PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 163
Ekonometria 30 2011

Marcin Pełka
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE DRZEW KLASYFIKACYJNYCH DLA DANYCH SYMBOLICZNYCH W OCENIE PREFERENCJI KONSUMENTÓW*

Streszczenie: W artykule przedstawiono analizę preferencji konsumentów z wykorzystaniem drzew klasyfikacyjnych dla danych symbolicznych. W tym celu zaprezentowano podstawowe pojęcia z zakresu analizy danych symbolicznych. W części empirycznej przedstawiono wyniki analizy preferencji konsumentów płatków śniadaniowych.

Słowa kluczowe: drzewa klasyfikacyjne dla danych symbolicznych, analiza preferencji, konsumenci płatków śniadaniowych.

1. Wstęp

Niezwykle ważnym zagadnieniem w teorii ekonomii, a zwłaszcza mikroekonomii i mikroekonometrii, jest pojęcie preferencji konsumenta. Kategoria preferencji jest odzwierciedleniem gustów i nawyków konsumentów ujawnionych w wyniku podjętych decyzji zakupowych lub deklarowanych w badaniach sondażowych. Preferencje wykorzystuje się do kwantyfikacji użyteczności, której w sposób bezpośredni nie można zmierzyć.

Preferencje konsumentów są jednym z istotniejszych zagadnień, które powinny być wzięte pod uwagę przez producenta. Znajomość kryteriów, pojęć, którymi operują konsumenci, pozwala zaproponować produkt zgodny z ich oczekiwaniami i przynoszący im maksymalną satysfakcję. Aby ustalić, jakie są oczekiwania konsumentów, czyli kryteria i wartości, które decydują o wyborze tego, a nie innego produktu, konieczna jest znajomość preferencji konsumentów i elementów, które na nie wpływają.

Rynek płatków zbożowych w 2009 r. osiągnął sprzedaż równą 586 mln zł, klienci zakupili ich 48 mln kg, z kolei wartość rynku w 2010 r. szacowano na 700 mln złotych (dane z marca 2010 r.). Ważną częścią rynku płatków zbożowych, o który wciąż trwa walka, jest rynek musli (mieszanka płatków zbożowych z dodatkami

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

suszonych owoców, orzechów). Według analiz, statystyczny Polak zjada tylko nieco ponad kilogram płatków zbożowych rocznie, jednakże analitycy zakładają, że rynek ten będzie się rozwijał dynamicznie (por. [Drewnowska 2010; 2009]). Duży wpływ na popularność tych produktów zbożowych ma z pewnością wzrost tempa życia polskich rodzin, a co za tym idzie – poszukiwanie dań, które są łatwiejsze w przygotowaniu. Innym ważnym czynnikiem decydującym o popularności płatków zbożowych jest rosnąca świadomość polaków na temat zdrowego odżywiania się.

Jedną z metod pozwalającą na zidentyfikowanie czynników (zmiennych) decydujących o takim, a nie innym zachowaniu konsumenta na rynku są drzewa klasyfikacyjne. Ideą konstrukcji drzewa klasyfikacyjnego jest rekurencyjny podział przestrzeni wielowymiarowej, gdzie znajdują się klasyfikowane obiekty tak, aby uzyskać rozłączne części (segmenty, klasy, grupy), w których znajdują się obiekty należące do tej samej klasy (zob. [Gatnar, Walesiak (red.), 2004, s. 103; Gatnar, Walesiak (red.), 2009, s. 238; Gatnar 2008, s. 37]).

Celem artykułu jest zaprezentowanie zastosowania drzew klasyfikacyjnych dla danych symbolicznych w analizie preferencji konsumentów. Aby to osiągnąć, artykuł przedstawia podstawowe pojęcia z zakresu analizy danych symbolicznych i drzew klasyfikacyjnych dla tego typu danych. W części empirycznej przedstawiono zastosowanie drzew klasyfikacyjnych w ocenie preferencji konsumentów płatków śniadaniowych, przy czym przykład ten ma jedynie charakter ilustracyjny.

2. Dane symboliczne

W przypadku obiektów symbolicznych możemy mieć do czynienia z rodzajami zmiennych, takimi jak [Bock, Diday 2000, s. 2-3; Billard, Diday 2006, s. 7-30]:

1) ilorazowe, przedziałowe, porządkowe, nominalne;

2) kategorie, np. biały, zielony;

3) interwałowe, czyli przedziały liczbowe, rozłączne lub nierozłączne, np. miesięczne wydatki na płatki zbożowe (12 zł; 24 zł);

4) wielowariantowe, przykładem mogą być kupowane płatki zbożowe: Kangoos, OwocoPłatki, Mlekołaki, Bio 4 zboża, Crunchy orkiszowe;

5) wielowariantowe z wagami (prawdopodobieństwami), gdzie oprócz listy kategorii występują wagi (prawdopodobieństwa), z jakimi obiekt ma wybraną kategorię, np. jeżeli wybrać zmienną preferowane marki płatków zbożowych: Mlekołaki (0,41), CiniMinis (0,29), Chocapic (0,20), płatki jęczmienne (0,10), to oznacza to, że respondent preferuje cztery rodzaje płatków zbożowych – najbardziej Mlekołaki, a najmniej płatki jęczmienne;

6) zmienne strukturalne [Bock, Diday 2000, s. 2-3; 33-37; Billard, Diday 2006, s. 30-34] – w literaturze przedmiotu wyróżnia się, oprócz wyżej wymienionych typów zmiennych, także zmienne strukturalne:

a) zmienne o zależności funkcyjnej lub logicznej pomiędzy poszczególnymi zmiennymi, gdzie *a priori* ustalono reguły funkcyjne lub logiczne decydujące o tym, jaką wartość przyjmie dana zmienna;

b) zmienne hierarchiczne, w których *a priori* ustalono warunki, od których zależy, czy zmienna dotyczy danego obiektu, czy też nie;

c) zmienne taksonomiczne, w których *a priori* ustalono systematykę, według której przyjmuje ona swoje realizacje.

W analizie danych symbolicznych wyróżnia się dwa rodzaje obiektów symbolicznych:

– obiekty symboliczne I rzędu – obiekt rozumiany w sensie „klasycznym" (obiekt elementarny), np. konsument, produkt, przedsiębiorstwo, kredytobiorca. Od obiektów w rozumieniu klasycznym odróżnia je to, że są one opisywane przez zmienne symboliczne,
– obiekty symboliczne II rzędu – obiekty utworzone w wyniku agregacji zbioru obiektów symbolicznych I rzędu lub agregacji obiektów w sensie „klasycznym", np. grupa konsumentów preferująca określoną markę produktu, kilka produktów jednego producenta, grupa kredytobiorców, która otrzymaną kwotę przeznaczyła na dany cel).

3. Drzewo klasyfikacyjne dla danych symbolicznych

Drzewa klasyfikacyjne dla danych symbolicznych można podzielić na drzewa klasyfikacyjne dla danych symbolicznych oparte na optymalnym podziale [Bock, Diday 2000, s. 245-261], warstwowe drzewa klasyfikacyjne (*strata decision trees*) (zob. [Bravo 2000; Bravo, García-Santesmases 2000; Noirhomme-Fraiture 2004, s. 273-283]) oraz bayesowskie drzewa klasyfikacyjne (*bayesian decision trees*) (zob. [Noirhomme-Fraiture 2004, s. 287-294]).

W przypadku drzew klasyfikacyjnych dla danych symbolicznych mamy do czynienia z uogólnionym algorytmem tworzenia drzewa klasyfikacyjnego (zob. [Bock, Diday 2000, s. 252-261]).

Drzewo klasyfikacyjne oparte na optymalnym podziale wymaga, by zmienną zależną identyfikującą klasy (grupy, segmenty) była zmienna nominalna. Natomiast zmienne niezależne mogą być zmiennymi symbolicznymi interwałowymi, wielowariantowymi lub wielowariantowymi z wagami lub zmiennymi klasycznymi.

Algorytm konstrukcji drzewa decyzyjnego dla danych symbolicznych można podzielić na następujące kroki [Bock, Diday 2000, s. 244-265]:

1. Budowa tablicy częstości dla zmiennych symbolicznych wielowariantowych, zmiennych nominalnych i porządkowych [Bock, Diday 2000, s. 246-249].

Dla zmiennych wielowariantowych tablica jest wynikiem zliczenia częstości zaobserwowania danej kategorii w poszczególnych obiektach.

W przypadku zmiennych wielowariantowych z wagami częstości są wagi przypisane poszczególnym kategoriom zmiennej.

Dla przedziałów liczbowych konieczne jest obliczenie wartości średniej arytmetycznej dla wszystkich możliwych kombinacji dolnych i górnych krańców przedziałów zmiennej.

Zmienne klasyczne o słabej skali pomiaru (nominalne, porządkowe) są traktowane jak listy kategorii, natomiast zmienne klasyczne o mocnej skali pomiaru (przedziałowe, ilorazowe) są traktowane jak przedziały liczbowe.

2. Ustalenie *a priori* wartości granicznej dla rozmiaru węzła n^* oraz wartości granicznej dla kryterium jakości podziału drzewa W^*. Wartości te są ustalane przez badacza na podstawie własnej wiedzy.

Jeżeli rozmiar węzła jest mniejszy od ustalonej wartości granicznej n^*, to węzeł taki jest węzłem końcowym. Jeżeli dla poszczególnego pytania binarnego wartość kryterium jakości jest większa od granicznej wartości W^*, to pytanie to może być zastosowane do podziału.

3. Konstrukcja pytań binarnych dla każdej z j zmiennych ($j = 1, ..., m$) i obliczenie prawdopodobieństw przydzielenia obiektu do lewego i prawego węzła drzewa.

Dla zmiennych interwałowych należy ustalić środki przedziałów obliczone na podstawie wszystkich możliwych dolnych i górnych krańców tej zmiennej. Środki te stanowić będą tzw. wartości odcięcia c (*cutting threshold*).

Jeżeli ustalona wartość c [Bock, Diday 2000, s. 249]:

a) znajduje się w przedziale zmiennej, wówczas prawdopodobieństwo przydzielenia obiektu do lewego węzła $p_k(l)$ wyraża się wzorem:

$$p_k(l) = \frac{c - \underline{v}_{kj}}{\underline{v}_{kj} - \overline{v}_{kj}}, \tag{1}$$

gdzie: $k = 1,\ldots,n$ – oznacza numer obiektu symbolicznego,
$\overline{v}_{kj}$ – górny kraniec przedziału j-tej zmiennej dla k-tego obiektu,
$\underline{v}_{kj}$ – dolny kraniec przedziału j-tej zmiennej dla k-tego obiektu.

b) jest mniejsza od dolnego krańca przedziału zmiennej, wówczas:

$$p_k(l) = 0\,, \tag{2}$$

c) jest większa od górnego krańca przedziału zmiennej, to:

$$p_k(l) = 1\,. \tag{3}$$

Dla zmiennych porządkowych wartość odcięcia c to poszczególne kategorie danej zmiennej (z wyłączeniem ostatniej kategorii). Dla każdego obiektu należy zsumować częstości wystąpienia tych wartości zmiennej, które są mniejsze od wartości c bądź jej równe. Otrzymana suma to prawdopodobieństwo przydzielenia k-tego obiektu do lewego węzła. Podobnie dla zmiennej nominalnej wartość odcięcia c to poszczególne kategorie tej zmiennej. Dla danego obiektu wartość c to częstość wystąpienia danej kategorii zmiennej.

Niezależnie od typu zmiennej prawdopodobieństwo przydzielenia k-tego obiektu do prawego węzła drzewa wyraża się wzorem [Bock, Diday 2000, s. 249]:

$$p_k(r) = 1 - p_k(l). \tag{4}$$

4. Obliczenie kryterium jakości podziału węzła dla każdego punktu *c* dla każdej ze zmiennych według wzoru [Bock, Diday i in. 2000, s. 254]:

$$W_j(t,c) = \log \prod_{k=1}^{n} [p_k(l) \cdot P_l(s) + p_k(r) \cdot P_r(s)], \tag{5}$$

gdzie: $j = 1,\ldots,m$ – numer zmiennej,
t – numer węzła,
c – wartość odcięcia,
$p_k(l)$ – prawdopodobieństwo przydzielenia *k*-tego obiektu do lewego węzła,
$p_k(r) = 1 - p_k(l)$ – prawdopodobieństwo przydzielenia *k*-tego obiektu do prawego węzła,
$P_l(s)$ – prawdopodobieństwo warunkowe, że w lewym węźle zaobserwowano klasę, do której należy *k*-ty obiekt (jest to iloraz sumy prawdopodobieństw przydzielenia obiektów z klasy *s* do lewego węzła oraz sumy prawdopodobieństw przydzielenia wszystkich obiektów do tego węzła),
$P_r(s)$ – prawdopodobieństwo warunkowe, że w prawym węźle zaobserwowano klasę, do której należy *k*-ty obiekt (jest to iloraz sumy prawdopodobieństw przydzielenia obiektów z klasy *s* do prawego węzła oraz sumy prawdopodobieństw przydzielenia wszystkich obiektów do tego węzła).

5. Wybór największego *W* spełniającego warunek:

$$W_j(t,c) > W^* \tag{6}$$

i podział węzła zgodnie z właściwym dla danej zmiennej sposobem, pod warunkiem że:

$$n_t > n^*, \tag{7}$$

gdzie n_t – rozmiar *t*-tego węzła.

6. Kroki 4-5 należy powtórzyć dla każdego węzła do momentu otrzymania węzłów końcowych. W dalszym podziałach nie bierze udziału pytanie wykorzystane we wcześniejszych etapach.

7. Ocena jakości reguł decyzyjnych przez obliczenie współczynnika trafności predykcji (*rate of correct predictions*). Oszacowanie współczynnika trafności predykcji wymaga obliczenia prawdopodobieństw przynależności poszczególnych obiektów do analizowanych klas.

Liczebność obiektów w węzłach końcowych decyduje o późniejszym opisie klas (grup, segmentów). Klasa o największej liczebności w węźle końcowym decyduje o tym, że węzeł ten oraz reguły decyzyjne go dotyczące stanowią charakterystykę tej, a nie innej klasy.

4. Preferencje konsumentów płatków zbożowych

W czerwcu i lipcu 2010 r. zebrano dane o preferencjach konsumentów płatków zbożowych z wykorzystaniem kwestionariusza ankiety. Ankietę rozprowadzono wśród 130 respondentów, po weryfikacji do ostatecznej analizy przyjęto 121 ankiet. Otrzymana próba ma charakter nielosowy, ponieważ dobór osób do próby miał charakter przypadkowy. Wybór przypadkowy nie gwarantuje reprezentatywności próby, ale w pewnym, choć ograniczonym zakresie umożliwia poznanie populacji (por. [Szreder 2004, s. 56]).

Respondenci w pierwszych pytaniach kwestionariusza odpowiadali na pytanie dotyczące konsumpcji płatków zbożowych. Udzielone odpowiedzi stały się podstawą do podzielenia respondentów na trzy grupy o różnych preferencjach:

a) konsumenci płatków kukurydzianych („konserwatyści") – osoby spożywające jedynie klasyczne kukurydziane płatki zbożowe, bez żadnych dodatków;

b) konsumenci płatków dla dzieci – osoby spożywające różnorodne płatki zbożowe z dodatkami (np. cukrem, czekoladą, cynamonem itp.);

c) konsumenci musli i innych płatków zbożowych – osoby spożywające musli oraz inne płatki zbożowe (np. płatki jęczmienne, otręby owsiane, płatki pszenne itp.).

Do identyfikacji czynników (zmiennych) różnicujących te grupy konsumentów wykorzystano drzewo klasyfikacyjne dla danych symbolicznych. Wśród zmiennych kandydatek znalazły się:

a) miesięczne dochody respondenta (zmienna interwałowa) – respondenci wskazywali przedział liczbowy, w którym mieszczą się ich dochody,

b) wydatki na żywność (zmienna interwałowa) – respondenci określali minimalne i maksymalne wydatki na żywność,

c) przeciętne jednorazowe wydatki na żywność (zmienna interwałowa) – respondenci określali minimalną i maksymalną kwotę pojedynczych zakupów,

d) liczba osób w gospodarstwie domowym (zmienna interwałowa) – respondenci wskazywali odpowiedni przedział liczbowy,

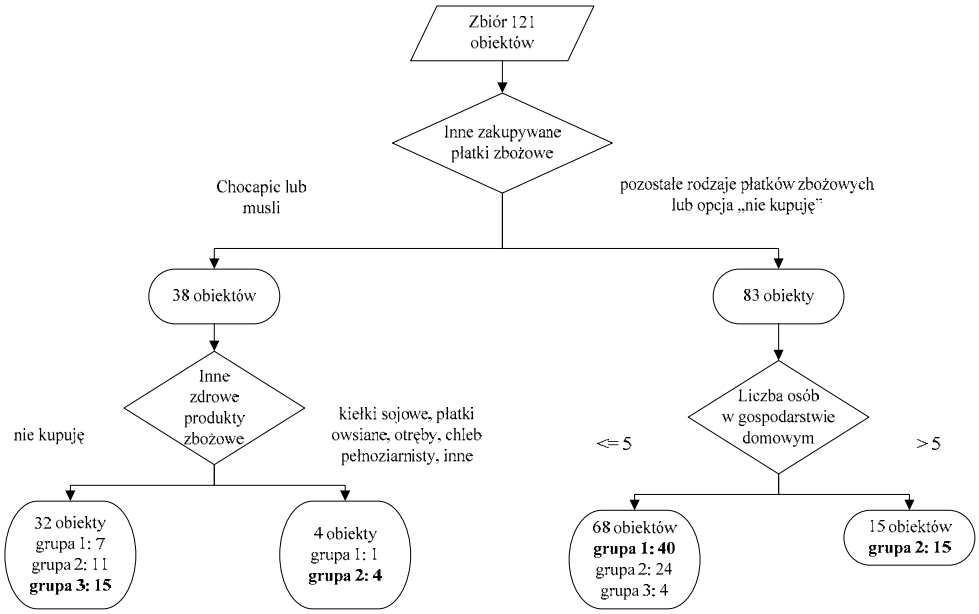
e) liczba dzieci w gospodarstwie domowym (zmienna interwałowa) – respondenci wskazywali przedział liczbowy,

f) miejsce zamieszkania (zmienna nominalna) – respondenci wskazywali jedną z dostępnych kategorii,

g) inne zakupywane płatki zbożowe (zmienna wielowariantowa) – respondenci wybierali wśród ośmiu różnych rodzajów płatków zbożowych oraz opcji „nie kupuję",

h) liczba zakupów w miesiącu (zmienna interwałowa) – respondenci wskazywali odpowiedni przedział,

i) miejsce dokonywania zakupów (zmienna wielowariantowa) – respondenci wybierali jedną z kategorii: sklep osiedlowy (wiejski), sklep sieciowy (np. „Żabka", „Biedronka", „ABC"), market,



Rys. 1. Drzewo klasyfikacyjne dla konsumentów płatków zbożowych

Źródło: obliczenia własne z wykorzystaniem programu R.

j) inne zdrowe produkty zbożowe (zmienna wielowariantowa) – respondenci wskazywali jedną spośród pięciu kategorii lub opcję „nie kupuję".

Wyniki analizy preferencji z zastosowaniem drzewa klasyfikacyjnego dla danych symbolicznych prezentuje rys. 1.

5. Wnioski

Najważniejszymi czynnikami pozwalającymi rozróżnić trzy ustalone *a priori* grupy konsumentów płatków zbożowych okazały się zmienne: inne zakupywane płatki zbożowe, inne zdrowe produkty zbożowe oraz liczba osób w gospodarstwie domowym.

Grupa 1: konsumenci płatków kukurydzianych („konserwatyści") to osoby rzadko kupujące inne produkty zbożowe, których gospodarstwo domowe liczy poniżej pięciu osób. Najczęściej są to osoby starsze, bezdzietne lub te, których dzieci opuściły gospodarstwo domowe.

Grupa 2: konsumenci płatków dla dzieci to głównie osoby kupujące oprócz płatków dla dzieci (z cukrem, czekoladowych, cynamonowych czy innych) także inne produkty zbożowe (np. musli czy klasyczne płatki kukurydziane). Konsumenci tej grupy czasem kupują inne zdrowe produktu zbożowe, takie jak chleb pełnoziarnisty czy otręby. Gospodarstwo domowe tej grupy konsumentów liczy więcej niż pięć osób.

Grupa 3: Konsumenci musli i innych płatków zbożowych to osoby, które nie kupują innych zdrowych produktów zbożowych ani innych płatków zbożowych. Jest to jednocześnie najmniej liczna grupa konsumentów.

Po oszacowaniu współczynnika trafności predykcji dla całego zbioru danych okazało się, że 38 ze 121 konsumentów zostało zaklasyfikowanych do niewłaściwej grupy (stanowi to 31,4% zbioru). Najwięcej błędnych predykcji trafiło się w przypadku konsumentów płatków dla dzieci (konsumenci grupy drugiej) – 30 przypadków. Wynika to z faktu, że w drzewie klasyfikacyjnym konsumenci tej grupy znajdują się w każdym węźle końcowym, co z kolei jest odbiciem dużej różnorodności w odpowiedziach tej grupy respondentów.

Literatura

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

Billard L., Diday E., *Symbolic Data Analysis. Conceptual Statistics and Data Mining*¸ Wiley, Chichester 2006.

Bravo M.C., *Strata Decision Tree Symbolic Data Analysis Software*, [w:] *Data Analysis Classification and Related Methods*, H.A.L Kiers., J.P. Rasson, P.J.F. Groenen, M. Schader (red.), Springer-Verlag, Berlin-Heidelberg 2000.

Bravo M.C., García-Santesmases J.M., *Symbol object description of strata by segmentation trees*, „Computational Statistics. Physica" 2000 vol. 15.

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

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

Gatnar E., Walesiak M. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

Drewnowska B., *Płatki zbożowe odporne na kryzys*, „Rzeczpospolita" nr 58 (8569) z 10 marca 2010.

Drewnowska B., *Bój o zyski z musli*, „Rzeczpospolita" nr 147 (8353) z 25 czerwca 2009.

Noirhomme-Fraiture M. (red.) 2004, *User manual for SODAS 2 software*, Software Report, Analysis System of Symbolic Official Data, Project Number IST-2000-25161.

Szreder M., *Metody i techniki sondażowych badań opinii*, PWE, Warszawa 2004.

APPLICATION OF SYMBOLIC DECISION TREES IN PREFERENCE ANALYSIS

Summary: The paper presents the application of symbolic decision trees in preference analysis. To obtain such a goal basic terms of symbolic data analysis are presented. In the empirical part the results of preferences of flake consumers analysis are presented.