziałowej lub ilorazowej, porządkowej, nominalnej (w tym dla zmiennych binarnych). Bardzo dobry przegląd różnych typów miar odległości przedstawiono m.in.

w pracach: [Cormack 1971], [T.F. Cox, M.A.A. Cox 1994, s. 10], [Gordon 1999, s. 20–21], [Anderberg 1973, s. 98–130], [Kaufman, Rousseeuw 1990, s. 4–37], [Walesiak 2002, s. 23–31];

– zastosowanej formuły normalizacji wartości zmiennych (zob. [Walesiak 2002, s. 29]);

– spełniania przez daną formułę dodatkowych własności (np. warunku nierówności trójkąta – miara odległości zwana jest wtedy metryką). Spośród miar odległości obiektów opisanych zmiennymi mierzonymi na skali przedziałowej i (lub) ilorazowej najczęściej wykorzystuje się z tego powodu odległość euklidesową i jej kwadrat;

– skal pomiaru zmiennych, gdy zbiór zmiennych zawiera zmienne mierzone na skalach różnych rodzajów. W tej sytuacji należy wykorzystać miary podobieństwa dopuszczające zmienne mierzone na różnych skalach. W literaturze miary takie zaproponowali autorzy następujących prac: [Gower 1971], [T.F. Cox, M.A.A. Cox 2000], [Bock, Diday 2000, s. 152], [Walesiak 2003]. Inne podejścia do rozwiązania tego problemu omówiono w pracach: [Kaufman, Rousseeuw 1990, s. 32–37], [Kolonko 1979], [Gordon 1981, s. 25–27], [Jajuga 1989], [Walesiak 1993].

7. Wybór metody klasyfikacji

Do rozwiązania problemu wyboru właściwej, dla danego typu danych empirycznych, metody klasyfikacji proponuje się w literaturze przedmiotu cztery zasadnicze podejścia (por. [Gordon 1987; 1996, s. 81–85]).

W pierwszym z nich poprawność poszczególnych metod ocenia się na podstawie zadanych typów struktur danych. Dana metoda klasyfikacji jest poprawna, jeśli wyniki klasyfikacji uzyskane za jej pomocą odpowiadają znanej strukturze danych. Przykłady zastosowania tego typu podejścia można znaleźć m.in. w pracach: [Milligan 1981; 1996, s. 355–361], [Grabiński 1990, 1992] oraz [Grabiński, Wydymus, Zeliaś 1989].

Podstawową wadą tego podejścia jest opieranie się na wygenerowanych strukturach danych, w których konfiguracje obiektów są na ogół przedstawiane w przestrzeniach dwuwymiarowych i trójwymiarowych. Trudno jest więc uogólnić wyniki na przypadek wielowymiarowy. Nawet wtedy, gdy podejście to opiera się na danych symulacyjnych (uzyskanych za pomocą odpowiednio skonstruowanych wielowymiarowych generatorów zmiennych losowych o zadanej postaci rozkładu), trudno jest uogólnić wyniki, ponieważ każda empirycznie uzyskana struktura danych jest inna i tak uzyskane wnioski mają ograniczony zasięg zastosowania.

Na podstawie wielu analiz poprawności odkrywania zadanych typów struktur danych G.W. Milligan [1996, s. 358] wskazał, że najlepsze wśród hierarchicznych metod aglomeracyjnych są metody Warda i giętka (β-*flexible*).

Drugie podejście polega na tym, że do klasyfikacji zbioru obiektów wykorzystuje się różne metody klasyfikacji, a następnie ocenia się zgodność wyników klasyfikacji i wybiera się te metody, które dają zbliżone wyniki. Wyniki klasyfikacji z użyciem tych metod podlegają w dalszej fazie syntetyzacji w celu wyłonienia zgodnej klasyfikacji.

Godne odnotowania propozycje mierników służących do porównywania wyników dwóch różnych podziałów podali: W.M. Rand [1971], L.J. Hubert i P. Arabie [1985], E.B. Fowlkes i C.L. Mallows [1983], J.C. Lerman [1988], L.A. Goodman i W.H. Kruskal [1979], D.L. Wallace [1983]. W literaturze polskiej propozycje takie przedstawili: E. Nowak [1985], C. Szmigiel [1976] i A. Sokołowski [1976].

Dany jest niepusty zbiór obiektów badania A o elementach A_i ($i = 1,..., n$) oraz dwa podziały tego zbioru na u i v klas otrzymane na podstawie jednolitej procedury klasyfikacyjnej: $P^{(q)} = \{P_1^{(q)}, ..., P_u^{(q)}\}$; $P^{(t)} = \{P_1^{(t)}, ..., P_v^{(t)}\}$. W literaturze najczęściej do oceny podobieństwa wyników klasyfikacji zbioru obiektów wykorzystywana jest miara Randa [1971]. W jej koncepcji porównuje się zaklasyfikowanie wszystkich par obiektów w podziałach $P^{(t)}$ i $P^{(q)}$ wyróżniając cztery typy par obiektów:

I: obiekty tworzące parę znajdują się w tych samych klasach w podziałach $P^{(t)}$ i $P^{(q)}$,

II: obiekty tworzące parę znajdują się w różnych klasach w podziałach $P^{(t)}$ i $P^{(q)}$,

III (IV): obiekty tworzące parę znajdują się w różnych klasach (w tej samej klasie) w $P^{(q)}$ i w tej samej klasie (różnych klasach) w $P^{(t)}$.

Typy (I) i (II) są interpretowane jako pary zgodne (Z) w obu klasyfikacjach $P^{(t)}$ i $P^{(q)}$, natomiast typy (III) i (IV) jako pary niezgodne (N). Intuicyjnie widać więc, że podobieństwo dwóch podziałów wzrasta w miarę wzrostu wartości Z. Na tej podstawie W.M. Rand [1971] skonstruował miarę pozwalającą oceniać podobieństwo wyników dwóch podziałów:

$$R = \frac{Z}{\binom{n}{2}}, \tag{4}$$

gdzie:

$$Z = \binom{n}{2} + \sum_{s=1}^{u} \sum_{r=1}^{v} n_{sr}^2 - \frac{1}{2}\left(\sum_{s=1}^{u} n_{s\bullet}^2 + \sum_{r=1}^{v} n_{\bullet r}^2 \right),$$

n_{sr} – liczba obiektów, które jednocześnie należą do klas $P_s^{(q)}$ i $P_r^{(t)}$,

$n_{\bullet r}$ – liczba obiektów w klasie $P_r^{(t)}$, $n_{s\bullet}$ – liczba obiektów w klasie $P_s^{(q)}$,

$r = 1, ..., v$; $s = 1, ..., u$; $v(u)$ – liczba klas w podziale $P^{(t)}(P^{(q)})$.

Przedział zmienności tej miary zaczyna się od 0, kiedy to dwa podziały $P^{(i)}$ i $P^{(q)}$ są zupełnie niepodobne (jeden podział zawiera tyle klas, ile jest obiektów, a drugi jedną klasę zawierającą wszystkie obiekty), a kończy na 1, kiedy podziały są identyczne. Miarę Randa (4) interpretuje się jako odsetek par obiektów zgodnych w obu klasyfikacjach $P^{(i)}$ i $P^{(q)}$ w ogólnej liczbie par obiektów określonych na zbiorze A.

Wadą miary Randa jest to, że wykazuje tendencję do wzrostu wartości w wypadku zwiększania liczby klas (por. [Everitt, Landau, Leese 2001, s. 182]). L.J. Hubert i P. Arabie [1985, s. 198] zaproponowali skorygowany indeks Randa:

$$R_{HA} = \frac{R - E(R)}{R_{\max} - E(R)}, \tag{5}$$

gdzie:

$R_{\max}$ – maksymalna wartość miary Randa ($R_{\max} = 1$),

$E(R)$ – wartość oczekiwana miary Randa określona wzorem:

$$E(R) = 1 + \frac{2\sum_r \binom{n_{\bullet r}}{2} \sum_s \binom{n_{s\bullet}}{2}}{\binom{n}{2}^2} - \frac{\sum_r \binom{n_{\bullet r}}{2} + \sum_s \binom{n_{s\bullet}}{2}}{\binom{n}{2}}.$$

Skorygowana miara Randa przyjmuje postać [Hubert, Arabie 1985, s. 198]:

$$R_{HA} = \frac{\sum_{r,s} \binom{n_{rs}}{2} - \frac{\sum_r \binom{n_{\bullet r}}{2} \sum_s \binom{n_{s\bullet}}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_r \binom{n_{\bullet r}}{2} + \sum_s \binom{n_{s\bullet}}{2}\right] - \frac{\sum_r \binom{n_{\bullet r}}{2} \sum_s \binom{n_{s\bullet}}{2}}{\binom{n}{2}}}. \tag{6}$$

Górną granicą miary R_{HA} jest 1. Miara $R_{HA} = 0$, gdy indeks Randa równy jest jego wartości oczekiwanej. Wartość oczekiwana skorygowanej miary Randa wynosi zero.

W trzecim podejściu (stosowanym dla metod klasyfikacji hierarchicznej) za właściwą, dla danego typu danych, metodę klasyfikacji należy uznać taką, która daje minimalne zniekształcenia przy transformacji wyjściowej macierzy odległo-

ści $[d_{ik}]$ w macierz wartości kofenetycznych $[h_{ik}]$ (inaczej wartości poziomu połączenia klas w dendrogramie). Wartości h_{ik} (dla każdego i, k) w macierzy $[h_{ik}]$ odczytuje się z dendrogramu, który wskazuje wartości poziomu połączenia klas P_i oraz P_k. W tabeli 2 przedstawiono trzy mierniki pomiaru zniekształcenia przy transformacji $[d_{ik}] \rightarrow [h_{ik}]$. Małe wartości D.2 i D.3 oraz duże wartości D.1 oznaczają małe zniekształcenia przy transformacji $[d_{ik}] \rightarrow [h_{ik}]$ przez daną metodę klasyfikacji.

Pewną słabą stroną oparcia się w wyborze właściwej (w odniesieniu do danego typu danych) metody klasyfikacji na miarach tego typu jest to, że na ogół metodą wybieraną przez D.1 jest metoda średniej klasowej [Sokal, Rohlf 1962], [Sneath 1969], a przez D.3 – metoda pojedynczego połączenia [Gordon 1987].

Tabela 2

Miary zniekształcenia przy transformacji $[d_{ik}] \rightarrow [h_{ik}]$

Lp.	Nazwa	Miara	Źródło
D.1	Współczynnik korelacji kofenetycznej	$\frac{\sum_{i,k}(d_{ik}-\bar{d})(h_{ik}-\bar{h})}{\left[\sum_{i,k}(d_{ik}-\bar{d})^2 \sum_{i,k}(h_{ik}-\bar{h})^2\right]^{0,5}}$	[Sokal, Rohlf 1962]
D.2	Suma kwadratów odchyleń	$\sum_{i,k} w_{ik}(d_{ik}-h_{ik})^2$	[Hartigan 1967]
D.3	Metryka Minkowskiego	$\begin{cases}\left[\sum_{i,k}\lvert d_{ik}-h_{ik}\rvert^{\frac{1}{\lambda}}\right]^{\lambda} & (0<\lambda\leq 1)\\ \max_{i,k}\{\lvert d_{ik}-h_{ik}\rvert\} & (\lambda=0)\end{cases}$	[Jardine, Sibson 1971]

w_{ik} – wagi (na ogół wszystkie odległości są jednakowo ważne, więc $w_{ik} = 1$)

Źródło: opracowanie własne na podstawie: [Gordon 1987; 1996, s. 82], [Cormack 1971].

W czwartym podejściu analizuje się formalne własności metod klasyfikacji, które mogą stanowić pomocne kryterium wyboru właściwej metody. Pierwsze własności formalne wypracowali N. Jardine i R. Sibson [1971]. Zostały one następnie uzupełnione w pracach: [Fisher, Van Ness 1971], [Van Ness 1973], a zwięzły ich przegląd w literaturze zawierają następujące monografie: [Gordon 1981; 1999, s. 98–100], [Pociecha 1982], [Ajvazjan, Bežaeva, Staroverov 1974]. W tabeli 3 w sposób syntetyczny przedstawiono formalne własności hierarchicznych metod aglomeracyjnych. Znajomość określonych własności po-

szczególnych metod klasyfikacji pozwala na właściwe ich wykorzystanie w badaniach empirycznych.

Tabela 3

Formalne własności wybranych hierarchicznych metod aglomeracyjnych

Metoda	Własności					
	A	B	C	D	E	F
Pojedynczego połączenia (*single-link*)	-	+	+	+	+	+
Kompletnego połączenia (*complete-link*)	-	+	+	+	+	+
Średniej klasowej (*group average-link*)	-	+	+	-	-	+
Powiększona suma kwadratów odległości (*incremental sum of squares*)	+	-	+	-	-	+
Środka ciężkości (*centroid*)	-	-	-	-	-	+

A – wypukłości, B – poprawnej struktury według klas, C – poprawnej struktury według drzewka połączeń, D – monotoniczności, E – powtarzania punktów, F – opuszczania klas, „+" – spełnia, „–" – nie spełnia

Źródło: opracowanie własne na podstawie: [Gordon 1981, s. 131], [Pociecha 1982], [Hussain 1982], [Walesiak 1993, s. 59], [Van Ness 1973, s. 424].

8. Ustalenie liczby klas

G.W. Milligan i M.C. Cooper [1985] przetestowali na podstawie zbiorów danych o znanej strukturze klas 30 procedur pozwalających wyznaczyć liczbę klas. Przedstawiony przegląd nie wyczerpuje zbioru istniejących sposobów wyznaczania liczby klas. Inne sposoby zawarte są m.in. w pracach: [Sokołowski 1992], [Walesiak 1988]. Większość metod przedstawionych w pracy [Milligan, Cooper 1985] opierała się, przy wyznaczaniu liczby klas, na wyjściowej macierzy danych. Niektóre z nich wykorzystywały odległości wewnątrzklasowe i międzyklasowe. Trzy najlepsze kryteria ogólne są następujące:

1. Indeks Calińskiego i Harabasza [1974]:

$$G1(u) = \frac{\frac{\mathrm{tr}(\mathbf{B})}{(u-1)}}{\frac{\mathrm{tr}(\mathbf{W})}{(n-u)}}, \tag{7}$$

gdzie:
B(W) – macierz kowariancji międzyklasowej (wewnątrzklasowej),
tr – ślad macierzy,
u – liczba klas,
n – liczba obiektów.

Indeks Calińskiego i Harabasza nazywany jest pseudostatystyką F (zob. [Lattin, Carroll, Green 2003, s. 291]).

2. Indeks Huberta i Levine [1976]:

$$G2(u) = \frac{D(u) - r \cdot D_{\min}}{r \cdot D_{\max} - r \cdot D_{\min}}, \tag{8}$$

gdzie:
$D(u)$ – suma wszystkich odległości wewnątrzklasowych,
r – liczba odległości wewnątrzklasowych,
$D_{\min}(D_{\max})$ – najmniejsza (największa) odległość wewnątrzklasowa.

3. Indeks Gamma Bakera i Huberta [1975]:

$$G3(u) = \frac{s(+) - s(-)}{s(+) + s(-)}, \tag{9}$$

gdzie:
$s(+)$ – liczba par odległości zgodnych,
$s(-)$ – liczba par odległości niezgodnych.

Przy obliczaniu indeksu Gamma (por. [Gordon 1999, s. 62]) porównuje się wszystkie odległości wewnątrzklasowe z wszystkimi odległościami międzyklasowymi. Liczba tych porównań wynosi więc $r \cdot c$, gdzie $r(c)$ to liczba odległości wewnątrzklasowych (międzyklasowych). Jeśli odległość wewnątrzklasowa jest mniejsza (większa) niż odległość międzyklasowa, to parę taką uznajemy za zgodną (niezgodną). Odległości wewnątrzklasowe równe międzyklasowym nie są uwzględniane we wzorze (9).

Maksymalna wartość $G1(u)$ i $G3(u)$ oraz minimalna $G2(u)$ wskazuje najlepszy podział zbioru obiektów, a zarazem wyznacza liczbę klas.

9. Walidacja wyników klasyfikacji

Testowanie braku struktury klas. Z logicznego punktu widzenia testowanie braku struktury klas powinno odbywać się w etapie trzecim procesu klasyfikacji (po ustaleniu zbioru obiektów, zbioru zmiennych i zebraniu danych statystycznych). W literaturze jednak testowanie braku struktury klas przeprowadza

się w etapie walidacji wyników klasyfikacji. Wynika to z ograniczonej użyteczności dostępnych testów (zob. [Everitt, Landau, Leese 2001, s. 180]), które w praktyce zwykle nie są stosowane.

W konstrukcji hipotezy zerowej mówiącej o braku struktury klas w badanym zbiorze obiektów wykorzystywane są m.in. modele (zob. [Gordon 1999, s. 186–188]):

– Poissona (zakładający, że obiekty są reprezentowane przez punkty, które są jednostajnie rozłożone w pewnym regionie m-wymiarowej przestrzeni),

– jednomodalny (zakładający, że m-wymiarowe obserwacje wygenerowane są z jednomodalnego rozkładu częstości),

– losowej macierzy odległości (zakładający, że elementy dolnego trójkąta macierzy odległości są uszeregowane w losowym porządku; wszystkie $(n(n-1)/2)!$ rankingi odległości są jednakowo prawdopodobne).

Zagadnienie testowania struktury klas omówiono w pracach: [Gordon 1999, 185–190], [Bock 1996, s. 377–453], [Everitt, Landau, Leese 2001, s. 180–181], [Gordon 1998, s. 22–39], [Arnold 1979, s. 545–551].

Analiza replikacji (*powtórzenie klasyfikacji*). Replikacja w wypadku zagadnienia klasyfikacji dotyczy przeprowadzenia procesu klasyfikacji zbioru obiektów na podstawie dwóch prób wylosowanych z danego zbioru danych, a następnie oceny zgodności otrzymanych rezultatów. Procedura replikacji składa się z następujących etapów [Milligan 1996, s. 368–369], [Gordon 1999, s. 184]:

– podzielić losowo zbiór danych (zbiór n obiektów opisanych m zmiennymi) na dwa podzbiory A i B,

– zastosować wybraną metodę klasyfikacji do podziału zbioru A na ustaloną liczbę klas u. Wyznaczyć środki ciężkości (*centroids*) dla poszczególnych klas,

– obliczyć odległości obiektów ze zbioru B od środków ciężkości klas wyznaczonych na podstawie podzbioru A,

– przydzielić obiekty z podzbioru B do klas zawierających najbliższy środek ciężkości. Prowadzi to do podziału podzbioru B na nie więcej niż u klas,

– zastosować tę samą metodę klasyfikacji do podziału podzbioru B na u klas,

– policzyć, np. za pomocą skorygowanej miary Randa, zgodność wyników dwóch podziałów podzbioru B. Poziom zgodności wyników dwóch podziałów podzbioru B odzwierciedla stabilność przeprowadzonej klasyfikacji zbioru obiektów.

Ocena jakości klasyfikacji. Syntetyczny miernik pozwalający mierzyć prawidłowość zaklasyfikowania poszczególnych obiektów do klas, prawidłowość wyodrębnienia poszczególnych klas oraz ogólną jakość klasyfikacji (relatywną zwartość i separowalność klas) zaproponował P.J. Rousseeuw w 1987 r. (zob. [Kaufman, Rousseeuw 1990, s. 83–88]). Wskaźnik Rousseeuwa (*silhou-*

ette index) pozwalający oceniać prawidłowość zaklasyfikowania poszczególnych obiektów do klas przyjmuje postać:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i); b(i)\}}, \tag{10}$$

gdzie:

$a(i)$ – średnia odległość obiektu i od obiektów należących do klasy P_r,

$a(i) = \sum_{k \in \{P_r / i\}} \frac{d_{ik}}{(n_r - 1)}$,

$b(i) = \min_{s \neq r} \{d_{iP_s}\}$; d_{iP_s} – średnia odległość obiektu i od obiektów należących do klasy $P_s \left(d_{iP_s} = \sum_{k \in P_s} \frac{d_{ik}}{n_s}\right)$,

$r, s = 1, ..., u$ – numer klasy,

u – liczba klas.

Indeks $S(i)$ przyjmuje wartości z przedziału $[-1; 1]$. Im bliżej wartości jeden, tym dany obiekt silniej należy do wyodrębnionej klasy. W wypadku klas jednoelementowych $S(i) = 0$. W programie *R* autorzy przyjęli dla klas jednoelementowych $S(i) = 1$.

Tabela 4

Interpretacja wartości miernika $S(P)$

$\hat{S}(P)$	Interpretacja
(0,70; 1,00]	silna struktura klas
(0,50; 0,70]	poważna struktura klas
(0,25; 0,50]	słaba struktura klas (należy zastosować inne metody klasyfikacji)
0,.25 i mniej	nie odkryto struktury klas

Źródło: [Kaufman, Rousseeuw 1990, s. 88].

Indeksy pozwalające mierzyć prawidłowość wyodrębnienia poszczególnych klas oraz ogólną jakość klasyfikacji (relatywną zwartość i separowalność klas) są następujące:

$$S(P_r) = \sum_{i \in P_r} \frac{S(i)}{n_r}, \tag{11}$$

$$S(P) = \sum_i \frac{S(i)}{n}. \tag{12}$$

Subiektywną ocenę przedziałów wartości miernika $S(P)$ zawiera tabela 4.

10. Opis (interpretacja) i profilowanie klas

W wyniku zastosowania do klasyfikacji zbioru obiektów wybranej metody klasyfikacji otrzymuje się podział tego zbioru na klasy $P_1, ..., P_u$. W badaniach marketingowych podstawowym zagadnieniem staje się w związku z tym:

– opis (interpretacja) otrzymanych wyników, tj. wskazanie cech charakterystycznych poszczególnych klas oraz wyjaśnienie, jakimi czynnikami różnią się wyodrębnione klasy. Podstawą opisu (interpretacji) wyodrębnionych klas są zmienne, które brały udział w procesie klasyfikacji zbioru obiektów. Dla ułatwienia interpretacji otrzymanych rezultatów klasyfikacji wyznacza się środki ciężkości poszczególnych klas (średnie arytmetyczne obliczone z wartości pierwotnych każdej zmiennej na podstawie obiektów tworzących daną klasę) oraz odchylenia standardowe zmiennych w poszczególnych klasach. Na ten sposób rozwiązania problemu interpretacji rezultatów klasyfikacji wskazują m.in.: J.E. Hair, R.E. Anderson, R.L. Tatham i W.C. Black [1995, s. 443], A. Sokołowski [1992, s. 47], K. Jajuga [1990, s. 134], F. Robles i R. Sarathy [1986]. Taki sposób opisu klas możliwy jest do zastosowania tylko wtedy, gdy zmienne użyte w zagadnieniu klasyfikacji zbioru obiektów są mierzone na skali przedziałowej lub ilorazowej (dla tych skal dopuszcza się użycie średniej arytmetycznej i odchylenia standardowego). Jeśli klasyfikacja jest przeprowadzana na podstawie zmiennych mierzonych na skali porządkowej lub nominalnej, to możliwe jest wyznaczenie opisowej (werbalnej) charakterystyki poszczególnych klas dla każdej zmiennej. Można wyznaczyć frakcje i odsetki występowania w danej klasie poszczególnych kategorii zmiennych;

– profilowanie klas. Celem profilowania klas jest wskazanie cech charakterystycznych poszczególnych klas pozwalających na wskazanie różnic pomiędzy nimi. Profilowanie klas przeprowadza się na podstawie zmiennych, które nie brały udziału w procesie klasyfikacji zbioru obiektów. Typowymi zmiennymi stosowanymi w profilowaniu klas w badaniach marketingowych są zmienne demograficzne, geograficzne, socjoekonomiczne, psychograficzne i in., które charakteryzują konsumentów (nabywców) poszczególnych klas. Profilowanie przeprowadza się zwykle z wykorzystaniem takich metod, jak (por. np. [Sagan 1998, s. 180], [*Multivariate Data*..., 1998, s. 501, 513–515]): analiza dyskryminacyjna, drzewa klasyfikacyjne, tabulacja krzyżowa (tablice kontyngencji).

11. Podsumowanie

W pierwszej części artykułu wyróżniono i scharakteryzowano cztery podstawowe problemy decydujące o skali trudności zagadnienia klasyfikacji. Następnie w syntetycznej formie zaprezentowano problemy decyzyjne wymagające rozstrzygnięcia w procesie klasyfikacji zbioru obiektów obejmującym osiem etapów. W kolejnych częściach, na podstawie światowej literatury klasyfikacyjnej oraz własnych doświadczeń, przedstawiono rekomendowane strategie postępowania. Opracowanie ma charakter porządkujący wiedzę z omawianego zakresu.

Literatura

Ajvazjan S.A., Bežaeva Z.I., Staroverov O.V. [1974], *Klassifikacija mnogomernych nabludenij*, Statistika, Moskva.

Analysis of Symbolic Data [2000], H.H. Bock., E. Diday (eds), Springer-Verlag, Berlin–Heidelberg.

Anderberg M.R. [1973], *Cluster Analysis for Applications*, Academic Press, New York–San Francisco–London.

Arnold S.J. [1979], *A Test for Clusters*, „Journal of Marketing Research", November, vol. 16.

Baker F.B., Hubert L.J. [1975], *Measuring the Power of Hierarchical Cluster Analysis*, „Journal of the American Statistical Asscociation", 70.

Bock H.H. [1996], *Probability Models and Hypotheses Testing in Patitioning Cluster Analysis* [w:] *Clustering and Classification*, P. Arabie., L.J. Hubert, G. de Soete (red.), World Scientific, Singapore.

Caliński R.B., Harabasz J. [1974], *A Dendrite Method for Cluster Analysis*, „Communications in Statistics", vol. 3.

Cormack R.M. [1971], *A Review of Classification (with Discussion)*, „Journal of the Royal Statistiçal Society", Ser. A, part 3.

Cox T.F., Cox M.A.A. [1994], *Multidimensional Scaling*, Chapman & Hall, London.

Cox T.F., Cox M.A.A. [2000], *A General Weighted Two-way Dissimilarity Coefficient*, „Journal of Classification", vol. 17.

Dąbrowski M., Laus-Mączyńska K. [1978], *Metody wyszukiwania i klasyfikacji informacji*, WNT, Warszawa.

Everitt B.S. [1974], *Cluster Analysis*, Heinemann, London.

Everitt B.S., Landau S., Leese M. [2001], *Cluster Analysis*, Edward Arnold, London.

Fisher L., Van Ness J.W. [1971], *Admissible Clustering Procedures*, „Biometrika", nr 1.

Fowlkes E.B., Gnanadesikan R., Kettenring J.R. [1988], *Variable Selection in Clustering*, „Journal of Classification", vol. 5.

Fowlkes E.B., Mallows C.L. [1983], *A Method for Comparing Two Hierarchical Clusterings*, „Journal of the American Statistical Association", nr 383.

Gnanadesikan R., Kettenring J.R., Tsao S.L. [1995], *Weighting and Selection of Variables for Cluster Analysis*, „Journal of Classification", vol. 12.

Goodman L.A., Kruskal W.H. [1979], *Measures of Association for Cross Classifications*, Springer-Verlag, New York–Heidelberg.

Gordon A.D. [1981], *Classification*, Chapman and Hall, London.

Gordon A.D. [1987], *A Review of Hierarchical Classification*, „Journal of the Royal Statistical Society", ser. A.
Gordon A.D. [1996], *Hierarchical Classification* [w:] *Clustering and Classification*, P. Arabie, L.J. Hubert, G. de Soete (eds),World Scientific, Singapore.
Gordon A.D. [1998], *Cluster Validation* [w:] *Data Science, Classification, and Related Methods*, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.H. Bock, Y. Baba (red.), Springer, Tokyo.
Gordon A.D. [1999], *Classification*, Chapman and Hall/CRC, London.
Gower J.C. [1971], *A General Coefficient of Similarity and Some of Its Properties*, „Biometrics" (27).
Grabiński T. [1990], *Problemy analizy poprawności procedur taksonomicznych* [w:] *Taksonomia – teoria i jej zastosowania*, Materiały z konferencji, red. J. Pociecha, Wydawnictwo AE w Krakowie, Kraków.
Grabiński T. [1992], *Metody taksonometrii*, Wydawnictwo AE w Krakowie, Kraków.
Grabiński T., Wydymus S., Zeliaś A. [1989], *Metody taksonomii numerycznej w modelowaniu zjawisk społeczno-gospodarczych*, PWN, Warszawa.
Hartigan J.A. [1967], *Representation of Similarity Matrices by Trees*, „Journal of the American Statistical Association", vol. 62.
Hubert L.J., Arabie P. [1985], *Comparing Partitions*, „Journal of Classification", nr 1.
Hubert L.J., Levine J.R. [1976], *Evaluating Object Set Partitions: Free Sort Analysis and Some Generalizations*, „Journal of Verbal Learning and Verbal Behaviour", 15.
Hussain M. [1982], *Taksonomiczne metody podziału zbiorów skończonych*, praca doktorska, AE w Krakowie, Kraków.
Jajuga K. [1989], *Podstawowe metody analizy wielowymiarowej w przypadku występowania zmiennych mierzonych na różnych skalach*, paca wykonana w ramach CPBP 10.09, AE we Wrocławiu, Wrocław.
Jajuga K. [1990], *Statystyczna teoria rozpoznawania obrazów*, PWN, Warszawa.
Jajuga K., Walesiak M. [2000], *Standardisation of Data set Under Different Measurement Scales* [w:] *Classification and Information Processing at the Turn of the Millennium*, R. Decker, W. Gaul (eds), Springer-Verlag, Berlin–Heidelberg.
Jardine N., Sibson R. [1971], *Mathematical Taxonomy*, Wiley, New York.
Kaufman L., Rousseeuw P.J. [1990], *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley, New York.
Kolonko J. [1979], *O wykorzystaniu w badaniach taksonomicznych danych pierwotnych mierzonych na skalach różnego typu* [w:] *Metody taksonomiczne i ich zastosowanie w badaniach ekonomicznych*, Materiały konferencyjne, materiał powielony, Szklarska Poręba, 25.10.1979.
Lattin J.M., Carroll J.D., Green P.E. [2003], *Analyzing Multivariate Data*, Brooks/Cole, Pacific Grove.
Lerman J.C. [1988], *Comparing Partitions (Mathematical and Statistical Aspects)* [w:] *Classification and Related Methods of Data Analysis*, H.H. Bock (ed.), North-Holland, Amsterdam.
Lira J., Wagner W., Wysocki F. [2002], *Mediana w zagadnieniach porządkowania liniowego obiektów wielocechowych* [w:] *Statystyka regionalna w służbie samorządu lokalnego i biznesu*, red. J. Paradysz, Internetowa Oficyna Wydawnicza, Centrum Statystyki Regionalnej, AE w Poznaniu, Poznań.
Milligan G.W. [1981], *A Review of Monte Carlo Tests of Cluster Analysis*, „Multivariate Behavioral Research", 16.
Milligan G.W. [1989], *A Validation Study of a Variable Weighting Algorithm for Cluster Analysis*, „Journal of Classification", vol. 6.
Milligan G.W. [1994], *Issues in Applied Classification: Selection of Variables to Cluster*, Classification Society of North America Newsletter, November, issue 37.

Milligan G.W. [1996], *Clustering Validation: Results and Iimplications for Applied Analyses* [w:] *Clustering and Classification*, P. Arabie, L.J. Hubert, G. de Soete (eds), World Scientific, Singapore.

Milligan G.W., Cooper M.C. [1985], *An Examination of Procedures for Determining the Number of Clusters in a Data Set*, „Psychometrika", nr 2.

Milligan G.W., Cooper M.C. [1988], *A Study of Standarization of Variables in Custer Analysis*, „Journal of Classification", no 2.

Multivariate Data Analysis with Readings [1995], J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, Prentice Hall, Englewood Cliffs.

Multivariate Data Analysis [1998], J.F. Hair, R.E. Anderson, R.L. Tatham, W.C. Black, Prentice Hall, Englewood Cliffs.

Nowak E. [1985], *Wskaźnik podobieństwa wyników podziałów*, „Przegląd Statystyczny", z. 1.

Pociecha J. [1982], *Kryteria oceny procedur taksonomicznych*, „Przegląd Statystyczny", z. 1/2.

Rand W.M. [1971], *Objective Criteria for the Evaluation of Clustering Methods*, „Journal of the American Statistical Association", nr 336.

Robles F., Sarathy R. [1986], *Segmenting the Commuter Aircraft Market with Cluster Analysis*, „Industrial Marketing Management", vol. 15.

Rousseeuw P.J. [1987], *Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis*, „Journal of Computational and Applied Mathematics", 20.

Rybaczuk M. [2002], *Graficzna prezentacja struktury danych wielowymiarowych* [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, K. Jajuga, M. Walesiak (red.), Taksonomia 10, Prace Naukowe AE we Wrocławiu, Wrocław, nr 942.

Sagan A. [1998], *Badania marketingowe. Podstawowe kierunki*, Wydawnictwo AE w Krakowie, Kraków.

Sneath P.H.A. [1969], *Evaluation of Clustering Methods (with Discussion)* [w:] *Numerical Taxonomy*, A.J. Cole (ed.), Academic Press, London.

Sokal R.R., Rohlf F.J. [1962], *The Comparison of Dendrograms by Objective Methods*, „Taxonomy", no 2.

Sokołowski A. [1976], *Metoda porównywania wyników podziału zbioru skończonego*, XII Konferencja Naukowa Ekonometryków, Statystyków i Matematyków Akademii Ekonomicznych Polski Południowej, Karpacz.

Sokołowski A. [1992], *Empiryczne testy istotności w taksonomii*, AE w Krakowie, Zeszyty Naukowe – Seria specjalna: Monografie, Kraków, nr 108.

Szmigiel C. [1976], *Wskaźnik zgodności kryteriów podziału*, „Przegląd Statystyczny", z. 4.

Van Ness J.W. [1973], *Admissible Clustering Procedures*, „Biometrika", 60.

Walesiak M. [1988], *Sposoby wyznaczania optymalnej liczby klas w zagadnieniu klasyfikacji hierarchicznej*, Prace Naukowe AE we Wrocławiu, Wrocław, nr 449.

Walesiak M. [1993], *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Prace Naukowe AE we Wrocławiu, nr 654, Seria: Monografie i Opracowania, nr 101.

Walesiak M. [2002], *Uogólniona miara odległości w statystycznej analizie wielowymiarowej*, Wydawnictwo AE we Wrocławiu, Wrocław.

Walesiak M. [2003], *Miara odległości obiektów opisanych zmiennymi mierzonymi na różnych skalach pomiaru*, Prace Naukowe AE we Wrocławiu, Wrocław, nr 1006.

Walesiak M. [2004], *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, Prace Naukowe AE we Wrocławiu, Wrocław nr 1010.

Walesiak M. [2005], *Problemy selekcji i ważenia zmiennych w zagadnieniu klasyfikacji* [w:] *Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Prace Naukowe AE we Wrocławiu, Taksonomia 12 (w druku).

Wallace D.L. [1983], *Comment*, „Journal of the American Statistical Association", vol. 78, nr 383.