PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Izabela Kurzawa, Feliks Wysocki
Akademia Rolnicza w Poznaniu

WYKORZYSTANIE ANALIZY KOSZYKOWEJ DO IDENTYFIKACJI ZACHOWAŃ KONSUMPCYJNYCH GOSPODARSTW DOMOWYCH W POLSCE

1. Wstęp

W przypadku analiz bardzo dużych zbiorów danych powstaje pytanie, w jaki sposób efektywnie i racjonalnie wykorzystać te nagromadzone informacje. W takich sytuacjach zastosowanie mogą znaleźć metody *data mining*, które umożliwiają odkrywanie nieznanych jeszcze zależności (prawidłowości) między danymi w nagromadzonych zbiorach danych [Lasek 2002]. Wśród metod *data mining* praktyczne zastosowanie znajdują metody odkrywania asocjacji – zależności między obiektami.

Celem pracy jest próba wykorzystania metod asocjacji – analizy koszykowej[1] do wykrycia zależności ukrytych w bazie danych, zawierających informacje o konsumpcji w gospodarstwach domowych w Polsce, i przedstawienie ich w postaci prostych reguł. Reguły te mogą wyrażać związki między cechami demograficzno-społecznymi gospodarstw domowych a ich wydatkami na wybrane dobra żywnościowe, nieżywnościowe i usługi. Odkryte w ten sposób prawidłowości opisują zachowania i zwyczaje zakupowe gospodarstw domowych.

Przeprowadzona analiza, oprócz aspektu poznawczego, ma zastosowanie praktyczne, stanowi bowiem cenne źródło informacji o zachowaniach konsumentów, a te mogą być wykorzystane do szybkiego budowania, aktualizowania i wdrażania skutecznych i trafnych strategii marketingowych w sektorze rolno-żywnościowym.

2. Metoda badań – analiza koszykowa

W pracy wykorzystano analizę koszykową, która polega na znajdowaniu związków między współwystępowaniem grup elementów w zbiorach danych –

[1] W pierwotnym zastosowaniu, poprzez badanie zawartości koszyków klientów, miała ona na celu usprawnienie funkcjonowania supermarketów. Analiza koszykowa wykorzystywana jest także często do analizy danych transakcyjnych pochodzących z obrotu towarowego.

odkrywaniu asocjacji [Agrawal i in. 1993 za Pasztyła 2005]. Wynikiem procesu odkrywania asocjacji jest zbiór reguł asocjacyjnych opisujących znalezione zależności lub korelacje w postaci:

jeśli wystąpi zdarzenie (cecha) X, to pociąga za sobą wystąpienie zdarzenia (cechy) Y.

Na przykład:

Jeżeli klient zakupił kawę i śmietankę, to kupi również ciastka;

Klienci banku w wieku od 49 do 60 lat najczęściej korzystają z kart płatniczych;

10% żonatych mężczyzn w wieku 40-60 lat posiada co najmniej dwa samochody w rodzinie.

Symbolicznie regułę asocjacyjną można zapisać:

JEŻELI X [poprzednik(*body*)], TO Y [następnik (*head*)]

$$X \rightarrow Y.$$

Jeżeli wiersz (pojedyncze gospodarstwo domowe) ze zbioru danych „pasuje" do reguły, czyli spełnia wszystkie warunki poprzednika i następnika, to oznacza, że reguła zawiera ten wiersz (gospodarstwo domowe), inaczej – wiersz wspiera regułę asocjacji.

Do oceny reguł asocjacyjnych stosuje się następujące miary [Tani i in. 2006]:

- Wsparcie reguły (X → Y) (*support*) oznacza stosunek liczby gospodarstw domowych do ogółu (*N*), które potwierdzają daną regułę:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}.$$

- Ufność reguły (X → Y) (*confidence*) oznacza stosunek liczby gospodarstw domowych zawierających daną regułę do liczby gospodarstw domowych zawierających cechę X:

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)},$$

gdzie: $\sigma(X \cup Y)$ – liczba gospodarstw domowych, dla których wystąpienie cechy X spowodowało wystąpienie cechy Y,

$\sigma(X)$ – liczba gospodarstw domowych z cechą X.

- Korelacja[2] (X → Y) oznacza siłę współwystępowania cech:

$$IS(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sqrt{\sigma(X) \cdot \sigma(Y)}} = \sqrt{\frac{s(X \cup Y)}{s(X)} \cdot \frac{s(X \cup Y)}{s(Y)}} = \sqrt{c(X \rightarrow Y) \cdot c(Y \rightarrow X)}.$$

Przed przystąpieniem do wyznaczania reguł asocjacyjnych należy określić minimum dla współczynnika wsparcia i ufności oraz korelacji, co będzie oznaczać poszukiwanie wszystkich takich reguł, które będą spełniały zadane warunki.

[2] Tan, Steinach i Kumar [2006] definiują ten współczynnik jako miarę uśredniającą dla niesymetrycznych danych, która jest średnią geometryczną ufności reguły X → Y oraz ufności reguły Y → X.

Reguły asocjacyjne oparte są na zmiennych jakościowych lub ilościowych (ciągłych). Dane ciągłe wymagają odpowiedniego przygotowania, tzw. dyskretyzacji. Najczęściej stosowane schematy dyskretyzacji cech ilościowych to:

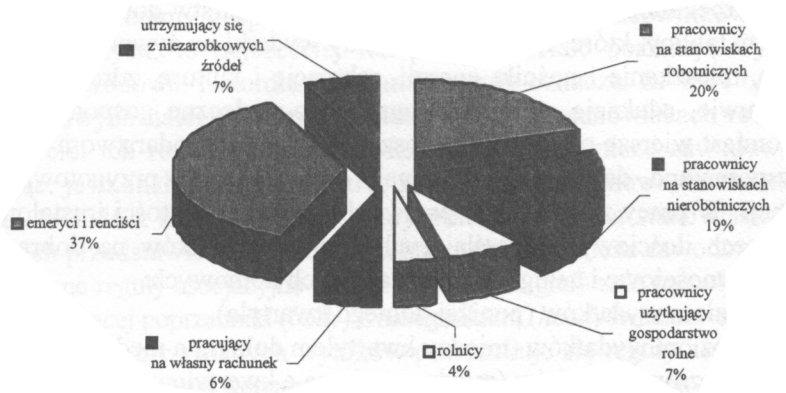
- przedziały o równej szerokości,
- przedziały o równej gęstości,
- grupowanie cech.

3. Dane wykorzystane w analizie koszykowej

Za podstawę analizy koszykowej przyjęto niepublikowane dane pochodzące z badań budżetów gospodarstw domowych prowadzonych przez Główny Urząd Statystyczny w Polsce w 2003 r. Analizowana próba roczna liczyła 32 452 gospodarstwa domowe[3].

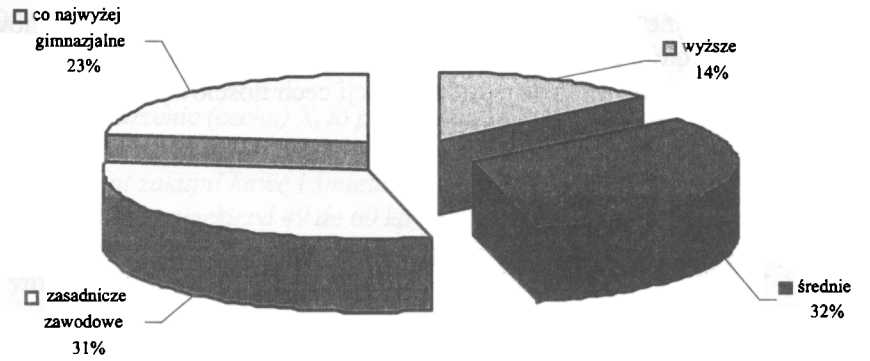
W pracy uwzględniono cechy demograficzno-społeczne gospodarstw domowych, tzn. ich klasy wyodrębnione według grup społeczno-ekonomicznych ludności (rys. 1) oraz wykształcenia głowy gospodarstwa domowego (rys. 2). Wykorzystano również dane dotyczące wydatków ponoszonych przez gospodarstwa domowe na żywność i napoje bezalkoholowe oraz mieszkanie i nośniki energii, rekreację i kulturę, zdrowie, transport, odzież i obuwie, edukację.

Na rys. 1 przedstawiono strukturę badanych gospodarstw domowych według grup społeczno-ekonomicznych. Największa część badanych gospodarstw domowych należała do grupy gospodarstw domowych emerytów i rencistów (37%). Natomiast najmniejszy odsetek gospodarstw domowych stanowili rolnicy (4%).



Rys. 1. Struktura gospodarstw domowych według grup społeczno-ekonomicznych w Polsce w 2003 r.
Źródło: opracowanie własne na podstawie niepublikowanych danych GUS z badań budżetów gospodarstw domowych w 2003 r.

[3] Badanie budżetów gospodarstw domowych prowadzone jest metodą reprezentacyjną, która daje możliwość uogólnienia uzyskanych wyników na wszystkie gospodarstwa domowe w kraju. Stosowany jest schemat losowania warstwowego, dwustopniowego. Jednostkami losowania pierwszego stopnia są terenowe punkty badań, które powarstwowano według województw, drugiego stopnia zaś – mieszkania w rejonach miejskich oraz wiejskich.



Rys. 2. Struktura gospodarstw domowych według grup społeczno-ekonomicznych w Polsce w 2003 r.
Źródło: opracowanie własne na podstawie niepublikowanych danych GUS z badań budżetów gospodarstw domowych w 2003 r.

Na rys. 2 przedstawiono strukturę badanych gospodarstw domowych według wykształcenia ich przedstawiciela. Największa część badanych gospodarstw domowych reprezentowana była przez osoby z wykształceniem średnim (32%) oraz zasadniczym zawodowym (31%). Najmniejszy odsetek gospodarstw domowych prowadzony był przez osoby z wykształceniem wyższym (14%).

4. Analiza koszykowa – wybrane wyniki badań[4]

Dane mikroekonomiczne z badań budżetów gospodarstw domowych z 2003 r. zestawiono w tablicy, której kolumny zawierały wydatki na żywność i napoje bezalkoholowe, mieszkanie i nośniki energii, rekreację i kulturę, zdrowie, transport, odzież i obuwie, edukację, cechy demograficzno-społeczne gospodarstw domowych, natomiast wiersze odpowiadały poszczególnym gospodarstwom domowym. Jak już wspomniano, dane ciągłe wymagały odpowiedniego przygotowania, tzw. dyskretyzacji. W pracy zastosowano przedziały o równej gęstości i ustalono 4 poziomy dla cech ilościowych określających poziom wydatków na dobra żywnościowe, nieżywnościowe i usługi w gospodarstwach domowych:

1) niski poziom wydatków (poniżej dolnego kwartyla),
2) średni poziom wydatków (między kwartylem dolnym a medianą),
3) wysoki poziom wydatków (między medianą a kwartylem górnym),
4) bardzo wysoki poziom wydatków (powyżej kwartyla górnego).

Następnie określono minimum dla współczynnika wsparcia (5%) i ufności (5%) oraz korelacji (20%), co oznacza poszukiwanie wszystkich takich reguł, które będą spełniały zadane warunki.

Otrzymano reguły asocjacyjne, których fragment zestawiono w tablicy wynikowej (rys. 3).

[4] W pracy wykonano obliczenia przy użyciu programu Statistica 7 – *Analiza koszykowa*.

Summary of association rules
Min. support = 5,0%, Min. confidence = 5,0%, Min. correlation = 20,0%
Max. size of body = 8, Max. size of head = 8

Body	==>	Head	Support(%)	Confidence(%)	Correlation(%)
gr społ_ekonom == pr_rob	==>	wykszt == zawodowe	13	61	49
gr społ_ekonom == pr_rob	==>	niskie_żyw_nap	8	40	36
gr społ_ekonom == pr_rob	==>	średnie_żyw_nap	6	29	26
gr społ_ekonom == pr_rob	==>	średnie_miesz_ener	7	35	32
gr społ_ekonom == pr_rob	==>	niskie_re_kul	6	30	28
gr społ_ekonom == pr_rob	==>	średnie_wyd_og	7	32	29
gr społ_ekonom == pr_rob	==>	średnie_re_kul	5	27	25
gr społ_ekonom == pr_rob	==>	niskie_zdrow	6	30	30
gr społ_ekonom == pr_rob	==>	niskie_miesz_ener	6	28	25

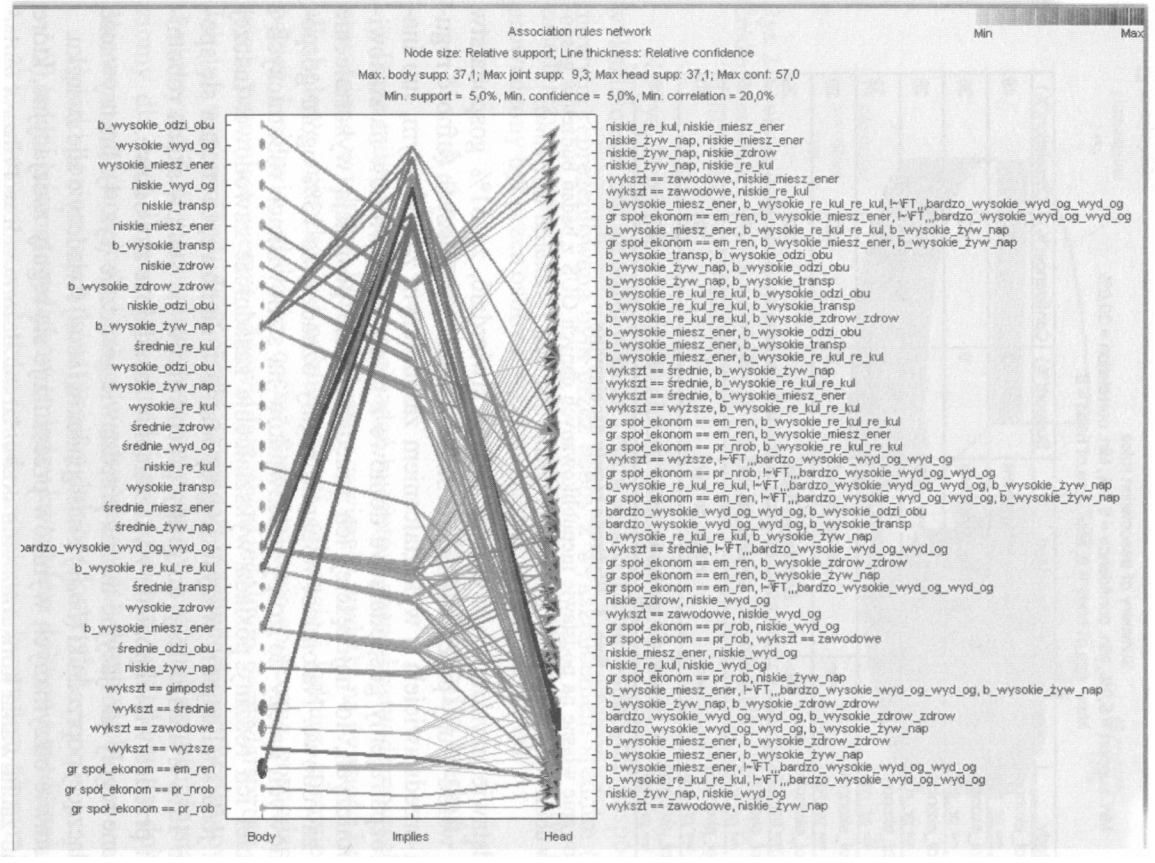
Rys. 3. Wybrane reguły asocjacyjne
Źródło: opracowanie własne na podstawie niepublikowanych danych GUS z badań budżetów gospodarstw domowych w 2003 r. i obliczeń z programu Statistica 7 – *Analiza koszykowa.*

Na podstawie uzyskanych wyników można stwierdzić, że 61% gospodarstw domowych wśród grupy pracowników na stanowiskach robotniczych (ufność reguły) posiada przedstawiciela z wykształceniem zasadniczym zawodowym, lub inaczej: jeśli rozpatrujemy gospodarstwa domowe z grupy pracowników na stanowiskach robotniczych, to najczęściej ich przedstawiciele posiadają wykształcenie zasadnicze zawodowe. Natomiast wsparcie reguły oznacza, że 13% ogółu gospodarstw domowych należy do grupy pracowników na stanowiskach robotniczych i przedstawiciel ich rodziny posiada wykształcenie zasadnicze zawodowe. Inaczej formułując: jeśli analizujemy wszystkie gospodarstwa domowe, to można się spodziewać, że 13% z nich należy do grupy pracowników na stanowiskach robotniczych, a ich przedstawiciel posiada wykształcenie zasadnicze zawodowe.

Otrzymane reguły asocjacyjne można przedstawić graficznie (rys. 4). Intensywność koloru linii łączącej poprzedniki (*body*) z następnikami (*head*) świadczy o sile związku.

Na podstawie otrzymanych wyników przeszukuje się reguły asocjacyjne, które mogą opisać zachowania konsumpcyjne gospodarstw domowych w Polsce i zadać np. pytanie: *Co charakteryzuje gospodarstwa domowe wybranej grupy społeczno-ekonomicznej ze względu na główne grupy wydatków na dobra żywnościowe, nieżywnościowe i usługi?* Wówczas na podstawie otrzymanych reguł asocjacyjnych można stwierdzić, że jeżeli gospodarstwo domowe należy np. do grupy emerytów i rencistów, to najczęściej pociąga za sobą:

- bardzo wysokie wydatki na mieszkanie i energię,
- wysokie lub bardzo wysokie wydatki na zdrowie,



Rys. 4. Przykład graficznej prezentacji reguł asocjacyjnych

Źródło: opracowanie własne na podstawie niepublikowanych danych GUS z badań budżetów gospodarstw domowych w 2003 r. i obliczeń z programu Statistica 7 – *Analiza koszykowa*.

- wysokie lub bardzo wysokie wydatki na żywność i napoje bezalkoholowe,
- niskie wydatki na rekreację i kulturę,
- niskie wydatki na transport,
- niskie wydatki na odzież i obuwie.

Dość charakterystyczną grupę stanowią gospodarstwa domowe pracowników na stanowiskach nierobotniczych. Dla tej grupy rodzin otrzymano następujące reguły:

jeżeli gospodarstwo domowe należy do wspomnianej grupy, to najczęściej pociąga za sobą:

- bardzo wysokie wydatki ogółem,
- co najmniej średnie wykształcenie głowy gospodarstwa domowego,
- wysokie lub bardzo wysokie wydatki na rekreację i kulturę,
- bardzo wysokie wydatki na transport,
- bardzo wysokie wydatki na odzież i obuwie.

Z kolei stawiając pytanie: *Co charakteryzuje gospodarstwa domowe osób z wybranym wykształceniem, jeśli weźmie się pod uwagę główne grupy wydatków na dobra żywnościowe, nieżywnościowe i usługi?*, można uzyskać następującą odpowiedź:

jeżeli głowa gospodarstwa domowego posiada np. wykształcenie zasadnicze zawodowe, to najczęściej pociąga za sobą:

- przynależność gospodarstwa domowego do grupy pracowników na stanowiskach robotniczych,
- średnie lub niskie wydatki na żywność i napoje bezalkoholowe.

5. Podsumowanie

Na podstawie przeprowadzonych badań można wskazać zalety i wady zastosowanej analizy koszykowej.

Do zalet można zaliczyć:

- fakt, że otrzymane reguły asocjacyjne mogą opisywać zachowania konsumpcyjne grup gospodarstw domowych,
- analiza koszykowa może stanowić punkt wyjścia do dalszych szczegółowych analiz,
- niewątpliwą korzyścią jest również prostota zastosowanej metody analizy koszykowej.

Natomiast do wad należy zaliczyć:

- fakt, że często przy przekształcaniu danych ilościowych traci się część informacji o badanej zbiorowości, otrzymując reguły asocjacyjne na wyższym poziomie abstrakcji,

- z drugiej zaś strony, szukanie asocjacji pośród nieprzekształconych danych ilościowych wpływa na niską efektywność tego algorytmu i przeszukiwanie zbyt dużej liczby zbiorów kandydujących do reguł,
- duża liczba cech przy wyszukiwaniu reguł wpływa również na złożoność obliczeniową, co pociąga za sobą zmniejszenie efektywności.

Literatura

Agrawal R., Imieliński T., Swami A. (1993), *Mining Association Rules between Sets of Items in Large Databases*, Proceedings of ACM SIGMOD International Conference on Management of Data, Washington D.C.

Lasek M. (2002), *Data Mining. Zastosowania w analizach i ocenach klientów bankowych*, Biblioteka Menedżera i Bankowca, Warszawa.

Pasztyła A. (2005), *Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej*, [w:] *Data Mining: Poznaj siebie i swoich klientów*, StatSoft Polska Sp. z o.o., Kraków.

Tan P., Steinach M., Kumar V. (2006), *Introduction to Data Mining*, Pearson Education, Inc., Boston.

THE APPLICATION OF BASKET ANALYSIS IN IDENTIFICATION OF CONSUMPTION BEHAVIOURS IN POLISH HOUSEHOLDS

Summary

The purpose of this analysis is to test the application of a certain type of data mining technique – association rule mining. This method is also called basket analysis. It is designed for finding interesting relationships in large set of data. In this research project, the association rules are based on micro-economic data concerning individual Polish household budgets in 2003. It is used to explore connections between household expenditures (food and no-food goods, services) and socio-demographic features of households. This can present consumption behaviours of household.