EKONOMETRIA ECONOMETRICS 2(40) • 2013

ISSN 1507-3866

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

PODEJŚCIE WIELOMODELOWE ANALIZY DANYCH SYMBOLICZNYCH W OCENIE POZYCJI PRODUKTÓW NA RYNKU

Streszczenie: Pozycjonowanie produktów to szeroka gama działań przedsiębiorstwa, które mają na celu projektowanie oferty i wizerunku firmy. Celem tych działań jest zajęcie wyróżniającego się na tle konkurencji miejsca w świadomości rynku docelowego. Celem artykułu jest zaprezentowanie i zastosowanie podejścia wielomodelowego analizy danych symbolicznych w zagadnieniu klasyfikacji na potrzeby oceny pozycji produktów na rynku. W części empirycznej przedstawiono zastosowanie podejścia wielomodelowego danych symbolicznych bazującego na macierzy współwystąpień oraz metodzie bagging w analizie danych z rynku motoryzacyjnego. Obydwa podejścia dały porównywalne wyniki w sensie skorygowanego indeksu Randa.

Słowa kluczowe: klasyfikacja wielomodelowa, analiza skupień danych symbolicznych, pozycjonowanie produktów.

1. Wstęp

Przedsiębiorstwa i usługodawcy prowadzą szeroką gamę działań, które mają na celu projektowanie oferty i wizerunku firmy. Celem tych działań jest zajęcie wyróżniającego się na tle konkurencji miejsca w świadomości rynku docelowego. Działania te nazywa się pozycjonowaniem produktów. Wynikiem pozycjonowania jest kreowanie zorientowanej na klienta propozycji wartości, powodów, dla których dany produkt czy usługa powinna być wybrana przez konsumenta (por. np. [Kotler 2005, s. 308-309; Kotler i in. 2002, s. 139]).

Pozycjonowanie obejmuje także metody określania, jak dany produkt czy usługa oferowane przez przedsiębiorstwo plasuje się na tle produktów czy usług przedsiębiorstw konkurencyjnych. Dąży się więc do wskazania tych cech produktów czy usług, które je wyróżniają (odróżniają) na tle konkurencji. Takimi cechami mogą być np. nowoczesność, ekologia, bezpieczeństwo, prestiż itp.

Pozycjonowanie może odbywać się z wykorzystaniem wielu różnych kryteriów, które charakteryzują produkt czy usługę. Mogą one dotyczyć samego produktu, ale także producenta, użytkowników, cech czy wizerunku (por. [Stanimir (red.) 2006, s. 232]).

W pozycjonowaniu produktów czy usług zastosowanie znajduje wiele różnych metod statystycznej analizy wielowymiarowej, wśród których szczególne miejsce zajmują metody regresji logistycznej, analizy czynnikowej, analizy skupień czy skalowania wielowymiarowego (zob. np. [Stanimir (red.) 2006, s. 236-255; Walesiak 1993, s. 20-22; Zaborski 2001, s. 30]).

W badaniach marketingowych coraz częściej oprócz danych klasycznych (o czterech skalach pomiaru – nominalnej, porządkowej, przedziałowej, ilorazowej) stosuje się także dane symboliczne (zob. np. [Pełka, Jefmański 2008]). Dane symboliczne pozwalają na dokładniejszy opis zjawisk marketingowych i ekonomicznych. Niemniej jednak wymagają one zastosowania odpowiednich technik, które pozwalają na wykorzystanie całości informacji, których dostarczają zmienne symboliczne.

Podejście wielomodelowe było dotychczas z powodzeniem stosowane z zagadnieniach dyskryminacyjnych i regresyjnych (zob. np. [Gatnar 2008]). Niemniej jednak idea podejścia wielomodelowego, tj. łączenia wyników wielu modeli, może być z powodzeniem zastosowana także w zagadnieniu klasyfikacji danych symbolicznych (zob. np. [Pełka 2012]). Podejście wielomodelowe w klasyfikacji to nic innego jak łączenie (czyli agregacja) N klasyfikacji (modeli) bazowych w jedną klasyfikację złożoną o k klasach (por. [Fred, Jain 2005; Gatnar 2008]).

Podstawowym celem artykułu jest zaprezentowanie propozycji zastosowania podejścia wielomodelowego analizy danych symbolicznych bazującego na macierzy współwystąpień oraz wykorzystującego ideę metody bagging (propozycji Hornika [2005]) na potrzeby pozycjonowania produktów na przykładzie danych rzeczywistych pochodzących z rynku samochodów osobowych.

2. Dane symboliczne

W analizie danych symbolicznych obiekty mogą być opisywane przez następujące rodzaje (typy) zmiennych (zob. np. [Bock, Diday (red.) 2000, s. 2-3; Billard, Diday (red.) 2006, s. 7-30]):

1) ilorazowe, przedziałowe, porządkowe, nominalne,

2) interwałowe, czyli przedziały liczbowe – np. preferowana cena w zł = [20; 50]; czas dojazdu do pracy w minutach [15; 45],

3) wielowariantowe – np. preferowany kolor samochodu = {czarny, czerwony, zielony},

4) wielowariantowe z wagami – np. preferowana marka samochodu = {Toyota (0,8), Skoda (0,2)} – co oznacza, że 80% swojego czasu, dochodów respondent jest gotów poświęcić na kupno Toyoty, a jedynie 20% na zakup Skody,

5) interwałowe z wagami – np. czas oczekiwania na produkt w minutach {[0; 15] (0,5), [15; 20] (0,3), [20; 30] (0,2)} – co oznacza, że 50% osób czeka do 15 minut, 30% od 15 do 20 minut, a 20% od 20 do 30 minut.

Oprócz tego zmienne symboliczne mogą być także zmiennymi strukturalnymi (por. [Bock, Diday (red.) 2000, s. 2-3, 33-37; Billard, Diday (red.) 2006, s. 30-34]).

Zmienne tego typu pozwalają zdefiniować zależności funkcyjne lub logiczne (decydujące o realizacji zmiennej); warunki, od jakich zależy, czy dana zmienna opisuje dany obiekt czy nie; a także systematykę realizacji zmiennej symbolicznej.

Szerzej o zmiennych symbolicznych, obiektach symbolicznych oraz o różnicach pomiędzy danymi klasycznymi i symbolicznymi piszą m.in.: [Noirhomme-Fraiture, Brito 2011; Bock, Diday (red.) 2000, s. 2-8, 24-53; Billard, Diday (red.) 2006, s. 7-66; Diday, Noirhomme-Fraiture 2008, s. 3-30; Dudek 2013, s. 35-43].

W analizie danych symbolicznych wyróżnia się dwa podstawowe rodzaje obiektów symbolicznych (por. np. [Bock, Diday (red.) 2000, s. 5-6, 18-19, 39-53; Noirhomme-Fraiture, Brito 2011; Dudek 2013, s. 39-41]):

1) obiekty symboliczne I rzędu – są to obiekty elementarne, np. konsument, produkt, przedsiębiorstwo. Od obiektów w rozumieniu klasycznym odróżnia je fakt, że są one opisywane przez zmienne symboliczne;

2) obiekty symboliczne II rzędu – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych I rzędu lub agregacji obiektów w sensie „klasycznym" – np. grupa produktów jednego przedsiębiorstwa, konsumenci preferujący jedną markę. W przykładzie empirycznym zamiast konkretnego modelu Skody Fabii (mającego jedną cenę, konkretne wyposażenie standardowe, zużycie paliwa) mamy do czynienia z obiektem zagregowanym, który opisuje wszystkie (w danym momencie dostępne na rynku) modele Skody Fabii (każdy z nich ma inną cenę, inne opcje standardowe, zużycie paliwa itd.).

3. Podejście wielomodelowe w analizie skupień danych symbolicznych

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 (spojrzenia) na zbiór danych,

2) łączenie wyników wielu klasyfikacji bazowych.

W ramach łączenia wyników wielu klasyfikacji bazowych szczególne miejsce zajmują propozycje bazujące na macierzy współwystąpień oraz adaptujące metodę bagging (por. [Hornik 2005; Fred, Jain 2005; Pełka 2012]).

Pierwszą propozycją adaptacji metody bagging jest propozycja Leischa [1999], która łączy w sobie metody iteracyjno-optymalizacyjne i hierarchiczne. Najpierw losowane są kolejne podpróby bootstrapowe, następnie na podstawie każdej podpróby określane są rezultaty klasyfikacji z zastosowaniem bazowej iteracyjno-optymalizacyjnej metody klasyfikacji. Centra skupień ze wszystkich podziałów są przekształcane w nowy zbiór danych, który jest poddawany podziałowi z zastosowaniem metod hierarchicznych. Uzyskany dendrogram jest cięty na poziomie określonym przez badacza. Każda obserwacja z pierwotnego zbioru danych jest przydzielana do grupy, której zalążek znajduje się najbliżej.

Kolejną propozycją w zakresie adaptacji metody bagging jest propozycja Dudoit i Fridlyand [2003]. Polega ona na utworzeniu prób bootstrapowych (np. przez losowanie ze zwracaniem). Następnie dla oryginalnego (pełnego) zbioru danych oraz prób bootstrapowych stosowany jest jeden (wybrany) algorytm iteracyjno-optymalizacyjny. W dalszej kolejności dokonuje się permutacji etykiet klas w poszczególnych próbach bootstrapowych tak, aby zachodziła jak największa zbieżność z etykietami przypisanymi obiektom z oryginalnego zbioru danych.

Ostatnią z propozycji jest propozycja Hornika [2006], która zakłada utworzenie *B* prób bootstrapowych, a następnie zastosowanie klasycznego algorytmu klasyfikacji (np. pam czy *k*-średnich) dla każdej z nich. Uzyskanie ostatecznego podziału dokonywane jest przez optymalizację funkcji (zob. [Hornik 2006, s. 9]):

$$\sum_{b=1}^{B} dist(c, c_b)^2 \rightarrow \min_{c \in C}, \tag{1}$$

gdzie: C – zbiór wszystkich możliwych klasyfikacji zagregowanych,
$dist$ – miara odległości euklidesowej,
$c_b \in (c_1, \ldots, c_B)$ – elementy klasyfikacji zagregowanej.

Macierz współwystąpień jest wynikiem łączenia wielu wyników klasyfikacji (modeli bazowych). Wiele różnorodnych wyników klasyfikacji 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 i Jain 2006, s. 44]):

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

gdzie: i, j – numery obiektów,
n_{ij} – wskazuje, ile razy obiekty i, j znajdują się w tej samej klasie we wszystkich N klasyfikacjach bazowych,
N – liczba klasyfikacji bazowych.

Ostateczny podział uzyskuje się przez wykorzystanie macierzy współwystąpień jako macierzy danych w dowolnej metodzie klasyfikacji (np. hierarchiczną czy iteracyjno-optymalizacyjną) (zob. np. [Fred i Jain 2005]). Ostateczną liczbę klas w klasyfikacji wielomodelowej można otrzymać, wykorzystując znane indeksy jakości klasyfikacji. W przypadku klasyfikacji hierarchicznych można wykorzystać także kryterium najdłuższego wiązania (*lifetime value*) (zob. [Fred i Jain 2005, s. 46-47]).

4. Przykład empiryczny

W celu dokonania pozycjonowania produktów z zastosowaniem podejścia wielomodelowego w klasyfikacji obiektów symbolicznych bazującego na macierzy współwystąpień i adaptacji metody bagging zaproponowanej przez Hornika [2005] zebrano dane pochodzące z rynku samochodów osobowych.

Zbiór danych zawiera 28 marek samochodów osobowych (obiekty symboliczne II rzędu[1]) (zob. tab. 1) opisywanych przez dziesięć zmiennych symbolicznych interwałowych:

x_1 – cena katalogowa w zł,
x_2 – rozstaw osi w mm,
x_3 – długość nadwozia w mm,
x_4 – szerokość nadwozia w mm,
x_5 – wysokość nadwozia w mm,
x_6 – moc silnika w KM,
x_8 – przyspieszenie do 100 km/h w s,
x_7 – prędkość maksymalna w km/h,
x_9 – spalanie w cyklu miejskim w l,
x_{10} – pojemność bagażnika w l.

Tabela 1. Wybrane marki i modele samochodów osobowych

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

* Wersja 5-drzwiowa.

Źródło: opracowanie własne na podstawie danych z oficjalnych witryn producentów (listopad 2011).

[1] Obiekty symboliczne II rzędu powstały w wyniku agregacji obiektów klasycznych. Na przykład obiekt Skoda Fabia powstał w wyniku połączenia (agregacji) informacji o wszystkich modelach Skody Fabii, różniących się pojemnością silnika, ceną, wyposażeniem itd.

W klasyfikacji wielomodelowej z zastosowaniem macierzy współwystąpień przygotowano 11 modeli (klasyfikacji) bazowych z zastosowaniem różnych metod klasyfikacji (hierarchicznych oraz iteracyjno-optymalizacyjnych). Liczba klas była wybierana losowo z przedziału [2; 20] Na podstawie wyników klasyfikacji bazowych zbudowana została macierz współwystąpień (o wymiarach 28 × 28), którą wykorzystano jako macierz danych w metodzie kompletnego połączenia. Do wyboru ostatecznej liczby klas zastosowano indeks sylwetkowy (zob. np. [Gatnar, Walesiak (red.) 2004, s. 342-343]). Ocenę stabilności klasyfikacji przeprowadzono z zastosowaniem skorygowanego indeksu Randa.

Najwyższa wartość indeksu sylwetkowego została osiągnięta dla dwóch klas (0,6199443). Skorygowany indeks Randa dla tej klasyfikacji wyniósł 0,7099146. Świadczy to o relatywnie stabilnym podziale 28 obiektów na dwie klasy. W klasie 1 znalazły się marki samochodów z segmentów A, B oraz C. W klasie 2 znalazły się marki samochodów z segmentu D oraz dwie marki z segmentu C.

W klasyfikacji wielomodelowej danych symbolicznych z wykorzystaniem idei metody bagging (propozycja Hornika [2005]) zbudowano 20 prób bootstrapowych przez losowanie ze zwracaniem. Jako algorytm bazowy wykorzystano metodę *k*-medoidów (pam). Najwyższą wartość indeksu sylwetkowego otrzymano dla 2 klas (0,8763210). Skorygowany indeks Randa dla tej klasyfikacji wyniósł 0,6266216. W klasie 1 znalazły się marki samochodów z segmentów A, B oraz C. W klasie 2 znalazły się wyłącznie marki samochodów z segmentu D.

Samochody z segmentu A to auta miejskie (mini) o niewielkich wymiarach oraz kosztach eksploatacji. Przykładami samochodów z tego segmentu są m.in. Fiat Panda, Citroen C1, Toyota Aygo. Samochody z segmentu B to również samochody małe, które jednakże oferują więcej miejsca dla pasażerów oraz bagażowego niż samochody z segmentu A. Samochody tego segmentu są często oferowane w dwóch wersjach nadwozia – hatchback oraz sedan. Przykładami samochodów z tego segmentu są m.in. Fiat Grande Punto (wersja 5-drzwiowa), Toyota Yaris, Skoda Fabia. Samochody z segmentu C (kompaktowe, klasa niższa-średnia) to samochody średnich wymiarów, oferujące odpowiednie miejsce dla pięciu dorosłych pasażerów wraz z bagażem oraz w miarę wygodne warunki podróży. Do aut tego segmentu zalicza się m.in. Fiata Bravo, Citroena C4. Samochody segmentu D (klasa średnia, samochody rodzinne) to samochody oferujące miejsce dla pięciu dorosłych osób wraz z bagażem, pozwalające na komfortowe podróżowanie na dalekich trasach – najczęściej dostępne są w wersji nadwozia sedan. Przykładowymi samochodami z tego segmentu są m.in. Toyota Avensis czy Volkswagen Passat.

5. Podsumowanie

Wybór jednej, odpowiedniej, metody analizy skupień jest zadaniem trudnym, ponieważ nieznana jest struktura i liczba klas, którą należy odkryć. Podejście wielomodelowe zmniejsza ryzyko wyboru niewłaściwej metody. Dodatkowo uniezależnienie

się od wyboru metody powoduje, że mamy do czynienia ze zwiększeniem stabilności klasyfikacji.

Wyniki empiryczne wskazują, że podejście wielomodelowe analizy skupień danych symbolicznych może znaleźć zastosowanie w pozycjonowaniu produktów. W badanym zbiorze danych podejście to pozwoliło na odkrycie struktury dwóch klas. Porównując wyniki podejścia opartego na macierzy współwystąpień oraz wykorzystującego metodę bagging, można powiedzieć, że dają one zbliżone wyniki (pod względem stabilności klasyfikacji mierzonej skorygowanym indeksem Randa). Niemniej jednak konieczne są dalsze badania symulacyjne pozwalające dokładniej ocenić obydwa podejścia.

Dotychczasowe badania symulacyjne wskazują, że podejście wielomodelowe analizy danych symbolicznych jest mniej wrażliwe na obecność zmiennych zakłócających czy obserwacji odstających w zbiorze danych (zob. np. [Pełka 2012]).

Literatura

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

Billard L., Diday E. (red.), *Symbolic Data Analysis. Conceptual Statistics and Data Mining*, John Wiley & Sons, Chichester 2006, s. 7-30.

De Carvalho F.A.T., Lechevallier, De Melo F.M., *Partitioning hard clustering algorithms based on multiple dissimilarity matrices*, „Pattern Recognition" 2012, no. 45(1), s. 447-464.

Diday E., Noihomme-Fraiture M., *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester 2008.

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

Dudoit S., Fridlyand J., *Bagging to improve the accuracy of a clustering procedure*, „Bioinformatics" 2003, vol. 19, no. 9, s. 1090-1099.

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

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

Gatnar E., Walesiak M. (red.), *Metody statystycznej analizy wielowymiarowej w badaniach marketingowych*, Wyd. AE we Wrocławiu, Wrocław 2004.

Hornik K., *A CLUE for CLUster ensembles*, „Journal of Statistical Software" 2005, vol. 14, s. 65-72.

Kotler P., *Marketing*, Rebis, Poznań 2005.

Kotler P., Armstrong G., Saunders J., Wong V., *Marketing. Podręcznik europejski*, PWE, Warszawa 2002.

Leisch F., *Bagged clustering*, „Adaptive Information Systems and Modeling in Economics and Management Science", Working Papers 1999, SFB, 51.

Norihomme-Fraiture M., Brito P., *Far beyond the classical data models: symbolic data analysis*, "Statistical Analysis and Data Mining" 2011, vol. 4, Issue 2, s. 157-170.

Pełka M., *Ensemble approach for clustering of interval-valued symbolic data*, „Statistics in Transition" 2012, vol. 13, no. 2, s. 335-342.

Pełka M., Jefmański B., *Zmienne symboliczne w badaniach marketingowych*, „Marketing i Rynek" 2008, nr 2, s. 22-25.

Stanimir A. (red.), *Analiza danych marketingowych. Problemy, metody, przykłady*. Wyd. AE we Wrocławiu, Wrocław 2006.
Walesiak M., *Statystyczna analiza wielowymiarowa w badaniach marketingowych*, Wyd. AE we Wrocławiu, Wrocław 1993.
Zaborski A., *Skalowanie wielowymiarowe w badaniach marketingowych*, Wyd. AE we Wrocławiu, Wrocław 2001.

ENSEMBLE LEARNING FOR SYMBOLIC DATA IN PRODUCT POSITIONING

Summary: Product positioning is a wide range of business activities. Positioning is the process by which marketers try to create an image or identity in the minds of their target market for its product, brand, or organization. The main aim of the paper is to preset and apply ensemble learning for symbolic data in cluster analysis in order to evaluate a product position. Empirical part of the paper presents the application of co-occurrence matrix and bagging algorithm in ensemble learning for symbolic data (car market data was used). These two approaches reached almost the same results when considering adjusted Rand index.

Keywords: ensemble clustering, cluster analysis of symbolic data, product positioning.