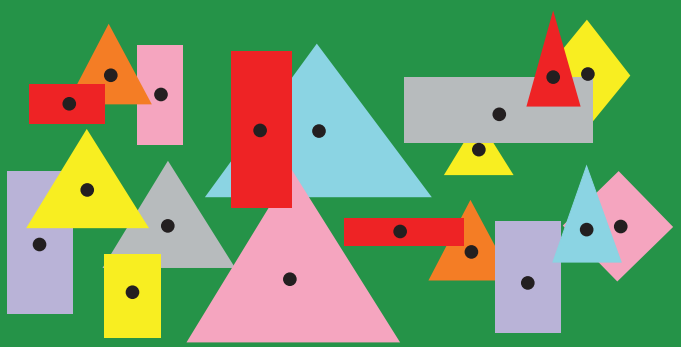
Jerzy Korzeniewski

Metody selekcji zmiennych w analizie skupień

Nowe procedury



WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

Jerzy Korzeniewski

Metody selekcji zmiennych w analizie skupień

Nowe procedury

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

ŁÓDŹ 2012

RECENZENT
Grażyna Trzpiot

PROJEKT OKŁADKI
Barbara Grzejszczak

Wydrukowano z gotowych materiałów dostarczonych do Wydawnictwa UŁ

ISBN 978-83-7525-695-6

Wydawnictwo Uniwersytetu Łódzkiego
90-131 Łódź, ul. Lindleya 8

Wydanie I. Nakład 141 egz. Ark. druk. 11,875
Zam. 5050/2012. Cena zł 26,– + VAT

SPIS TREŚCI

1. Wprowadzenie

1.1. Cele pracy

Analiza skupień (*cluster analysis*) zwana też nauczaniem bez nadzoru (*unsupervised learning*) lub taksonomią (*taxonomy*) poświęcona jest zagadnieniu pogrupowania obiektów zbioru danych w rozłączne grupy (skupienia) spójne wewnętrznie i jak najbardziej zróżnicowane pomiędzy sobą. Poprawnie przeprowadzona analiza skupień umożliwia podzielenie zbioru danych na grupy, celem lepszego zrozumienia informacji w nich zawartych i ustalenia własności grup obiektów do siebie podobnych oraz ich syntetycznej charakterystyki. Analiza skupień odgrywa bardzo ważną rolę w różnych dziedzinach nauki, w tym w analizach zjawisk społeczno-ekonomicznych. Typowym przykładem może być segregacja klientów banku lub firmy na grupy klientów podobnych do siebie pod względem cech istotnych dla banku lub firmy. Analiza skupień rozwija się od ponad 50 lat. W tym czasie opracowano wiele różnych metod realizujących zadania cząstkowe analizy skupień, które nazywane są dzisiaj etapami analizy skupień (por. § 1.2). W ostatnich latach, wzorem analogicznej drogi w przypadku klasyfikacji (nauczania z nadzorem) rozwija się podejście zagregowane do analizy skupień (*ensemble clustering*), którego ideą jest łączenie wyników uzyskanych przy pomocy różnych metod, mające na celu poprawienie wyników. Realizacja poszczególnych etapów analizy skupień musi być wykonywana w oparciu o zbiór zmiennych charakteryzujących obserwacje. W zagadnieniach praktycznych zawsze stoimy przed koniecznością określenia i wyboru zbioru zmiennych, którymi będziemy się posługiwać. Jeśli zmienne zostaną niewłaściwie wybrane, to fakt ten może w znacznym stopniu obniżyć jakość otrzymanych wyników. Jak zaznacza się w literaturze wybór zmiennych w klasyfikacji jest jednym z najważniejszych a zarazem najtrudniejszych zadań. Zmienne, które tworzą strukturę skupień w zbiorze danych nazywać będziemy istotnymi (*relevant features, true features*) zaś pozostałe zmienne nieistotnymi lub zakłócającymi (*irrelevant features, masking features, noisy features).*

Cele rozprawy można ująć następująco:

1. Ocena dotychczasowego dorobku naukowego w zakresie wybierania zmiennych tworzących strukturę skupień w zbiorze danych;
2. Zaproponowanie nowych rozwiązań, które będą efektywniejsze od metod istniejących, nieobarczone tyloma założeniami co najlepsze dotychczas skonstruowane metody, założeniami takimi jak:

- konieczność znajomości liczby skupień w zbiorze danych;
- ograniczenia możliwości stosowania w przypadku występowania zmiennych mierzonych na różnych skalach;
- ograniczenie do wyboru o charakterze porównawczym pomiędzy dwoma zbiorami zmiennych, czego konsekwencją jest brak możliwości wyszukiwania wielokrotnych struktur danych;
- konieczność odwołania się do subiektywnej oceny stosowanych wskaźników.

3. Zbadanie efektywności nowych rozwiązań w zastosowaniu ich do selekcji zmiennych w empirycznych zbiorach danych o charakterze ekonomiczno-socjologicznym.

Pierwszy rozdział pracy jest poświęcony zagadnieniom wstępnym i zawiera charakterystykę etapów analizy skupień, ogólne omówienie uwarunkowań zagadnienia wybierania zmiennych w analizie skupień oraz stosowane w dalszym ciągu rozprawy charakterystyki pojedynczych zmiennych, miary podobieństwa podziałów zbioru danych, miary korelacji zmiennych, miary jakości wyników dokonanego wyboru zmiennych. Ponadto, przedstawione są zbiory danych generowanych stosowane później w eksperymentach symulacyjnych.

Zawartość drugiego oraz trzeciego rozdziału podzielona została według kryterium odnoszącego się do tego czy metoda oparta jest na wnioskowaniu statystycznym na podstawie modelu czy też jest algorytmem czysto heurystycznym nie opartym na modelu. Alternatywnym podziałem mógłby być, na przykład, podział na metody dokonujące selekcji zmiennych i metody ważące zmienne. Autor niniejszej monografii stoi na stanowisku, że taki podział byłby o wiele mniej różnicujący, gdyż selekcja zmiennych jest szczególnym przypadkiem ważenia zmiennych. Ponadto, bardzo istotne jest też to, że problematyka metod modelowych jest podobna dla wszystkich tego typu metod, gdyż problemem zasadniczym jest estymacja parametrów modelu, na ogół, dokonywana przy wykorzystaniu tego samego algorytmu. Inną alternatywą mógłby być, na przykład, podział metod ze względu na rodzaj skali pomiarowej zmiennych opisujących obiekty zbioru. Ta alternatywa wydaje się mniej rozsądna ze względu na to, że niektóre metody można stosować zarówno do zbiorów danych z silnymi jak i słabymi skalami pomiarowymi.

Czwarty rozdział zawiera propozycję autorską nowych metod wyboru zmiennych w analizie skupień oraz wyniki badania tej metody za pomocą eksperymentów Monte Carlo. Ideą nowej metody jest zastąpienie korelacji pomiędzy zmiennymi opisującymi obserwacje korelacją pomiędzy odległościami par obiektów oraz zastąpienie grupowania na określoną liczbę skupień grupowaniem wielokrotnym na dwa skupienia. Uzyskujemy dzięki temu

uniwersalną metodę, którą możemy stosować zarówno do słabych jak i silnych skal pomiarowych mierzących wartości zmiennych. Ponadto, do pewnego stopnia uwalniamy się od konieczności znajomości liczby skupień w zbiorze danych.

Piąty rozdział poświęcony jest stronie aplikacyjnej zagadnienia tj. ocenie porównawczej zaproponowanych nowych metod oraz najlepszych dotychczas skonstruowanych metod na podstawie efektywności uzyskanej w zastosowaniu do empirycznych zbiorów danych. Do tej pory nie udało się skonstruować metody nieobarczonej co najmniej kilkoma założeniami i takiej, która dobrze wybierałaby zmienne dla zbiorów danych opisanych przez zmienne mierzone na różnych skalach i dawała dobre efekty zarówno dla dużych (w sensie liczby zmiennych) jak i dla małych zbiorów danych. Próba podjęta przez autora ma na celu znalezienie takiej metody, która dla badaczy będzie najbardziej wiarygodna – odporna na niektóre zakłócenia.

Problem selekcji zmiennych w analizie skupień można rozpatrywać w różnych ujęciach w odniesieniu do założeń mniej lub bardziej ograniczających stosowalność konstruowanych metod. Niektóre ograniczenia zostały wymienione w celach rozprawy. Należy jeszcze wspomnieć o tym, że w swej najogólniejszej postaci, w problemie selekcji dopuszcza się możliwość istnienia kilku struktur skupień w tym samym zbiorze danych, przy czym zbiory zmiennych tworzących te struktury nie muszą być rozłączne. Ta najogólniejsza forma problemu jest na tyle skomplikowana, że metody, które ją rozważają spisują się bardzo słabo w zawężonych formach problemu, na przykład, w najprostszej postaci, w której zakładamy, że istnieje tylko jeden zbiór zmiennych tworzących strukturę skupień. Propozycje autora mają przede wszystkim na uwadze ujęcie problemu, w którym zakładamy rozłączność zbiorów zmiennych tworzących różne struktury skupień.

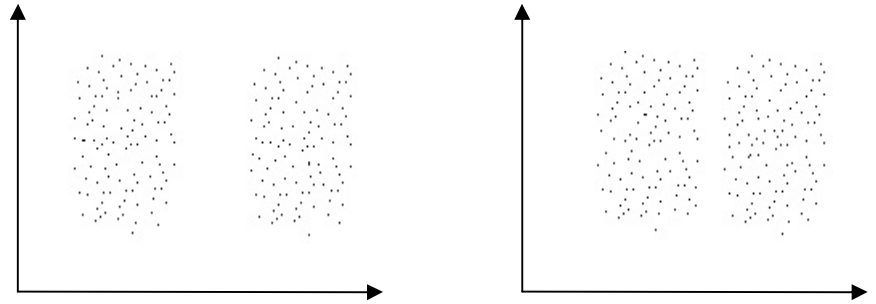
W zakończeniu przedstawione są wnioski z przeprowadzonych badań podsumowujące wszystkie uzyskane oceny metod istniejących oraz propozycji autora.

Rozprawa niniejsza była wspomagana środkami z grantu habilitacyjnego nr 4323/B/H03/2009/37.

1.2. Charakterystyka analizy skupień. Wybór mierników

Analiza skupień jest dziedziną statystyki bardzo istotną pod względem aplikacyjnym. Konsekwencją potrzeby opracowania metod niezbędnych do segmentacji zbiorów danych na spójne podgrupy obiektów przy braku jednoznacznej i precyzyjnej definicji takich podgrup jest pewien chaos w dorobku tej dziedziny. Niewiele jest wyników teoretycznych, na przykład, o wiele mniej niż w pokrewnej dziedzinie statystyki jaką jest klasyfikacja ze

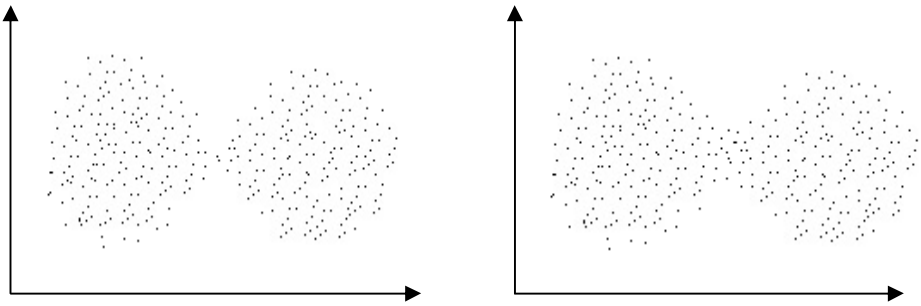
zbiorem uczącym, natomiast bardzo wiele różnych algorytmów i metod heurystycznych.



Rys.1.1. Dwa przykłady zbiorów punktów dwuwymiarowych, z których jeden tworzy bardzo wyraźną strukturę dwóch skupień.

Ź r ó d ł o: opracowanie własne.

Przyczyną takiego stanu rzeczy jest to, że nie ma jednoznacznej definicji skupienia i, co za tym idzie, często trudno zdecydować czy w danym zbiorze istnieje struktura skupień czy też nie. Na rysunkach 1.1 oraz 1.2 przedstawione są przykłady czterech zbiorów punktów dwuwymiarowych. Zbiory po lewej stronie tworzą wyraźną lub bardzo wyraźną strukturę skupień. Zbiór po prawej stronie składa się z podobnych podzbiorów punktów ale orzeczenie o istnieniu struktury skupień zależy od wizualnych upodobań patrzącego.



Rys.1.2. Dwa przykłady zbiorów punktów dwuwymiarowych, z których jeden tworzy dość wyraźną strukturę dwóch skupień.

Ź r ó d ł o: opracowanie własne.

Powyższe, dość lapidarne sformułowania, precyzyjniej ujął światowej klasy ekspert w dziedzinie informatyki, sztucznej inteligencji i możliwości

wykorzystania komputerów do rozwiązywania problemów świata realnego, Jon Kleinberg. Wyróżnił on (*Kleinberg,* 2002) trzy własności, które w analizie skupień byłyby bardzo pożyteczne. Pierwsza własność to niezmienniczość grupowania względem skali odległości (*scale invariance*) polegająca na tym, że dla dowolnego skalara *c* i dowolnej miary odległości *d* funkcja *f* decydująca o grupowaniu obiektów powinna mieć własność, którą symbolicznie można ująć jako $f(d) = f(c \cdot d)$. Druga własność to zupełność grupowania (*richness*) polegająca na tym, że dla jakiegokolwiek podziału *P* powinna istnieć odległość *d* taka, że $f(d) = P$. Trzecia własność to zgodność grupowania (*consistency*) polegająca na tym, że $f(d) = P$ implikuje $f(d^*) = P$, gdzie odległość d^* jest transformacją odległości *d* indukowaną przez podział *P* w następujący sposób: odległość indukowana dowolnych dwóch obiektów z tego samego skupienia w podziale *P* nie przekracza odległości *d* tych obiektów oraz odległość indukowana dowolnych dwóch obiektów z różnych skupień w podziale *P* jest co najmniej taka jak odległości *d* tych obiektów. Następnie Kleinberg udowodnił twierdzenie o niemożności (*impossibility theorem*): nie ma funkcji decydującej o grupowaniu obiektów, która miałaby wszystkie trzy własności. Można polemizować z tym czy sformułowane własności są rzeczywiście niezbędne w każdej aplikacji empirycznej analizy skupień, ale, z pewnością, twierdzenie o niemożności rozwiewa wiele wątpliwości nękających wszystkich badaczy starających się znaleźć jak najlepszą metodę do grupowania obiektów badanego zbioru danych.

W czasie kilkudziesięciu lat rozwoju analizy skupień wypracowano zgodę co do tego, że pełna analiza skupień zbioru danych powinna objąć następujące etapy:

1. Wybór obiektów i zmiennych.
2. Wizualizacja obiektów (lub zmiennych).
3. Normalizacja zmiennych.
4. Wybór miary odległości pomiędzy obiektami.
5. Wybór metody grupowania obiektów.
6. Ustalenie liczby skupień.
7. Grupowanie obiektów – właściwy etap analizy skupień.
8. Ocena wyników grupowania.
9. Opis i profilowanie klas.

Sokołowski (1992) ujmuje zadania analizy skupień jeszcze ogólniej, włączając do charakterystyk zbioru czynnik czasowy. Następnie dzieli on zadania taksonomiczne na proste (np. grupowanie obiektów), złożone (grupowanie obiektów i cech) i kompleksowe.

Wybór obiektów zbioru danych polega na ustaleniu czy istnieją w zbiorze jakieś obserwacje przypadkowe będące, na przykład konsekwencją błędów pomiarowych lub sposobu zbierania danych, które nie powinny być uwzględniane w dalszej analizie. Taka procedura łączy się z pojęciem odporności i, jak wiadomo, jest ściśle powiązana z wyborem cech opisujących obiekty a nawet doborem metod do końcowych etapów analizy skupień, na przykład, metod grupowania obiektów. Może się bowiem okazać, że dla niektórych metod grupowania, uwzględnienie wszystkich obiektów prowadzi do mniejszych strat jakości niż usuwanie obiektów podejrzanych o to, że są obiektami nietypowymi (*outliers*). Ponadto wybór obiektów jest ściśle związany z wyborem zmiennych opisujących obiekty. Selekcja zmiennych jest jednym z najważniejszych i jednocześnie najtrudniejszych etapów analizy skupień. Uwarunkowania tego etapu zostały przedstawione bardziej szczegółowo w następnym paragrafie.

Etap wizualizacji danych jest ważny dla analizy skupień, gdyż jest pomocny w odkryciu ewentualnej struktury skupień w zbiorze danych, ich liczby, a nawet przy wyborze właściwych algorytmów grupowania obiektów (biorąc pod uwagę ich własności). Wizualizacji danych można dokonać nie tylko dla wartości oryginalnych zmiennych, lecz również dla macierzy odległości między zmiennymi. W tym zakresie, oprócz tradycyjnych wykresów dwuwymiarowych, pomocne są (por. np. *Everitt i inni,* 2001): wielowymiarowe wykresy rozrzutu; trójwymiarowe wykresy zmiennych; metody skalowania wielowymiarowego i sieci Kohonena, umożliwiające graficzne przedstawienie danych w przestrzeni o mniejszej liczbie wymiarów. Program R pozwala dodatkowo tworzyć bardziej złożone wykresy (*Gatnar, Walesiak,* 2009): wielowymiarowy wykres rozrzutu z funkcją gęstości poszczególnych zmiennych; wykres rozrzutu dla 3 zmiennych metrycznych w przestrzeni 2-wymiarowej *(bubbleplot)*. Metody graficzne można nawet, w przypadkach prostych zbiorów danych, wykorzystać we wcześniejszym etapie analizy skupień tj. etapie doboru zmiennych, np. wykres pudełkowy (ramka-wąsy, *box-whiskers*).

Ewentualna normalizacja zmiennych jest traktowana jak szczególny przypadek ważenia zmiennych, gdyż standaryzacja (lub inne przekształcenie cech) wpływa na rozmieszczenie obiektów w przestrzeni euklidesowej. Już Cormack (1971) zauważył ten problem i doszedł do wniosku, że na ogół standaryzacja zmiennych zmniejsza efektywność analizy skupień. Cel standaryzacji czyli wyeliminowanie różnic w skalach pomiaru cech, jest niespójny z zasadniczym celem analizy skupień, gdyż różnice pomiędzy cechami mogą wynikać z ich naturalnych własności implikujących istnienie ewentualnej struktury skupień. Dowiedziono, że nadawanie zmiennym wag odwrotnie proporcjonalnych do całkowitej zmienności cech – czyli standaryzacja zmiennych przez ich całkowite odchylenie – jest nieefektywny, a

wręcz niewskazany, gdyż utrudnia rozróżnianie grup obiektów podobnych. Tezę o tym, że normalizowanie pojedynczych zmiennych może mieć negatywny wpływ na zachowanie oryginalnej struktury skupień tzn., może tę strukturę zniekształcić lub nawet zniszczyć stawia również Stoddard (1979).

Milligan (1996) wskazuje na błędne przekonanie wielu badaczy, że fakt występowania znacznej różnicy zmienności cech w analizie skupień jest podstawą do przeprowadzenia standaryzacji zmiennych z obawy o to, by cechy o dużej zmienności nie miały nadmiernego wpływu na wyniki analizy skupień. Podkreśla, że normalizacja zmiennych jest kwestią indywidualną, a nie rutynowym przekształceniem jak również, że nie zawsze uzasadnione jest twierdzenie, że standaryzacja może ukryć ewentualną strukturę skupień występującą w zbiorze danych. Powołując się na wcześniejsze badania porównawcze, Milligan rozważa również ewentualne procedury normalizacji zmiennych. Z reguły przyjmuje się klasyczną normalizację zmiennych, jeśli natomiast rozważymy separowalność i wewnątrzgrupową zmienność skupień otrzymanych na podstawie cech unormowanych, to okazuje się, że na tle innych przekształceń jest to rozwiązanie nieefektywne. Zbadano osiem sposobów normalizacji (*Milligan, Cooper*, 1988), (0) brak normalizacji, (1) normalizacja za pomocą klasycznej standaryzacji, (2) standaryzacja przez odchylenie standardowe, (3) przekształcenie ilorazowe w oparciu o wartość maksymalną, (4) unitaryzacja przez rozstęp zmiennej, (5) unitaryzacja zerowana, (6) przekształcenie ilorazowe w oparciu o sumę wariantów cechy i (7) rangowanie. Najlepsze wyniki pod względem stopnia zgodności klasyfikacji z właściwą strukturą skupień (ocena indeksem Randa) za pomocą metod aglomeracyjnych uzyskano dla przekształceń normalizacyjnych opartych na rozstępie zmiennych (typ (4) i (5)).

Zasadność takiej tezy dotyczącej ważenia (normalizacji) zmiennych na potrzeby analizy skupień potwierdzają wyniki badań, które przedstawili Gnanadesikan i inni (1995). Autorzy zastosowali dziewięć procedur ważenia cech, m.in.: (0) brak ważenia, (1) standaryzację zmiennych z wykorzystaniem samego odchylenia standardowego, (2) unitaryzację opartą tylko na rozstępie, (3) macierz odwrotną wewnątrzklasowej zmienności cech, (4) macierz odwrotną elementów diagonalnych macierzy wewnątrzklasowej zmienności cech, (5) iloczyn elementów diagonalnych macierzy międzyklasowej i elementów diagonalnych macierzy odwrotnej do wewnątrzklasowej, spośród których najlepsze, tj. dające najmniejszy błąd klasyfikacji okazały się metody oparte na wewnętrznej zmienności cech w znalezionych skupieniach. Zastosowanie wag dla zmiennych w postaci ich wewnętrznej (lub zewnętrznej) zmienności cech w skupieniach (typ (3), (4) i (5)), może poprawiać wynik klasyfikacji przy silnej strukturze skupień. Spośród wszystkich formuł najlepszy okazał się typ (3), a wyniki pośrednie potwierdziły, że system wag równych (0) oraz standaryzacja

cech przez skalowanie (1) są zdecydowanie nieefektywne. Unitaryzacja zmiennych przy pomocy ich rozstępu (2) jest rozwiązaniem pośrednim, dającym mniejszy błąd klasyfikacji niż (0) czy (1), lecz większy niż formuły (3), (4) i (5). Na podstawie cytowanych wyników można by sądzić, że gdyby standaryzacja przez skalowanie była przeprowadzona w obrębie każdego ze znanych skupień, to jej efektywność wyodrębniania skupień byłaby wyższa. Problemem jest jednak to, że na tym etapie analizy skupień nie znamy dokładnej struktury skupień czyli przynależności obiektów do skupień.

Dla grupowania metodą Warda i metodą *k*-średnich efekty różnych formuł standaryzacyjnych badał również Steinley (2004). Formułę standaryzacyjną będącą kompromisem pomiędzy wymogiem jak najlepszego zachowania struktury skupień a nadawaniem większych wag zmiennym z większą zmiennością, zaproponowali Steinley i Brusco (por. wzór (3.37)). Formuła ta, zdaniem autorów, poprawia efektywność metody HINoV, ale ten wniosek można kwestionować (por. rozdz. 3 i 5).

Wyczerpujący przegląd formuł normalizacyjnych dla zmiennych ciągłych można znaleźć w pracy Pawełek (2008).

Kierując się cytowanymi wynikami badań, spośród zbadanych popularnych formuł standaryzacyjnych postaci

$$x_j' = \frac{x_j - a}{b} \quad \text{dla} \quad j = 1,\ldots,V \tag{1.1}$$

do standaryzowania zmiennych mierzonych na skali interwałowej oraz ilorazowej wybrano formułę w postaci unitaryzacji zerowanej przy zastosowaniu rozstępu, czyli

$$x_j' = \frac{x_j - \min x_j}{r_j} \quad \text{dla} \quad j = 1,\ldots,V \tag{1.2}$$

Wybór określonej miary odległości zdeterminowany jest przez skalę pomiarową, ale zależy również od sposobu standaryzacji (lub normalizacji) zmiennych. Miarą odległości powszechnie wykorzystywaną w analizie skupień w przypadku zmiennych mierzonych na silnych skalach (interwałowej i ilorazowej), ze względu na najlepiej zbadane jej własności i prostą interpretację geometryczną, jest odległość euklidesowa lub kwadrat tej odległości. Taką miarę (por. tab. 1.1) będziemy stosować dla zmiennych mierzonych na skali ilorazowej i interwałowej, w zbiorach danych występujących w eksperymentach symulacyjnych.

W przypadku zmiennych mierzonych na skali porządkowej rozsądnym wyborem, ze względu na posiadane własności jest odległość *GDM* (*Walesiak*, 2002). Odległość *GDM* dana jest wzorem

$$d_{ik} = \frac{1}{2} - \frac{\sum_{j=1}^{V} a_{ikj} b_{kij} + \sum_{j=1}^{V} \sum_{\substack{l=1 \\ l \neq i,k}}^{n} a_{ilj} b_{klj}}{2\sqrt{\sum_{j=1}^{V} \sum_{l=1}^{n} a_{ilj}^2 \cdot \sum_{j=1}^{V} \sum_{l=1}^{n} b_{klj}^2}} \quad , \tag{1.3}$$

gdzie: d_{ik} jest odległością pomiędzy obserwacjami o numerach *i, k*; symbole a_{ikj}, b_{kij}, a_{ilj}, b_{klj} są obliczane w zależności od skali pomiarowej. Dla skali ilorazowej i interwałowej stosuje się wzory

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \quad \text{dla} \quad p = k, l \\ b_{krj} &= x_{kj} - x_{rj} \quad \text{dla} \quad r = i, l \end{aligned} \tag{1.4}$$

Dla skali porządkowej stosuje się wzory

$$a_{ipj} = \begin{cases} 1 & \text{dla} \quad x_{ij} > x_{pj} \\ 0 & \text{dla} \quad x_{ij} = x_{pj} \\ -1 & \text{dla} \quad x_{ij} < x_{pj} \end{cases} \tag{1.5}$$

dla *p=k, l,* oraz

$$b_{krj} = \begin{cases} 1 & \text{dla} \quad x_{kj} > x_{rj} \\ 0 & \text{dla} \quad x_{kj} = x_{rj} \\ -1 & \text{dla} \quad x_{kj} < x_{rj} \end{cases} \tag{1.6}$$

dla *r=i, l.* Dla skali nominalnej stosuje się wzory

$$\sum_{j=1}^{V} \sum_{l=1}^{n} a_{ilj}^2 = \sum_{j=1}^{V} \sum_{l=1}^{n} b_{klj}^2 = V(n-1) \tag{1.7}$$

$$a_{ikj} b_{kij} = \begin{cases} 1 & \text{dla} \quad x_{ij} = x_{kj} \\ -1 & \text{dla} \quad x_{ij} \neq x_{kj} \end{cases} \tag{1.8}$$

oraz

$$a_{ilj} b_{klj} = \begin{cases} 1 & \text{dla} \quad x_{ij} = x_{kj} \wedge \left(x_{ij}, x_{kj} = x_{lj} \vee x_{ij}, x_{kj} \neq x_{lj} \right) \\ -1 & \text{dla} \quad x_{ij} \neq x_{kj} \wedge \begin{pmatrix} (x_{ij} \neq x_{lj}\ ,\ x_{kj} = x_{lj}) \vee \\ (x_{ij} = x_{lj}\ ,\ x_{kj} \neq x_{lj}) \vee (x_{ij}, x_{kj} \neq x_{lj}) \end{pmatrix} \end{cases}$$

(1.9)

dla $l \neq i, k$. Odległość dana wzorem (1.3) będzie stosowana zarówno w procesie grupowania obserwacji jak i przy obliczaniu wartości współczynników

korelacji. Miara *GDM* spełnia warunek symetryczności, zwrotności, nieujemności, lecz nie zawsze spełnia warunek „nierówności trójkąta". Odległość *GDM* nie zmienia swojej wartości w wyniku transformacji wartości zmiennych za pomocą dozwolonego w danej skali przekształcenia oraz jest unormowana na przedziale [0;1].

W przypadku zmiennych mierzonych na skali nominalnej (w szczególności binarnej), na ogół, stosowana jest miara Sokala-Michenera (por. tab. 1.1). Ta odległość będzie używana w niniejszej monografii.

Do pomiaru odległości pomiędzy obiektami opisanymi zmiennymi mierzonymi na różnych skalach pomiarowych zastosowana zostanie formuła (por. *Gatnar, Walesiak*, 2004, *Walesiak,* 2011)

$$d = \frac{w_1 d_1 + w_2 d_2 + w_3 d_3 + w_4 d_4}{w_1 + w_2 + w_3 + w_4} \quad , \qquad (1.10)$$

gdzie: $w_1, w_2, w_3, w_4 \in \{0,1,\ldots,V\}$ są wagami przypisanymi odległościom mierzonym na podzbiorze zmiennych ze skalą nominalną (d_1), podzbiorze zmiennych ze skalą porządkową (d_2), podzbiorze zmiennych ze skalą interwałową (d_3) oraz podzbiorze zmiennych ze skalą ilorazową (d_4). Wagi te są określone przez liczbę zmiennych odpowiadającą danej skali, wobec czego spełniają warunek $w_1 + w_2 + w_3 + w_4 = V$. Odległości d_1, d_2, d_3, d_4 obliczane będą według formuł przyjętych dla pojedynczych zmiennych.

Tabela 1.1. Miary odległości stosowane w pracy, *m* oznacza liczbę wszystkich zmiennych danego rodzaju skali w określonym podzbiorze zmiennych.

Typ skali pomiarowej	Nazwa odległości	Formuła odległości
Binarna, nominalna	Sokala-Michenera	$\frac{m_r}{m}$, m_r – liczba współrzędnych, na których oba obiekty różnią się, m – liczba wszystkich współrzędnych
Porządkowa	GDM	formuła (1.3) (przy $V=m$) z podstawieniami (1.5) i (1.6)
Ilorazowa	euklidesowa	$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2}$, m – liczba wszystkich współrzędnych
Ilorazowa	GDM	formuła (1.3) (przy $V=m$) z podstawieniem (1.4)

Ź r ó d ł o: opracowanie własne na podstawie *Gatnar, Walesiak* (2009).

Formuła (1.10) budzi najmniej zastrzeżeń (por. *Jajuga,* 1989, *Walesiak,* 1993) spośród wszystkich prób ujednolicenia miar odległości różnych skal pomiarowych. Miary odległości pomiędzy parami obiektów, jakie zastosowano do różnych skal pomiarowych, przedstawione są w tabeli 1.1.

W literaturze przedmiotu proponowane są trzy podejścia w zakresie wyboru metod grupowania obiektów. W pierwszym wybór oparty jest na analizie własności poszczególnych metod przy wykorzystaniu informacji niezależnych od badanego zbioru danych, w drugim przy wykorzystaniu macierzy danych. Trzecie podejście zwane jest strategią grupowania, polega na syntetyzacji wyników otrzymanych przy pomocy różnych metod.

Sposób pierwszy opiera się na porównaniu wyników grupowania poszczególnych metod ze znaną przynależnością obiektów do skupień. Takie podejście wykorzystuje się w badaniu własności algorytmów grupowania w eksperymentach symulacyjnych. W przypadku empirycznych zbiorów danych liczba skupień i przynależność obiektów do skupień nie jest znana, ale znając własności algorytmów grupowania można dokonać wstępnego wyboru algorytmu. Takie podejście, nie daje całkowitej gwarancji, że wybrana metoda będzie efektywna dla konkretnego, badanego zbioru danych. Dlatego też, niezbędny jest nadzór badacza oraz odpowiednie zaprojektowanie kilkuwariantowej analizy w celu wybrania najlepszego rozwiązania.

W podejściu drugim wykorzystywana jest macierz danych. To podejście polega na formalnej ocenie cech algorytmów, wśród których wymienia się m.in. własności (*Pociecha,* 1982): 1) najlepszego obrazu – wynik grupowania nie zależy od kolejności rozważania obiektów, 2) wypukłości zbioru grupowanych obiektów, 3) połączenia obiektów na wykresie drzewa (połączenia obiektów nie przecinają się), 4) poprawnej struktury grup – gdy w wyróżnionych grupach wszystkie odległości wewnętrzne są mniejsze niż wszystkie odległości zewnętrzne (struktura grupowa) lub gdy można ustalić kolejność podobieństwa między obiektami (struktura hierarchiczna), 5) poprawnej struktury połączeń drzewa (dla procedur hierarchicznych), jeżeli rezultaty grupowania dają się przedstawić w postaci drzewa połączeń zgodnego z kolejnością podobieństwa obiektów zbioru, 6) powtarzania obiektów – dodanie kilku obiektów identycznych do już występujących w zbiorze danych, nie zmienia granicy wyróżnionych klas, 7) powtarzania grup – gdy dodanie obiektów identycznych (z już występującymi) w ramach jednego skupienia nie zmienia wyniku klasyfikacji, tzn. obiekty zostaną przypisane do tego samego skupienia, 8) opuszczania grup – po usunięciu w całości wybranego skupienia, nie zmieni się wynik grupowania, tzn. wszystkie pozostałe obiekty ponownie zostaną przydzielone do tych samych skupień, 9) monotoniczności – monotoniczna transformacja macierzy odległości nie zmienia wyników grupowania.

Tabela 1.2. Wybrane własności niektórych metod aglomeracyjnych grupowania obiektów.

Metoda aglomeracyjna	Własność:					
	wypukłości	poprawnej struktury grup	poprawnej struktury połączeń drzewa	powtarzania obiektów	opuszczania grup	monotoniczności
Pojedynczego połączenia	nie	tak	tak	tak	tak	tak
Pełnego połączenia	nie	tak	tak	tak	tak	tak
Średniej klasowej	nie	tak	tak	nie	tak	nie
Warda	tak	nie	tak	nie	tak	nie
Środka ciężkości	nie	nie	nie	nie	tak	nie

Źródło: Gordon (1999), Everitt i in. (2001), Mikulec (2010), Pociecha (1982).

Wykorzystanie własności formalnych poszczególnych procedur do wyboru najlepszej wymaga jednak ich adaptacji do badanego zbioru danych. Nie wszystkie własności podane w tabeli 1.2 są jednakowo ważne i nie wszystkimi musi cechować się stosowana metoda aglomeracyjna. Kierując się cytowanymi wynikami badawczymi, w niniejszej monografii, spośród metod aglomeracyjnych stosowane będą metody: Warda, pełnego połączenia i średniej klasowej.

Z powodu ograniczonych możliwości stosowania podejścia drugiego (opartego na macierzy danych) oraz nie dającego całkowitej gwarancji wyboru najlepszej metody grupowania podejścia pierwszego (niezależnego od danych) proponuje się trzeci sposób wyboru metod analizy skupień – strategię grupowania (por. *consensus trees, Gordon,* 1999). Ten pomysł polega na zastosowaniu różnych metod grupowania obiektów a następnie porównania uzyskanych wyników w celu wyboru rezultatu najlepszego bądź uogólnienia uzyskanych wyników. To podejście różni się nieco od dwóch poprzednich : po pierwsze, wykorzystuje się zbiór danych; po drugie, przeprowadza się kolejne wariantów grupowania dla różnej liczby skupień; po trzecie, nacisk kładzie się na ocenę uzyskanego wyniku. Podejście takie stało się popularne dzięki rozwojowi techniki komputerowej oraz dostępności oprogramowania. Mając na uwadze to wielowariantowe podejście do badanego zbioru danych analiza

skupień jest również traktowana jako najbardziej rozwinięta forma analizy kombinatorycznej (*combinatorial data analysis)*.

Metody grupowania obiektów można podzielić na kilka rodzajów (por. *Gatnar i Walesiak ,* 2004), najpopularniejsze z nich to: metody partycjonujące (podziałowe) wśród których można wyróżnić metody obszarowo-gęstościowe oraz metody optymalizujące wstępny podział zbioru danych, metody aglomeracyjne (bądź deglomeracyjne) oraz metody wizualizacji danych.

Obszerne opracowanie wyników symulacji *Monte Carlo* dla procedur aglomeracyjnych – dokonanych przez kilkunastu różnych autorów posługujących się różnej wielkości zbiorami danych pod względem liczby analizowanych cech i obiektów – w zakresie ich zdolności do wykrywania struktury skupień wygenerowanych zbiorów danych, przy występowaniu różnych czynników mogących zakłócać tę strukturę, zawiera praca Milligana i Coopera (1987). Przegląd metod wskazuje, że spośród metod aglomeracyjnych najwyżej ocenione zostały procedury pełnego wiązania, średniej grupowej, oraz Warda, przy czym na pierwszym miejscu zdecydowanie najczęściej wymieniana była metoda Warda.

Dla metod podziałowych oprócz wyboru algorytmu bardzo istotny jest dobór punktów startowych, a więc określenie k obiektów w zbiorze danych inicjujących algorytm, których wybór, na ogół, ma zasadnicze znaczenie dla uzyskanego wyniku. W przypadku metody *k*-średnich najnowsze badania (*Steinley i Brusco,* 2007) wskazują jako najlepsze rozwiązania metodę aglomeracyjną Warda, którą można zastosować do wybrania punktów startowych oraz metodę wielokrotnego losowego doboru punktów startowych i przyjęciu wyniku minimalizującego sumę kwadratów odchyleń wartości zmiennych od centrów skupień (por. wzór (1.15). Ciekawym rezultatem jest to, że lepsze wyniki uzyskuje się dzięki połączeniu metod różnych rodzajów.

Należy jednak zauważyć, że wszystkie cytowane badania przeprowadzone były przy założeniu znanej (i poprawnej) liczby skupień. Wyniki tracą więc nieco na wartości w przypadku grupowania obiektów empirycznych zbiorów danych, gdy poprawnej liczby skupień nie znamy.

Etap ustalający liczbę skupień w zbiorze danych znajduje się przed etapem grupowania obiektów ale do jego przeprowadzenia na ogół konieczne jest uzyskanie podziału zbioru na wszystkie początkowe (tj. do pewnej liczby np. 20 skupień) liczby skupień, gdyż większość znanych indeksów liczby skupień ma charakter optymalizacyjny tzn. wskazuje na optymalną liczbę skupień dla danej metody grupowania. Wśród najczęściej stosowanych wymienić należy indeksy; Bakera-Huberta, Calińskiego–Harabasza, Dunna, Daviesa-Bouldina, Hartigana, Huberta-Levine'a, Krzanowskiego-Lai, indeks sylwetkowy indeks gap. Osobną grupę tworzą indeksy opracowane tylko pod kątem metod aglomeracyjnych np.

indeks Mojeny (1977). Dla metod aglomeracyjnych, Sokołowski (1992) wyróżnia aż pięć różnych grup indeksów liczby skupień.

Efektywność wymienionych i innych indeksów badana była przez wielu autorów m.in. *Milligan i Cooper* (1985), *Migdał-Najman i Najman* (2005), *Korzeniewski* (2005) Wybór właściwych indeksów służących do oceny liczby skupień nie jest łatwy. Spośród wymienionych najlepszą opinią cieszą się indeks gap, Daviesa-Bouldina, Calińskiego–Harabasza. Słabsze od nich okazały się indeksy Dunna i Hartigana. Jak stwierdzają sami wynalazcy indeksów, niektóre z indeksów mogą okazać się nieefektywne, w przypadku niektórych zbiorów danych. Dlatego też, nie należy wyników badań traktować z całkowitą pewnością, a jednym z często stosowanych w praktyce rozwiązań jest wykorzystanie wskazań kilku indeksów z następną syntezą ich wskazań (*Gordon,* 1999). Odnotujmy wzory wykorzystywane przy obliczaniu indeksu sylwetkowego, gdyż zostaną one później użyte przy walidacji efektów analizy skupień. Wartość indeksu sylwetkowego *i*-ego obiektu dana jest wzorem

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad , \qquad (1.11)$$

gdzie wielkość $a(i) = 1/(n_k - 1) \sum_{j \in C_k \; j \neq i} d(i, j)$ to średnia odległość obiektu i od pozostałych obiektów należących do skupienia C_k (które zawiera obiekt *i*); zaś wielkość $b(i) = \min_{C_r \neq C_k} d(i, C_r)$, gdzie $d(i, C_r) = (1/n_r) \sum_{q \in C_r} d(i, q)$, można określić jako minimalną średnią odległość obiektu $i \in C_k$ od pozostałych skupień.

W literaturze etap oceny wyników grupowania jest określany jako walidacja *(validation)*, przy czym stosowane są w nim różne podejścia do analizy wyników, opierające się na: 1) testowaniu losowości obiektów (czyli braku struktury klas); 2) testowaniu kompletności struktury grup; 3) ocenie poszczególnych skupień; 4) ocenie całego wyniku grupowania; 5) ocenie struktury klasyfikacji hierarchicznej (*Gordon,* 1999). Na przykład w zakresie oceny poprawności poszczególnych skupień w ujęciu stochastycznym może być wykorzystana wartość statystyki *U* Manna-Withneya (por. *Domański,* 1979) i symulacje *Monte Carlo*. Z kolei w testowaniu uzyskanego wyniku grupowania z punktu widzenia braku struktury klas, w konstrukcji hipotezy zerowej wykorzystywane są trzy modele (por. np. *Gordon,* 1999): rozkład Poissona – w przypadku analizy czy obiekty reprezentowane przez punkty w przestrzeni *V*-wymiarowej są ułożone równomiernie w pewnym jej obszarze; rozkład jednomodalny – zakładający, że badane obiekty pochodzą właśnie z takiego *V*-wymiarowego rozkładu, tzn. tworzą jedno skupienie oraz analiza losowości

zbioru, np. macierzy odległości – dla oceny czy odległości pomiędzy obiektami są losowe, tzn. czy elementy dolnego trójkąta macierzy odległości są uporządkowane losowo. Warto zauważyć, że do oceny braku struktury klas można wykorzystać indeks przerwy (*gap index, Tibshirani i inni,* 2001), który został skonstruowany z myślą o ocenie liczby skupień w zbiorze danych ale ma tę własność, że można go stosować gdy ta liczba jest równa zeru.

W praktyce, najczęściej oceny wyniku grupowania dokonuje się za pomocą odpowiednio dobranych miar jakości klasyfikacji oraz przy pomocy replikacji klasyfikacji.

Replikacja, czyli wielokrotne powtórzenie klasyfikacji może być traktowana jako sprawdzanie krzyżowe wyniku *(cross-validation)*, dotyczy bowiem badania jak daleko identyfikacja skupień na podstawie dwóch podprób wylosowanych z analizowanego zbioru danych odpowiada ostatecznemu wynikowi grupowania, tj. przynależności obiektów do skupień otrzymanej na podstawie całego zbioru danych. W literaturze przedmiotu można znaleźć kilka propozycji miar oceny zgodności wyników grupowania (*Denoeud* i inni, 2005): indeks oparty na CER, indeks Randa, skorygowany indeks Randa, indeks Jaccarda, Wallace'a, Lermana, Fowlkesa i Mallowsa (dla metod hierarchicznych) czy wskaźnik Nowaka (1985).

Najprostszym indeksem zgodności dwóch podziałów jest chyba indeks oparty na CER (*classification error rate*) czyli odsetku błędnych klasyfikacji. Formuła tego indeksu to suma liczby par obiektów przypisanych do tego samego skupienia lub do różnych skupień w obu podziałach (czyli $t_1 + t_2$ por. (1.15)) odniesiona do liczby wszystkich par. Wartość tego indeksu jest równa wartości indeksu Randa (nieskorygowanego). Wyniki badań empirycznych dla kolejnych pięciu wymienionych indeksów w zakresie ich poprawności (dokładności) dokonane przez Denoeuda i in. wskazały na podobne zachowanie się indeksu Jaccarda, skorygowanej miary Randa, Wallace'a i Lermana. Natomiast przy szczegółowej interpretacji ich wyników spośród wymienionych miar najbardziej prawidłowymi, stabilnymi okazały się miary Jaccarda i Wallace'a, a trzecią w kolejności skorygowana miara Randa. Skorygowany indeks Randa ze względu na swoją uniwersalność oraz powszechność jego stosowania a także to, iż jest on podstawą działania innych metod, np. metody *HINoV,* będzie wykorzystywany w niniejszej pracy.

Skorygowany indeks Randa (por. *Hubert i Arabie*, 1985), dla dwóch różnych podziałów P_1, P_2 zbioru danych ma postać:

$$RI(P_1, \mathrm{P}_2) = \frac{\binom{n}{2}(t_1 + t_2) - \left[(t_1 + t_3)(t_1 + t_4) + (t_3 + t_2)(t_4 + t_2)\right]}{\binom{n}{2}^2 - \left[(t_1 + t_3)(t_1 + t_4) + (t_3 + t_2)(t_4 + t_2)\right]} \tag{1.12}$$