EKONOMETRIA ECONOMETRICS 34 • 2011

ISSN 1507-3866

Justyna Wilk
Uniwersytet Ekonomiczny we Wrocławiu

KLASYFIKACJA DANYCH SYMBOLICZNYCH W ŚRODOWISKU R[1]

Streszczenie: Celem artykułu jest zaprezentowanie możliwości programu R w zakresie klasyfikacji danych symbolicznych. W pierwszej części artykułu przedstawiono podstawowe pojęcia analizy danych symbolicznych, problemy decyzyjne i metody statystyczne w procesie klasyfikacji danych symbolicznych. W części drugiej wskazano specyfikę programu R oraz pakiety i funkcje, które mają zastosowanie w klasyfikacji danych symbolicznych. W ostatniej części zaprezentowano przykład klasyfikacji danych symbolicznych w środowisku R.

Słowa kluczowe: analiza skupień, klasyfikacja danych symbolicznych, program R.

1. Wstęp

Klasyfikacja jest przedmiotem wielu badań, m.in. z zakresu ekonomii, medycyny, chemii, informatyki itd. Interesującym kierunkiem rozwoju metod klasyfikacji jest analiza danych symbolicznych (ADS). Mianem danych symbolicznych określa się przede wszystkim dane w formie przedziałów liczbowych, zbiorów kategorii oraz struktur udziałowych. W porównaniu z danymi w ujęciu klasycznym, reprezentowanymi przez pojedynczą kategorię lub wartość liczbową, charakteryzują się one złożoną strukturą. Z tego względu w klasyfikacji danych symbolicznych zastosowanie znajdują wybrane metody statystyczne. Użyteczność metodologii ADS w badaniach empirycznych jest w dużym stopniu determinowana dostępnością oprogramowania komputerowego.

Jednym z najważniejszych na rynku i coraz bardziej popularnym w Polsce pakietem statystycznym jest program R. Możliwości jego zastosowania w klasyfikacji danych symbolicznych są znacznie większe niż w przypadku innych programów stosowanych w ADS (np. SPSS, STATISTICA i SODAS). Celem artykułu jest zaprezentowanie możliwości programu R w zakresie klasyfikacji danych symbolicznych.

W pierwszej części artykułu przedstawiono podstawowe pojęcia ADS, problemy decyzyjne oraz metody statystyczne w procesie klasyfikacji danych symbolicz-

[1] Praca naukowa finansowana ze środków na naukę w latach 2009-2012 jako projekt badawczy nr N N111 446037.

nych. W części drugiej wskazano specyfikę programu R oraz pakiety i funkcje, które mają zastosowanie w klasyfikacji obiektów opisanych zmiennymi symbolicznymi. W ostatniej części zaprezentowano przykład klasyfikacji danych symbolicznych w środowisku R.

2. Podstawowe zagadnienia w klasyfikacji danych symbolicznych

Specyfika danych symbolicznych. W ADS wyróżnia się obiekty i zmienne symboliczne. Zbiór realizacji zmiennych symbolicznych dla obiektów tworzy tablicę danych symbolicznych. Wśród zmiennych symbolicznych wymienia się, oprócz metrycznych i niemetrycznych, zmienne (por. [Bock 2000]):

– interwałowe (*interval-valued variable*) – realizacjami tych zmiennych są przedziały wartości ciągłych ze zbioru liczb rzeczywistych,
– wielowariantowe (*multivalued variable*) – zbiory kategorii (równorzędnych lub uporządkowanych), wartości skokowych bądź przedziałów liczbowych,
– udziałowe (*modal variable*) – zbiory kategorii z przypisanymi indeksami wagowymi, prawdopodobieństwami, częstościami lub udziałami procentowymi,
– strukturalne (*dependent variable*) – w sytuacji, gdy między zmiennymi występują logiczne powiązania (taksonomiczne, hierarchiczne lub funkcyjne).

Ze względu na stopień agregacji danych można wyróżnić (por. [Bock 2000]):

– obiekty symboliczne I rzędu (*first order symbolic objects*) – obiekty w ujęciu klasycznym, tj. elementarne jednostki badania,
– obiekty symboliczne II rzędu (*second order symbolic objects*) – obiekty złożone (wtórne), będące wynikiem agregacji zbioru obiektów w ujęciu klasycznym bądź opisu klas uzyskanego z wykorzystaniem techniki CLINT.

Klasyfikacja danych symbolicznych. Klasyfikacja polega na podziale zbioru obiektów na względnie jednorodne klasy na podstawie zestawu kryteriów. Jest złożonym procesem, którego wyniki zależą od wyborów dokonanych w każdym etapie. Klasyfikacja obiektów opisanych zmiennymi symbolicznymi ze względu na ich nietypową strukturę wymaga zastosowania specyficznych metod statystycznych. Typowa procedura klasyfikacyjna obejmuje następujące kroki (por. [Walesiak 2004]):

1. Wybór obiektów oraz opisujących je zmiennych.
2. Określenie liczby klas oraz grupowanie obiektów[2].
3. Ocena wyników klasyfikacji.
4. Interpretacja (opis) oraz profilowanie klas.

W zależności od celu badania należy wskazać jednostkę badawczą; w ADS obiekty mogą mieć charakter pierwotny lub wtórny. Jeśli badaniem obejmuje się tylko część populacji, to należy również ustalić strukturę i liczebność próby. Następnie, na podstawie wiedzy merytorycznej, określa się zestaw zmiennych, na podstawie

[2] Jeśli metoda analizy skupień bazuje na macierzy odległości, należy wykonać pomiar odległości obiektów.

których przeprowadzona zostanie klasyfikacja. W uzasadnionych sytuacjach w selekcji zmiennych stosuje się również algorytmy formalne. Dla zmiennych symbolicznych zastosowanie mają m.in. metoda grafowa Ichino i adaptacja metody *HINoV* Carmone'a, Kary i Maxwell.

Wybór liczby klas jest podyktowany wiedzą merytoryczną bądź wsparty metodami formalnymi. W ADS zastosowanie mają indeksy bazujące na macierzy odległości (np. indeks Bakera i Huberta) oraz tablicy danych symbolicznych (np. indeks $Q(P)$ Verde, Lechevalliera i Chavent). Wielość i różnorodność algorytmów analizy skupień powoduje, że nie można wskazać metody uniwersalnej dla wszystkich problemów badawczych. Procedury hierarchiczne dają w wyniku hierarchię klas, którą uzyskuje się w drodze aglomeracji bądź deglomeracji zbioru obiektów. Metody niehierarchiczne dokonują podziału ze względu na przyjęte kryterium jakości podziału. Do popularnych procedur niehierarchicznych należą metody optymalizacyjne, które dzielą zbiór obiektów na zadaną liczbę klas. W klasyfikacji danych symbolicznych zastosowanie mają metody taksonomii numerycznej bazujące na macierzy odległości oraz metody taksonomii symbolicznej (por. [Wilk 2010; 2011]). Jeśli metoda analizy skupień bazuje na macierzy odległości, to przed klasyfikacją należy dokonać pomiaru odległości obiektów. Wybór miary odległości zależy od charakteru zmiennych opisujących obiekty. W ADS zastosowanie mają m.in. miary zaproponowane przez Ichino i Yaguchiego oraz de Carvalho.

Uzyskany podział zbioru obiektów poddaje się weryfikacji formalnej, aby określić, na ile wyniki klasyfikacji odwzorowują rzeczywistą strukturę zjawiska. Wśród metod oceny klasyfikacji danych symbolicznych można wskazać indeks *silhouette* Rousseeuwa, adaptację analizy replikacji z indeksem Randa oraz metodę Bertranda i Bel-Mufti.

W wielu badaniach, oprócz ustalenia liczby i liczebności klas oraz przynależności obiektów do klas, istotne jest również rozpoznanie cech charakterystycznych i czynników różnicujących klasy. Klasy interpretuje się na podstawie zmiennych biorących udział w grupowaniu obiektów. W tym celu w ADS stosuje się technikę CLINT. W profilowaniu uczestniczą natomiast zmienne, które nie brały udziału w klasyfikacji i zastosowanie mają metody wielowymiarowej analizy statystycznej, m.in. drzewa klasyfikacyjne.

3. Pakiety i funkcje programu R w klasyfikacji danych symbolicznych

R to środowisko do obliczeń statystycznych i jednocześnie język programowania działający w tym środowisku. Jest on programem bezpłatnym (również do zastosowań komercyjnych) na licencji GNU/GPL i może działać w systemach operacyjnych Linux, Windows i MacOS. Ma tekstowy interfejs, dlatego jego użytkowanie wymaga posiadania przynajmniej podstawowej wiedzy z zakresu programowania. Udostępnia otwarty kod źródłowy, dający możliwość modyfikacji procedur i tworzenia

własnych programów. Korzystanie z oprogramowanych metod wymaga załadowania pakietów zawierających odpowiednie funkcje (por. [Dudek 2009, s. 13-14]).

W programie R można przeprowadzić kompletną procedurę klasyfikacji danych symbolicznych. Pakietem dedykowanym ADS jest symbolicDA. Wiele funkcji z tego zakresu zawiera również pakiet clusterSim. Zastosowanie mają także inne biblioteki, np. cluster i stats. Pakiety i funkcje programu R, które można wykorzystać w procesie klasyfikacji danych symbolicznych, zaprezentowano w tab. 1.

Tabela 1. Pakiety i funkcje programu R w klasyfikacji danych symbolicznych

Wyszczególnienie	Pakiet	Funkcja	Uwagi
Zbiór danych symbolicznych	clusterSim	data	generowanie zbioru obiektów symbolicznych
	symbolicDA	generate.SO	
		parse.SO	wczytanie tablicy danych symbolicznych
Metody selekcji zmiennych	clusterSim	HINoV.Symbolic	dla obiektów opisanych zmiennymi interwałowymi
	symbolicDA	HINoV.SDA	dla obiektów opisanych dowolnym rodzajem zmiennych symbolicznych
		IchinoFS.SDA	
Indeksy wyboru liczby klas	clusterSim	index.G2	indeks Bakera i Huberta
		index.G3	indeks Huberta i Levine
	symbolicDA	index.G1d	adaptacja indeksu Calińskiego i Harabasza
Miary odległości	symbolicDA	dist.SDA	dla obiektów opisanych zmiennymi interwałowymi
	clusterSim	dist.Symbolic	dla obiektów opisanych dowolnym rodzajem zmiennych symbolicznych
Metody analizy skupień	cluster	diana	deglomeracyjna metoda Macnaughtona-Smitha
		pam	optymalizacyjna metoda *k*-medoidów
		agnes	metody aglomeracyjne
	stats	hclust	
	symbolicDA	DClust	metody optymalizacyjne
		SClust	
Metody oceny klasyfikacji	clusterSim	index.S	indeks *silhouette* Rousseeuwa
	symbolicDA	replication.SDA	adaptacja analizy replikacji z indeksem Randa
Metody opisu klas	symbolicDA	cluster.Description.SDA	technika CLINT
		.zoomStar	wykres rozgwiazdy
Metody profilowania klas	symbolicDA	kernel.SDA	jądrowa analiza dyskryminacyjna
		decissionTree.SDA	algorytm TREE drzew klasyfikacyjnych

Źródło: opracowanie na podstawie pracy: [Wilk 2011].

Do wczytania tablicy danych symbolicznych[3] w formacie xml służy funkcja parse.SO (tab. 2). Zbiór obiektów symbolicznych o zadanej strukturze klas można wygenerować z wykorzystaniem funkcji generate.SO lub data.

Tabela 2. Składnie funkcji służących do wczytania lub generowania zbioru danych symbolicznych

1. Wczytanie tablicy danych symbolicznych w formacie xml	
parse.SO(file)	
file	nazwa pliku (bez rozszerzenia xml) zawierającego tablicę danych symbolicznych
2. Generowanie zbioru obiektów symbolicznych o zadanej strukturze klas	
data(data_symbolic)	
data_symbolic	zbiór 125 obiektów opisanych 6 zmiennymi interwałowymi, o strukturze 5 klas
generate.SO(numObjects, numClusters, numIntervalVariables, numMultivaluedVariables)	
numObjects	wektor wskazujący liczbę obiektów w poszczególnych klasach
numClusters	liczba klas
numIntervalVariables	liczba zmiennych interwałowych
numMultivaluedVariables	liczba zmiennych wielowariantowych

Źródło: opracowanie na podstawie dokumentacji programu R.

Tabela 3. Składnie funkcji metod selekcji zmiennych symbolicznych

1. Adaptacja metody *HINoV*	
HINoV.Symbolic(x, u=NULL, distance="H", method="pam", Index="cRAND")	
x	tablica danych symbolicznych (zbiór obiektów opisanych zmiennymi interwałowymi)
u	liczba klas
distance	miara odległości: „M", „H", „S"
method	metoda klasyfikacji: „single", „ward", „complete", „average", „mcquitty", „median", „centroid", „pam" (domyślnie)
Index	"cRAND" – skorygowany indeks Randa (domyślnie), „RAND" – indeks Randa
HINoV.SDA(table.Symbolic, u=NULL, distance="H", method="pam", Index="cRAND")	
table.Symbolic	tablica danych symbolicznych
distance	miara odległości: "U_2", „U_3", „U_4", „C_1", „SO_1", „SO_2", „SO_3", „SO_4", „SO_5", „L_1", „L_2"
2. Metoda grafowa Ichino	
IchinoFS.SDA(table.Symbolic)	

Źródło: opracowanie na podstawie dokumentacji programu R.

[3] Do tworzenia i edycji tablic danych symbolicznych w formacie xml przeznaczony jest program SDAEditor A. Dudka, który można pobrać ze strony http://wgrit.ae.jgora.pl/keii/sdaeditor/index.html.

W pakietach symbolicDA (funkcja HINoV.SDA) i clusterSim (funkcja HINoV. Symbolic) dostępna jest adaptacja metody *HINoV* służącej selekcji zmiennych przy zadanej liczbie klas (tab. 3). W pakiecie symbolicDA oprogramowano również grafową metodę Ichino (funkcja IchinoFS.SDA).

Charakterystykę trzech indeksów służących wyborowi liczby klas obiektów symbolicznych zawiera tab. 4. Wskazane indeksy bazują na macierzy odległości wyznaczonej dla zbioru obiektów symbolicznych.

Tabela 4. Składnie funkcji indeksów wyboru liczby klas obiektów symbolicznych

1. Indeks Bakera i Huberta	
`index.G2(d, cl)`	
`d`	macierz odległości
`cl`	wektor informujący o przynależności obiektów do klas
2. Indeks Huberta i Levine	
`index.G3(d, cl)`	
3. Adaptacja pseudostatystyki *F* Calińskiego i Harabasza	
`index.G1d(d, cl)`	

Źródło: opracowanie własne na podstawie dokumentacji programu `R`.

Tabela 5. Składnie funkcji miar odległości obiektów symbolicznych

`dist.Symbolic(data, type="U_2", gamma=0.5, power=2)`	
`data`	tablica danych symbolicznych
`type`	miara odległości: „M", „H", „S", `U_2"`
`gamma`	parametr określany dla miary `U_2`
`power`	parametr *q* określany dla miary `U_2`
`dist.SDA(table.Symbolic, type="U_2", subType=NULL, gamma=0.5, power=2, probType="J", probAggregation="P_1", s=0.5, p=2, variableSelection=NULL, weights=NULL)`	
`table.Symbolic`	tablica danych symbolicznych
`type`	miara odległości: „U_2", „U_3", „U_4", „C_1", „SO_1", „SO_2", „SO_3", „SO_4", „SO_5", „L_1", „L_2"
`subType`	funkcja porównująca dla miar `C_1` oraz `SO_1`: „D_1", „D_2", „D_3", „D_4", „D_5"
`probType`	składowa miara odległości: "J", „CHI", „REN", „CHER", „LP"
`probAggregation`	miara agregatowa dla odległości składowych: „P_1", „P_2"
`s`	parametr dla miary Renyi `(REN)`
`p`	parametr dla metryki Minkowskiego (`LP`)
`variableSelection`	wektor wskazujący numery zmiennych uwzględnionych podczas obliczania odległości obiektów
`weights`	wagi zmiennych w metryce Minkowskiego

Źródło: opracowanie na podstawie dokumentacji programu `R`.

Pomiaru odległości obiektów symbolicznych można dokonać z wykorzystaniem funkcji `dist.SDA` oraz `dist.Symbolic`. Wybór miary odległości jest determinowany rodzajem zmiennych symbolicznych opisujących zbiór obiektów (zob. tab. 5):

– zmienne interwałowe – miary odległości: `M`, `H`, `S`,
– zmienne interwałowe i wielowariantowe – miary odległości: `U_2`, `U_3`, `U_4`, `C_1`, `SO_1`, `SO_2`, `SO_3`, `SO_4`, `SO_5`,
– zmienne interwałowe, wielowariantowe i udziałowe – miary odległości: `L_1`, `L_2`,
– zmienne udziałowe – składowe miary odległości: `J`, `CHI`, `REN`, `CHER`, `LP` oraz agregatowe miary odległości: `P_1, P_2`.

Metody analizy skupień mające zastosowanie w ADS znajdują się w pakietach `stats`, `cluster` i `symbolicDA` (tab. 6).

Tabela 6. Składnie funkcji metod klasyfikacji danych symbolicznych*

1. Metody hierarchiczne	
1.1. Metody aglomeracyjne	
`hclust(d, method="complete")`	
`d`	macierz odległości
`method`	metoda: „`single`", „`complete`", „`ward`", „`average`", „`mcquitty`", „`median`", „`centroid`"
`agnes(x, diss=TRUE, method="average", par.method)`	
`x`	macierz odległości
`diss`	opcja wymagana w przypadku danych symbolicznych: `TRUE`
`method`	metoda klasyfikacji: „`average`", „`single`", „`complete`", „`ward`", „`weighted`", „`flexible`"
`par.method`	wektor zawierający parametry dla metody „`flexible`" (1, 3 lub 4)
1.2. Metoda deglomeracyjna (metoda Macnaughtona-Smitha)	
`diana(x, diss=TRUE)`	
2. Metody optymalizacyjne	
`pam(x, k, diss=TRUE, medoids=NULL, trace.lev=0)`	
`k`	liczba klas
`medoids`	`NULL` (domyślnie) lub wektor zawierający początkowe zalążki klas
`SClust(table.Symbolic, cl, iter=100, variableSelection=NULL, objectSelection=NULL)`	
`table.Symbolic`	tablica danych symbolicznych
`cl`	liczba klas lub wektor zawierający początkowe zalążki klas
`iter`	maksymalna liczba iteracji
`variableSelection`	`NULL` (domyślnie) dla wszystkich zmiennych lub wektor wskazujący numery zmiennych uwzględnionych podczas klasyfikacji obiektów
`objectSelection`	`NULL` (domyślnie) dla wszystkich obiektów lub wektor wskazujący numery obiektów uwzględnionych podczas klasyfikacji obiektów
`DClust(dist,cl,iter=100)`	
`dist`	macierz odległości

* Opis niektórych funkcji zawiera podstawowe argumenty.

Źródło: opracowanie własne na podstawie dokumentacji programu R.

Składnie funkcji metod stosowanych w ocenie wyników klasyfikacji zaprezentowano w tab. 7. Indeks *silhouette* Rousseeuwa bazuje na macierzy odległości, natomiast analizę replikacji przeprowadza się, bazując na tablicy danych symbolicznych.

Tabela 7. Składnie funkcji metod oceny wyników klasyfikacji*

1. Indeks *silhouette* Rousseeuwa	
`index.S(d, cl)`	
`d`	macierz odległości
`cl`	wektor informujący o przynależności obiektów do klas
2. Adaptacja analizy replikacji z indeksem Randa	
`replication.SDA(table.Symbolic, u=2, method="SClust", S=10, fixedAsample=NULL)`	
`table.Symbolic`	tablica danych symbolicznych
`u`	liczba klas
`method`	metoda zastosowana w klasyfikacji: „`SClust`"
`S`	liczba symulacji
`fixedAsample`	numery obiektów dobrane losowo do podzbioru `A` (`NULL`) lub numery obiektów dobrane arbitralnie do podzbioru `A`

* Opis niektórych funkcji zawiera podstawowe argumenty.

Źródło: opracowanie własne na podstawie dokumentacji programu `R`.

Do wyznaczenia charakterystyk poszczególnych klas obiektów symbolicznych można wykorzystać funkcje `cluster.Description.SDA` oraz `.zoomStar` z pakietu `symbolicDA`. Funkcja `cluster.Description.SDA` służy do opisu klas metodą CLINT. Dla zmiennej interwałowej określa się najmniejszy przedział liczbowy, obejmujący wszystkie przedziały realizacji zmiennej charakteryzujące obiekty należące do klasy. Z kolei dla zmiennych wielowariantowych wyznacza się zbiór kategorii, w skład którego wchodzą wszystkie kategorie zmiennej charakteryzujące obiekty reprezentujące klasę. W przypadku zmiennych udziałowych wyznaczane są minimalne, maksymalne i średnie wagi uzyskiwane przez obiekty dla każdej kategorii zmiennej oraz suma tych wag.

Za pomocą funkcji `.zoomStar` można wykonać wykres rozgwiazdy, na którym zaprezentowane zostaną realizacje zmiennych zaobserwowane dla obiektów danej klasy.

Składnie funkcji metod profilowania klas obiektów symbolicznych oprogramowanych w pakiecie `symbolicDA` zostały scharakteryzowane w tab. 9.

Tabela 8. Składnie funkcji służących interpretacji klas obiektów symbolicznych*

1. Technika opisu klas obiektów symbolicznych CLINT	
`cluster.Description.SDA(table.Symbolic, clusters)`	
`table.Symbolic`	tablica danych symbolicznych
`clusters`	wektor wskazujący przynależność obiektów do klas
2. Wizualizacja charakterystyk klas obiektów symbolicznych – wykres rozgwiazdy	
`.zoomStar(table.Symbolic, j, variableSelection=NULL)`	
`j`	`.medoid(d, classes, i)`, przy czym `d` oznacza macierz odległości, `classes` – wektor wskazujący przynależność obiektów do klas, `i` – numer klasy pokazywanej na wykresie
`variableSelection=NULL`	`NULL` (domyślnie) dla wszystkich zmiennych lub wektor wskazujący numery zmiennych pokazywanych na wykresie

* Opis niektórych funkcji zawiera podstawowe argumenty.

Źródło: opracowanie własne na podstawie dokumentacji programu R.

Tabela 9. Składnie funkcji służących profilowaniu klas obiektów symbolicznych*

1. Jądrowa analiza dyskryminacyjna	
`kernel.SDA(sdt, formula, testSet, h, type)`	
`sdt`	tablica danych symbolicznych
`formula`	zob. funkcja `lm`
`testSet`	wektor zawierający numery obiektów zakwalifikowane do zbioru testowego
`h`	szerokość pasma jądra
`type`	miara odległości: "`U_2`", "`U_3`", "`U_4`", "`C_1`", "`SO_1`", "`SO_2`", "`SO_3`", "`SO_4`", "`SO_5`"
2. Algorytm TREE drzew klasyfikacyjnych	
`decisionTree.SDA(sdt, formula, testSet)`	
`testSet`	wektor zawierający numery obiektów zakwalifikowane do zbioru uczącego

* Opis niektórych funkcji zawiera podstawowe argumenty.

Źródło: opracowanie własne na podstawie dokumentacji programu R.

4. Przykład

Zgromadzono dane charakteryzujące wybrane modele rodzinnych samochodów osobowych dostępne na polskim rynku, zaliczane do klasy niższej i średniej. Celem badania jest klasyfikacja samochodów na podstawie zestawu zmiennych charakteryzujących parametry techniczne i gabaryty. Może być ona szczególnie użyteczna dla potencjalnych nabywców, którzy mając pewne oczekiwania i preferencje, muszą dokonać wyboru samochodu. W analizie uwzględniono 30 obiektów symbolicznych

II rzędu, które powstały w wyniku agregacji obiektów I rzędu, np. „Seat Exeo" obejmuje wszystkie dostępne wersje silnika i nadwozia (tab. 10).

Tabela 10. Zbiór obiektów symbolicznych II rzędu

Lp.	Marka	Model	Lp.	Marka	Model	Lp.	Marka	Model
1	Škoda	Nowa Fabia	11	Toyota	Aygo	21	Chevrolet	Aveo
2	Škoda	Nowa Octavia	12	Toyota	Yaris	22	Chevrolet	Lacetti
3	Fiat	Panda	13	Toyota	Corolla	23	Seat	Ibiza
4	Fiat	Grande Punto	14	Toyota	Avensis	24	Seat	Leon
5	Fiat	Bravo	15	Opel	Corsa	25	Seat	Exeo
6	Peugeot	308	16	Opel	Astra	26	Honda	Jazz
7	Peugeot	407	17	Volkswagen	Nowe Polo	27	Honda	Civic 5D
8	Citroën	C1	18	Volkswagen	Golf	28	Honda	Accord Sedan
9	Citroën	Nowy C3	19	Volkswagen	Passat Limousine	29	Nissan	Micra
10	Citroën	C4	20	Chevrolett	Nowy Spark	30	Nissan	Tiida

Źródło: opracowanie własne.

Zbiór zmiennych symbolicznych opisujących obiekty zaprezentowano w tab. 11.

Tabela 11. Zbiór zmiennych symbolicznych

Lp.	Nazwa zmiennej symbolicznej	Rodzaj zmiennej symbolicznej	Zbiór realizacji zmiennej symbolicznej	Jednostka miary
1	cena	interwałowa	[27990, 144500]	zł
2	długość nadwozia	interwałowa	[3415; 4765]	mm
3	szerokość nadwozia	interwałowa	[1465; 2033]	mm
4	pojemność skokowa	wielowariantowa	{1,0; 1,1; 1,2; 1,3; 1,4; 1,6; 1,7; 1,8; 1,9; 2,0; 2,2; 2,4}	dm^3
5	moc silnika	interwałowa	[54; 270]	KM
6	maksymalna prędkość	interwałowa	[150; 247]	km/h
7	przyspieszenie 0-100 km/h	interwałowa	[6; 18]	s
8	rodzaj paliwa	wielowariantowa	{benzyna; diesel}	–

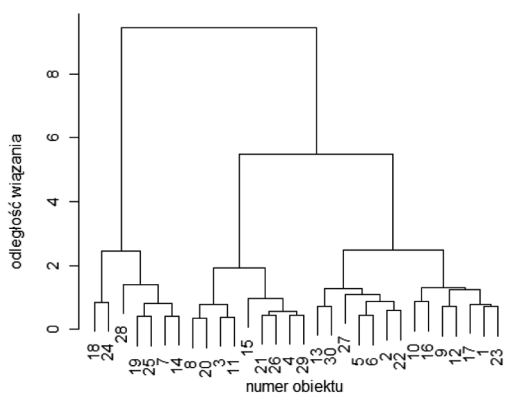
Źródło: opracowanie własne.

W programie R zastosowano następującą procedurę obliczeniową:

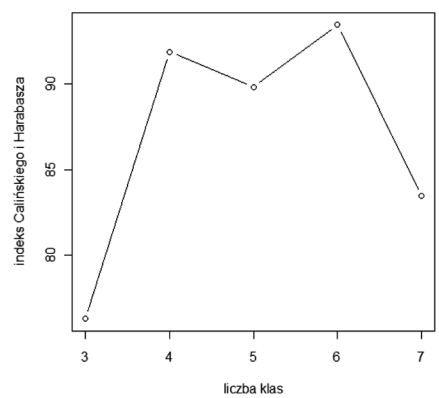
```
library(symbolicDA)
library(stats)
options(OutDec=",")
tds<-parse.SO(„samos")
d<-dist.SDA(tds, type="U_3", gamma=0.4, power=2)
```

```
k<-hclust(d, method="ward")
plot(k, main=NULL, sub="", xlab="numer obiektu", ylab="odległość
wiązania")
klasy3<-cutree(k, 3)
klasy4<-cutree(k, 4)
klasy5<-cutree(k, 5)
klasy6<-cutree(k, 6)
klasy7<-cutree(k, 7)
ch3<-index.G1d(d, klasy3)
ch4<-index.G1d(d, klasy4)
ch5<-index.G1d(d, klasy5)
ch6<-index.G1d(d, klasy6)
ch7<-index.G1d(d, klasy7)
x<-c(3, 4, 5, 6, 7)
y<-c(ch3, ch4, ch5, ch6, ch7)
plot(x, y, type="b", xlab="liczba klas", ylab="indeks Calińskiego
i Harabasza")
set.seed(123)
r<-replication.SDA(tds, u=6, method="ward", S=15)
print(r)
print(klasy6)
o<-cluster.Description.SDA(tds, klasy6)
print(o)
```

W pierwszej kolejności wyznaczono macierz odległości z wykorzystaniem normalizowanej miary odległości Ichino-Yaguchiego U_3 (funkcja `dist.SDA` z pakietu `symbolicDA`). Grupowanie obiektów symbolicznych przeprowadzono metodą Warda (funkcja `hclust` z pakietu `stats`). W wyniku otrzymano dendrogram klas (rys. 1a).



a) dendrogram klas otrzymany metodą Warda



b) wartości indeksu Calińskiego i Harabasza

Rys. 1. Dendrogram klas otrzymany metodą Warda

Źródło: opracowanie własne w programie R.

Na podstawie wzrokowej oceny wykresu jako potencjalną liczbę klas wskazano 3, 4, 5, 6, 7. Dla każdego podziału wyznaczono wartości indeksu Calińskiego i Harabasza[4] (funkcja `index.G1d` w pakiecie `symbolicDA`). Indeks osiągnął najwyższą wartość (93,5) w podziale obiektów na 6 klas (rys. 1b). Następnie przeprowadzono ocenę wyników klasyfikacji z wykorzystaniem analizy replikacji z indeksem Randa (funkcja `replication.SDA` w pakiecie `symbolicDA`). Indeks uzyskał wartość 0,48, co oznacza relatywnie dobrą stabilność klasyfikacji.

W ostatnim kroku sporządzono opis klas z wykorzystaniem techniki CLINT (funkcja `cluster.Description.SDA` z pakietu `symbolicDA`). Charakterystykę klas zaprezentowano w tab. 12.

Tabela 12. Charakterystyka klas

Numer klasy		1	2	3	4	5	6
Obiekty należące do klasy		Nowa Fabia, Nowy C3, C4, Ibiza, Nowe Polo, Yaris, Astra	Nowa Octavia, Bravo, 308, Corolla, Lacetti, Civic 5D, Tiida	Panda, C1, Aygo, Nowy Spark	Grande Punto, Corsa, Aveo, Jazz, Micra	407, Avensis, Exeo, Accord Sedan, Passat Limousine	Golf, Leon
Nazwa zmiennej symbolicznej	cena	[33500; 79500]	[43030; 106900]	[27990; 45990]	[35990; 64840]	[72900; 144500]	[54990;138770]
	długość nadwozia	[3785; 4515]	[4255; 4597]	[3415; 3640]	[3719; 4030]	[4661; 4765]	[4199; 4315]
	szerokość nadwozia	[1465; 2033]	[1470; 1815]	[1578; 1630]	[1660; 1944]	[1772; 1840]	[1768; 1779]
	pojemność skokowa*	1,1 (3); 1,2 (3); 1,3 (2); 1,4 (7); 1,6 (6); 1,7 (1); 1,9 (1)	1,2 (1); 1,4 (6); 1,6 (4); 1,8 (4); 1,9 (1); 2,0 (4); 2,2 (1)	1,0 (2); 1,1 (2); 1,2 (2)	1,0 (1); 1,2 (5); 1,3 (2); 1,4 (5)	1,4 (1); 1,6 (4); 1,8 (4); 2,0 (5); 2,2 (2); 2,4 (1)	1,2 (1); 1,4 (2); 1,6 (2); 1,8 (1); 1,9 (1); 2,0 (2)
	moc silnika	[60; 150]	[73; 165]	[54; 81]	[65; 101]	[102; 200]	[59; 270]
	maksymalna prędkość	[155; 210]	[173; 215]	[150; 164]	[150; 182]	[190; 241]	[172; 247]
	przyspieszenie 0-100 km/h	[6; 16]	[7; 14]	[12; 15]	[11; 18]	[6; 12]	[6; 13]
	rodzaj paliwa*	benzyna (7); diesel (6)	benzyna (7); diesel (7)	benzyna (4); diesel (1)	benzyna (5)	benzyna (4); diesel (4)	benzyna (2); diesel (2)

* Dla zmiennych wielowariantowych w nawiasie podano liczebność kategorii.

Źródło: opracowanie własne na podstawie wyników otrzymanych w programie `R`.

Do klasy 1 należą samochody o średnich gabarytach, przeznaczone głównie do jazdy w mieście. Dostępne są w różnych wersjach silnika, o mocy w granicach

[4] Indeks Calińskiego i Harabasza: $G1(u) \in R_+$, $\hat{u} = \arg\max_u \{G1(u)\}$, gdzie u – liczba klas.

60-150 KM i pojemności najczęściej 1,4 lub 1,6 dm^3. Są oferowane w relatywnie niskich cenach, w granicach 33,5-79,5 tys. zł. Skupienie 2 stanowią auta kompaktowe, zapewniające względny komfort jazdy dla 4 dorosłych osób. Ze względu na parametry techniczne są one przeznaczone do jazdy zarówno po mieście, jak i poza miastem. Stanowią rodzaj kompromisu pomiędzy autami małymi i dużymi, za umiarkowaną cenę (ok. 43-107 tys. zł).

Grupę 3 reprezentują auta miejskie, które, ze względu na najmniejsze gabaryty, można określić jako „mini" (długość nadwozia nie przekracza 3,6 m). Charakteryzują się najsłabszymi parterami silnika; moc silnika nie przekracza 81 KM, a maksymalna prędkość deklarowana przez producentów wynosi 164 km/h. Te parametry przekładają się na relatywnie niską cenę, nieprzekraczającą 46 tys. zł. Klasa 4 skupia samochody o średnich rozmiarach; mniejsze niż auta zaliczone do klasy 1, jednak większe niż auta klasy 3. Cechują się one nieco słabszymi parametrami technicznymi od aut z grupy 1, ale zbliżonym do nich przedziałem cenowym.

Do skupienia 5 zostały zakwalifikowane samochody rodzinne, o największych gabarytach, zapewniające w miarę komfortowe warunki podróżowania 5 dorosłym osobom na dłuższych dystansach. Ze względu na dobre osiągi techniczne oraz wielkość są oferowane w relatywnie wysokich cenach, przekraczających 70 tys. zł. Samochody zaliczone do klasy 6 charakteryzują średnie gabaryty i dosyć duże zróżnicowanie pod względem technicznym. W zależności od parametrów silnika mogą być użytkowane jako samochody kompaktowe (podobnie jak auta klasy 4) albo sportowe. Ich ceny są zbliżone do aut klasy 2 i klasy 5.

5. Podsumowanie

Program R daje bardzo duże możliwości w zakresie analizy skupień oraz ADS. Jest obecnie jedynym programem, w którym można zrealizować kompletną procedurę klasyfikacji danych symbolicznych. Atuty programu stanowi bezpłatny dostęp (w tym do celów komercyjnych), a także otwarty kod źródłowy, który ułatwia rozwój procedur i dostosowanie do potrzeb użytkownika. W przygotowaniu jest kolejny pakiet wspomagający klasyfikację danych symbolicznych `clamix`.

Literatura

Bock H.H., *Symbolic Data*, [w:] *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*, red. H.H. Bock, E. Diday, Springer-Verlag, Berlin-Heidelberg 2000.

Dudek A., *Wprowadzenie do programu R*, [w:] *Statystyczna analiza danych z wykorzystaniem programu R*, M. Walesiak, E. Gatnar (red.), PWN, Warszawa 2009.

Walesiak M., *Problemy decyzyjne w procesie klasyfikacji zbioru obiektów*, [w:] *Zastosowania metod ilościowych*, red. J. Dziechciarz, Ekonometria 13, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1010, AE, Wrocław 2004.

Wilk J., *Cluster Analysis Methods in Symbolic Data Analysis*, [w:] *Data Analysis Methods in Economic Investigations*, red. J. Pociecha, Studia i Prace UE w Krakowie nr 11, Kraków 2010.
Wilk J., *Analiza skupień na podstawie danych symbolicznych*, [w:] *Analiza danych jakościowych i symbolicznych z wykorzystaniem programu R*, E. Gatnar, M. Walesiak (red.), PWN, Warszawa 2011.

SYMBOLIC DATA CLASSIFICATION IN R ENVIRONMENT

Summary: The aim of this paper is to present the functionalities of R software in the area of symbolic data classification. In the first part of the article the basic concepts of symbolic data analysis, decision problems and statistical methods in symbolic data classification procedure were discussed. In the second part the R software, its packages and functions useful in symbolic data classification were characterized. In the last part of the paper an example of symbolic data classification by means of R software was presented.

Keywords: cluster analysis, symbolic data classification, R software.