PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu

RESEARCH PAPERS
of Wrocław University of Economics

Nr 327

Taksonomia 22
Klasyfikacja i analiza danych – teoria i zastosowania

Redaktorzy naukowi
Krzysztof Jajuga, Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2014

Redaktor Wydawnictwa: Barbara Majewska
Redaktor techniczny: Barbara Łopusiewicz
Korektor: Barbara Cibis
Łamanie: Beata Mazur
Projekt okładki: Beata Dębska

Publikacja jest dostępna w Internecie na stronach:
www.ibuk.pl, www.ebscohost.com,
w Dolnośląskiej Bibliotece Cyfrowej www.dbc.wroc.pl,
The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

Tytuł dofinansowany ze środków Narodowego Banku Polskiego
oraz ze środków Sekcji Klasyfikacji i Analizy Danych PTS

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS nr 327 • 2014
Taksonomia 22 ISSN 1899-3192
Klasyfikacja i analiza danych – teoria i zastosowania

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

KLASYFIKACJA POJĘCIOWA DANYCH SYMBOLICZNYCH W PODEJŚCIU WIELOMODELOWYM

Streszczenie: Celem artykułu jest zaproponowanie nowego podejścia w klasyfikacji wielomodelowej danych symbolicznych z wykorzystaniem klasyfikacji pojęciowej jako klasyfikatora bazowego. W artykule przedstawiono podstawowe pojęcia z zakresu analizy danych symbolicznych, podejścia wielomodelowego oraz klasyfikacji pojęciowej. W części empirycznej omówiono wyniki badań symulacyjnych dla sztucznych i rzeczywistych zbiorów danych.

Słowa kluczowe: analiza danych symbolicznych, klasyfikacja wielomodelowa, klasyfikacja pojęciowa.

1. Wstęp

Obiekty symboliczne, w odróżnieniu od obiektów w ujęciu klasycznym, mogą być opisywane przez wiele różnych typów zmiennych. Oprócz zmiennych w ujęciu klasycznym (metrycznych lub niemetrycznych) mogą być opisywane przez zmienne interwałowe, zmienne wielowariantowe i zmienne wielowariantowe z wagami, a także zmienne strukturalne [zob. np. Bock, Diday (red.) 2000, s. 2-3]. Pozwala to na dokładniejszy opis obiektów, ale utrudnia analizę skupień.

Podejście wielomodelowe było dotychczas z powodzeniem stosowane w zagadnieniach dyskryminacyjnych i regresyjnych [zob. np. Gatnar 2008]. Niemniej idea podejścia wielomodelowego, tj. łączenia wyników wielu modeli, może być z powodzeniem zastosowana także w zagadnieniu klasyfikacji danych symbolicznych. Podejście wielomodelowe w klasyfikacji to nic innego jak łączenie (czyli agregacja) wielu klasyfikacji (modeli) bazowych w jedną klasyfikację złożoną [por. Fred, Jain 2005].

Celem artykułu jest zaproponowanie nowego podejścia w klasyfikacji wielomodelowej danych symbolicznych, która wykorzystuje klasyfikację pojęciową dla tego typu danych jako klasyfikator bazowy. W części empirycznej przedstawiono wyniki badań symulacyjnych, w których zastosowano rzeczywiste i sztuczne zbiory danych symbolicznych.

2. Dane symboliczne w zagadnieniu klasyfikacji

Obiekty symboliczne mogą być opisywane zmiennymi symbolicznymi różnego typu [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday 2006, s. 7-30; Dudek 2013, s. 35-36]. W tabeli 1 zawarto najważniejsze typy zmiennych symbolicznych wraz z przykładami.

Tabela 1. Przykładowe zmienne symboliczne wraz z realizacjami

Zmienna	Realizacje	Typ zmiennej symbolicznej
Preferowana cena samochodu (w zł)	<27000, 42000>; <35000, 50000> <20000, 30000>; <25000, 37000>	interwałowa (przedziały nierozłączne)
Rozważana pojemność silnika (w cm^3)	<1000, 1200>; <1300, 1400> <1500, 1800>; <1900, 2200>	interwałowa (przedziały rozłączne)
Wybrany kolor	{niebieski, czerwony, żółty} {zielony, czarny, szary, biały}	wielowariantowa
Preferowana marka samochodu	{Toyota (0,3); Volvo (0,7)} {Audi (0,6); Skoda (0,4)} {VW (1,0)}	wielowariantowa z wagami

Źródło: opracowanie własne.

Szerzej o zmiennych symbolicznych, obiektach symbolicznych oraz o różnicach pomiędzy danymi klasycznymi a symbolicznymi piszą m.in.: Dudek [2013, s. 42-43], Bock, Diday (red.) [2000, s. 2-8, 24-53], Billard, Diday [2006, s. 7-66], Noirhomme-Fraiture, Brito [2011], Diday, Noirhomme-Fraiture [2008, s. 3-30].

3. Klasyfikacja pojęciowa w podejściu wielomodelowym

W analizie danych symbolicznych w podejściu wielomodelowym w analizie skupień wyróżnia się dwa rozwiązania [por. Pełka 2012; de Carvalho i in. 2012; Fred, Jain 2005]:

1. łączenie wielu macierzy odległości – każda z nich postrzegana jest jako osobny punkt widzenia na zbiór danych,
2. łączenie wyników wielu klasyfikacji bazowych.

Wśród technik łączenia wyników klasyfikacji bazowych stosuje się różnorodne rozwiązania [por. Gathemi i in. 2009; Pełka 2012] – m.in. bazujące na hipergrafach, macierzy współwystąpień czy miarach informacyjnych. W metodach łączenia wielu klasyfikacji bazowych zwykle stosuje się metody iteracyjno-optymalizacyjne i hierarchiczne jako klasyfikatory bazowe dla danych klasycznych i danych symbolicznych [zob. np. Pełka 2012; Fred, Jain 2005].

Jednak klasyfikatorem bazowym w analizie danych symbolicznych mogą być także metody klasyfikacji pojęciowej. „Pojęcie jest poznawczą reprezentacją skończonej liczby wspólnych cech, które w jednakowym stopniu przysługują wszyst-

kim reprezentantom (desygnatom) danej klasy" [cyt. za: Gatnar 1998, s. 71]. Oznacza to, że w przeciwieństwie do tradycyjnych metod klasyfikacji, gdzie postuluje się, by obiekty w jednej klasie były jak najbardziej podobne, a obiekty z różnych klas jak najmniej podobne. W klasyfikacji pojęciowej obiekty w tej samej klasie mają pewne wspólne cechy.

Wynikiem klasyfikacji pojęciowej są zwykle [por. np. Gatnar 1998]:

- etykiety klas,
- pojęcia reprezentujące klasy,
- reguły przynależności obiektów do klas.

Przykładem wyniku klasyfikacji pojęciowej może być wynik działania algorytmu COBWEB, zaprezentowany przez Gatnara [1998, s. 115-118], gdzie klasyfikacji poddano pięć obiektów opisanych pięcioma zmiennymi (są to uogólnienia typów wyborców) – zob. tabela 2.

Tabela 2. Wynik klasyfikacji otrzymany za pomocą algorytmu COBWEB

Klasa	Zamieszkanie	Wykształcenie	Dochody	Partia
1	miasto	wyższe	wysokie	UW
2	miasto	wyższe	wysokie	SLD
3	miasto	średnie	przeciętne	ZChN
4	wieś	średnie	niskie, zasiłek	PSL, KPN

Źródło: opracowanie własne na podstawie [Gatnar 1998, s. 117-118].

W części empirycznej zastosowano metodę hierarchiczną/piramid P. Brito. W podejściu zaproponowanym przez E. Didaya i P. Brito [1989] każda klasa odpowiada opisującemu ją syntetycznemu obiektowi symbolicznemu (pojęciu). Jest to metoda klasyfikacji pojęciowej, która może być zastosowana do tworzenia klasyfikacji nierozłącznych (metoda piramid) lub rozłącznych (metoda hierarchiczna).

Konstrukcja dendrogramu klas w pierwszym kroku, podobnie jak w tradycyjnych hierarchicznych metodach aglomeracyjnych, zakłada, że poszczególne obiekty symboliczne tworzą klasy jednoelementowe. W następnych krokach spośród obiektów (klas) poszukuje się takich par P_i oraz P_j, aby klasa powstała w wyniku ich połączenia (P_t) była kompletna, a spośród nich wybiera się to połączenie, dla którego współczynnik uogólnienia jest najmniejszy [Dudek 2013, s. 77-78]:

$$G(P_t) = \prod_{k=1}^{m} \frac{\mu(AS_{tk})}{\mu(AS_{\max k})}, \tag{1}$$

gdzie: $AS_{\max}$ – syntetyczny obiekt symboliczny odpowiadający zbiorowi danych $\mathbf{E}$,
$\mu(\cdot)$ – długość przedziału dla zmiennych symbolicznych interwałowych, liczebność zbioru dla zmiennych symbolicznych wielowariantowych,

E – zbiór danych symbolicznych poddawany klasyfikacji,
$AS_{t\,k}$ – syntetyczny obiekt symboliczny opisujący *t*-tą klasę (powstałą z połączenia obiektów P_i oraz P_j) w dendrogramie (piramidzie) *k*,
$AS_{\max k}$ – najbardziej ogólny obiekt symboliczny opisujący dany dendrogram (piramidę) klas.

Współczynnik uogólnienia określony wzorem 1 oznacza stopień „podobieństwa" obiektów połączonych w klasę, tj. im mniejsza jest jego wartość (lub przyrost wartości), tym bardziej „podobne" obiekty zostaną połączone w jedną klasę. Oznacza to, że pojęcie opisujące taką klasę obejmować będzie tylko te obiekty, które się w niej znajdują, i nie obejmie innych obiektów spoza klasy.

Wynikiem klasyfikacji z wykorzystaniem metody hierarchicznej P. Brito jest dendrogram klas, pojęcia reprezentujące klasy oraz etykiety klas.

Zastosowanie klasyfikacji pojęciowej w podejściu wielomodelowym dla danych symbolicznych wymaga rozwiązania problematyki agregacji (łączenia) wyników klasyfikacji bazowych. W niniejszym artykule do łączenia wyników klasyfikacji pojęciowej obiektów symbolicznych proponuje się wykorzystanie macierzy współwystąpień.

Macierz współwystąpień jest wynikiem łączenia wielu wyników klasyfikacji (modeli bazowych). Wiele różnorodnych wyników klasyfikacji pojęciowej można otrzymać m.in. przez zastosowanie jednej metody klasyfikacji, ale z różnymi parametrami, wykorzystanie podzbiorów obiektów lub wykorzystanie różnych metod klasyfikacji.

Współwystępowanie pary obiektów w tych samych klasach (grupach) stanowi wskazówkę istnienia związku między nimi. Elementy macierzy współwystąpień, która ma wymiary $n \times n$, są zdefiniowane w następujący sposób [por. np. Fred, Jain 2005, s. 44]:

$$C(i,j) = \frac{n_{ij}}{N}, \tag{2}$$

gdzie: i, j – numery obiektów,
n_{ij} – wskazuje, ile razy obiekty o numerach *i*-tym oraz *j*-tym znajdują się we wszystkich N klasyfikacjach bazowych,
N – liczba klasyfikacji bazowych.

Ostateczną klasyfikację otrzymuje się przez zastosowanie macierzy współwystąpień jako macierzy danych w dowolnej metodzie klasyfikacji (np. iteracyjno-optymalizacyjnej) [por. Fred i Jain 2005]. Liczbę klas w tym przypadku ustala się, podobnie jak w klasyfikacji z wykorzystaniem jednej metody, z zastosowaniem indeksów jakości klasyfikacji. Fred oraz Jain dodatkowo dla klasyfikacji hierarchicznej proponują zastosowanie kryterium najdłuższego wiązania (*lifetime value*) [zob. Fred, Jain 2005, s. 46-47].

Algorytm klasyfikacji wielomodelowej danych symbolicznych z wykorzystaniem metody hierarchicznej P. Brito jako klasyfikatora bazowego oraz łączeniem wyników z zastosowaniem macierzy współwystąpień przedstawia się następująco:

1. Uzyskanie S różnych klasyfikacji bazowych na podstawie zbioru danych $\mathbf{E}$ (np. przez zastosowanie metody hierarchicznej P. Brito z różnymi parametrami czy wykorzystanie podzbiorów obiektów).
2. Utworzenie na podstawie S różnych klasyfikacji bazowych macierzy współwystąpień zgodnie ze wzorem 2.
3. Zastosowanie macierzy współwystąpień jako macierzy danych w metodzie *k*-średnich lub *pam*.
4. Otrzymanie ostatecznej klasyfikacji przez zastosowanie indeksów jakości klasyfikacji.

Wynikiem zastosowania tego algorytmu będzie nowa zagregowana klasyfikacja bazująca na metodzie hierarchicznej P. Brito. Klasyfikacja wynikowa w tym przypadku nie będzie już klasyfikacją pojęciową.

4. Badania symulacyjne

Na potrzeby badań symulacyjnych przygotowano w programie R trzy zbiory danych o znanej strukturze klas. Zbiór wygenerowano z wykorzystaniem pakietu `mlbench` (funkcje `mlbench.cuboids`, `mlbench.smiley` oraz `mlbench.cassini`). W celu otrzymania danych symbolicznych interwałowych z wykorzystaniem pakietu `mlbench` otrzymane dane traktowane są jako środki zmiennej symbolicznej interwałowej. Rozstęp jest dobierany w taki sposób, aby zachować oryginalny kształt danych. Najczęściej jest on dobierany losowo z przedziału $[0,1]$.

Dodatkowo w badaniach wykorzystano dwa rzeczywiste zbiory danych. Pierwszym jest zbiór danych przygotowany przez M. Ichino (oleje i tłuszcze). Zbiór ten opisują cztery zmienne symboliczne interwałowe charakteryzujące wybrane własności fizyczne i chemiczne ośmiu kwasów tłuszczowych – sezamowego, lnianego, pachnotki, bawełnianego, kameliowego, oliwy z oliwek, smalcu wieprzowego i wołowego [zob. Ichino 1988].

Drugi zbiór danych dotyczy 28 marek samochodów osobowych (obiektów symbolicznych drugiego rzędu) różnych marek opisywanych przez dziesięć zmiennych symbolicznych interwałowych [por. Pełka 2013] – cena, długość samochodu, rozstaw osi, szerokość samochodu, wysokość pojazdu, moc silnika, prędkość maksymalna, przyspieszenie, zużycie paliwa, pojemność bagażnika.

Dla każdego ze zbiorów danych wykorzystano 20 symulacji i obliczono średnią wartość skorygowanego indeksu Randa (M_R) dla klasyfikacji wielomodelowej oraz pojedynczej klasyfikacji z wykorzystaniem metody *k*-medoidów (*pam*), w której zastosowano nieznormalizowaną odległość Ichino-Yaguchiego (`U_2`). Porównanie wyników klasyfikacji dla sztucznych i rzeczywistych zbiorów danych zawarto w tab. 3.

Tabela 3. Wyniki symulacji dla rzeczywistych i sztucznych zbiorów danych

Zbiór danych	Porównywany element	PAM	Klasyfikacja wielomodelowa
Cuboids	- rozważane podziały	[2; 20]	–
	- modele bazowe	–	[2; 20] + 20 losowych z przedziału [21; 200]
	- ostateczna liczba klas	2	5
	- M_R	0,2338	0,8525
Smiley	- rozważane podziały	[2; 20]	–
	- modele bazowe	–	[2; 20] + 50 losowych z przedziału [21; 200]
	- ostateczna liczba klas	5	4
	- M_R	0,7861	1,000
Cassini	- rozważane podziały	[2; 20]	–
	- modele bazowe	–	[2; 20] + 50 losowych z przedziału [21; 200]
	- ostateczna liczba klas	2	3
	- M_R	0,5150	0,9876
Ichino	- rozważane podziały	[2; 8]	–
	- modele bazowe	–	[2; 8]
	- ostateczna liczba klas	2	2
	- M_R	1,000	1,000
Samochody	- rozważane podziały	[2; 28]	–
	- modele bazowe	–	[2; 28]
	- ostateczna liczba klas	2	3
	- M_R	0,9873	1,000

„–” nie dotyczy.

Źródło: obliczenia własne z wykorzystaniem programu R.

W przypadku sztucznych zbiorów danych o nietypowej strukturze klas (tj. cuboids, simely i cassini) podejście wielomodelowe osiąga znacznie lepsze wyniki niż w przypadku pojedynczej klasyfikacji z zastosowaniem metody *k*-medoidów (PAM). W przypadku rzeczywistych zbiorów danych (tj. samochodów i zbioru Ichino), które mają łatwą do odkrycia strukturę klas, zarówno klasyfikacja wielomodelowa, jak i metoda *k*-medoidów osiągają identyczne (albo prawie identyczne) wyniki.

Oznacza to, że w przypadku zbiorów danych o nietypowych strukturach klas oraz zbiorach danych o dużej liczbie obiektów podejście wielomodelowe z zastosowaniem klasyfikacji pojęciowej jest o wiele bardziej skutecznym narzędziem niż pojedyncze metody klasyfikacji. W przypadku zbiorów danych o niezbyt skomplikowanych strukturach klas zarówno podejście wielomodelowe, jak i pojedyncze metody klasyfikacji osiągają podobne wyniki. Należy jednakże dodać, że podejście wielomodelowe nie wymaga w tym przypadku znacznie większych nakładów obliczeniowych, niż pojedyncza metoda klasyfikacji.

5. Podsumowanie

Podejście wielomodelowe danych symbolicznych, bazujące na macierzy współwystąpień oraz klasyfikacji pojęciowej jako klasyfikatorze bazowym, może zostać z powodzeniem zastosowane w analizie danych symbolicznych różnych typów.

Klasyfikacja wielomodelowa okazała się bardziej skutecznym i użytecznym narzędziem analizy danych w przypadku sztucznych zbiorów danych przy uwzględnieniu wartości średniego skorygowanego indeksu Randa. W przypadku rzeczywistych zbiorów danych osiągnęła ona podobne wyniki jak pojedyncza metoda klasyfikacji *k*-medoidów (por. tab. 3).

W artykule zaproponowano łączenie wyników klasyfikacji pojęciowej z zastosowaniem macierzy współwystąpień. Ostateczne wyniki klasyfikacji zagregowanej nie są w tym przypadku pojęciami.

Celem przyszłych badań powinno stać się poszukiwanie metod łączenia (agregacji) wyników klasyfikacji pojęciowej w taki sposób, aby wyniki klasyfikacji zagregowanej były także pojęciami. Odrębnym zagadnieniem będzie łączenie innych elementów z bazowych klasyfikacji pojęciowych – np. reguł klasyfikacji.

Literatura

Bock H.-H., Diday E. (red.) (2000), *Analysis of Symbolic Data. Explanatory Methods for Extracting Statistical Information from Complex Data*, Springer Verlag, Berlin – Heidelberg.

Billard L., Diday E. (2006), *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester.

De Carvalho F.A.T., Lechevallier Y., de Melo F.M. (2012), *Partitioning hard clustering algorithms based on multiple dissimilarity matrices*, „Pattern Recognition" 45(1), s. 447-464.

Diday E., Brito P. (1989), *Symbolic cluster analysis*, [w:] O. Opitz (red.), *Conceptual and Numerical Analysis of Data*, Springer-Verlag, Berlin – Heidelberg, s. 45-84.

Diay E., Noirhomme-Fraiture M. (2008), *Symbolic data analysis. Conceptual statistics and data mining*, Wiley, Chichester.

Dudek A. (2013), *Metody analizy danych symbolicznych w badaniach ekonomicznych*, Wyd. UE we Wrocławiu, Wrocław.

Fred A.L.N., Jain A.K. (2005), *Combining multiple clustering using evidence accumulation*, „IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 27, s. 835-850.

Gathemi R., Sulaiman N., Ibrahim H., Mustapha N. (2009), *A survey: Clustering ensemble techniques*, „Proceedings of World Academy of Science, Engineering and Technology", vol. 38, s. 636-645.

Gatnar E. (1998), *Symboliczne metody klasyfikacji danych*, Wydawnictwo Naukowe PWN, Warszawa.

Gatnar E. (2008), *Podejście wielomodelowe w zagadnieniach dyskryminacji i regresji*, Wydawnictwo Naukowe PWN, Warszawa.

Ichino M. (1988), *General metrics for mixed features – the Cartesian space theory for pattern recognition*, [w:] *Proceedings of the 1988 IEEE International Conference on Systems, Man and Cybernetics*, vol. 1, International Academic Publishers, Beijing, s. 494-497.

Noirhomme-Fraiture M., Brito P. (2011), *Far beyond the classical data models: symbolic data analysis*, „Statistical Analysis and Data Mining", vol. 4, issue 2, s. 157-170.
Pełka M. (2012), *Ensemble approach for clustering of interval-valued symbolic data*, „Statistics in Transition", vol. 13, no. 2, s. 335-342.
Pełka M. (2013), *Podejście wielomodelowe analizy danych symbolicznych w ocenie pozycji produktów na rynku*, Ekonometria 2(40), Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wrocław, s. 95-102.

THE ENSEMBLE CONCEPTUAL CLUSTERING FOR SYMBOLIC DATA

Summary: The main aim of the paper is to present a proposal of a new ensemble clustering for symbolic data with the application of conceptual learning which is applied as the base classifier. The paper presents basic terms of symbolic data, ensemble learning and conceptual clustering. In the empirical part the results of simulation study with artificial and real data sets are presented and compared.

Keywords: symbolic data analysis, ensemble clustering, conceptual clustering.