PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176
Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Małgorzata Śniegocka-Łusiewicz
Uniwersytet Mikołaja Kopernika w Toruniu

ANALIZA KOSZYKOWA W BADANIU CYKLICZNOŚCI REGUŁ ASOCJACJI W HANDLU DETALICZNYM

Streszczenie: W pracy zakłada się istnienie zmienności w czasie preferencji konsumentów odnośnie do wyboru produktów. Wiedza na temat reguł wyboru następników oraz prawdopodobieństwa wyboru poprzedników w poszczególne dni oraz miesiące ułatwi przygotowanie reklamy skierowanej do konkretnych reklamobiorców. Analiza koszykowa umożliwiała przebadanie empirycznego zbioru danych i uwidoczniła zmiany w poziomie ufności i wsparcia w poszczególnych okresach. Z badania wynika istnienie najsilniejszych reguł w październiku oraz w soboty przy równoczesnym istnieniu największej liczby transakcji. Ponadto w artykule wykazano, że w badaniach empirycznych kryteria minimalne analizy koszykowej muszą być ustalone na niskim poziomie.

Słowa kluczowe: analiza koszykowa, algorytm *a priori*, zachowania konsumenckie, *data mining*.

1. Wstęp

Dla reklamodawców najistotniejsze jest trafienie z przekazem do odpowiednich osób, w odpowiednim czasie i z właściwym produktem. Istnieje wiele sposobów na zbliżenie się do tego ideału. Niniejsza praca ma na celu zaprezentowanie jednego z nich. Przedstawiona zostanie metoda zdobycia wiedzy na temat preferencji konsumentów wybranego sklepu. Dzięki zależnościom, jakie można odnaleźć w historii sprzedaży pojedynczych transakcji, istnieje możliwość odpowiedniego przygotowania akcji marketingowych w zależności od miesiąca oraz dnia tygodnia.

Głównym celem badania jest zaobserwowanie występowania zjawiska zmienności w cyklu rocznym i tygodniowym w zachowaniach konsumentów – klientów wybranego sklepu. Podstawowym narzędziem badawczym będzie analiza koszykowa pozwalająca, za pomocą wyznaczonych reguł asocjacji, ocenić zmienność zachowań konsumenckich pod względem preferencji zakupowych.

Prezentacja ma na celu przedstawienie wyników analizy koszykowej wybranej bazy danych. Zaprezentowane zostaną cechy charakterystyczne omawianych danych i sposób ich przygotowania do analizy oraz wyniki badań zmienności miar asocjacji dla dużej bazy transakcyjnej sklepu mięsnego. Ponadto przedstawione zostaną możliwe sposoby wykorzystania analizy asocjacji do celów marketingowych.

Zgodnie z definicją [Hansen 1972, s. 15] zachowania konsumenckie jest to „ogół działań i percepcji konsumenta składających się na przygotowanie decyzji wyboru produktu, dokonanie owego wyboru oraz konsumowanie". Temat ten jest niezwykle szeroki, istnieje też wiele metod badania poczynań konsumenckich [Rudnicki 2000, s. 240-317]. Jednym ze sposobów jest analiza danych transakcyjnych, którą można sprowadzić do analizy paragonów. Dysponując bazą danych zawierających informacje z paragonów kolejnych klientów sklepu mięsnego, można pokusić się o próby wykrycia zależności i wpływów zakupów jednego produktu na zakup innego. Dzięki temu można zapoznać się, choć częściowo, z decyzjami, jakie konsument podejmuje, oraz − przy odpowiednim podejściu − spróbować na te decyzje wpłynąć.

Badana baza danych zawiera informacje z ponad dwóch lat sprzedaży w sklepie mięsnym − razem 554 546 rekordów (wierszy odpowiadających zakupowi jednego produktu, kilka wierszy składa się na jeden paragon). Wśród znajdujących się tam informacji są następujące: nazwa produktu, zakupiona ilości, cena oraz szczegóły sprzedaży (nr paragonu, data i godzina sprzedaży, sprzedawca obsługujący, podatek naliczony do produktu, kod produktu i wartość produktu). Dzięki takim danym można przeprowadzić badanie w grupach paragonów odnośnie do kształtowania się decyzji konsumenckich. Do tejże analizy można zastosować jedną z metod data miting − analizę koszykową.

Analizą koszykową nazywamy analizę zawartości koszyka klienta. Każdy klient typowego sklepu dysponuje koszykiem, w którym gromadzi kupowane produkty. Kluczową sprawą jest to, że on sam decyduje, co, w jakiej ilości oraz w jakiej kolejności znajdzie się w koszyku. Analiza koszykowa polega na rozpoznaniu reguł, którymi kierują się klienci przy zapełnianiu koszyka, zwyczajów danego klienta, prawidłowości w korzystaniu z danego typu usług, badaniu, jakie produkty kupowane są razem lub w określonej sekwencji. W analizie koszykowej termin „koszyk" jest bardzo umowny. Równie dobrze przedmiotem badania może nie być sklep, lecz punkt usługowy albo sklep internetowy. Za pomocą analizy koszykowej możemy poznać, jakie miary asocjacji występują w danym badaniu oraz jakie wartości przybierają. Do badania posiadanej bazy danych zastosowany zostanie algorytm *a priori* analizy koszykowej.

Algorytm *a priori* [Larose 2006, s. 189] wydobywa zestaw reguł z danych, wybierając reguły z najwyższą zawartością informacji. Metoda *a priori* oferuje 5 różnych metod selekcji reguł i wykorzystuje wyrafinowany model indeksowania do wydajnego przetwarzania dużych baz danych. Metoda *a priori* wymaga, aby

wszystkie dane wejściowe i wyjściowe były jakościowe. Algorytm ten wykorzystuje właściwość *a priori*, która mówi, że jeżeli zbiór zdarzeń *Z* nie jest pusty, to dla dowolnego elementu *A*, gdzie $Z \cup A$, także nie będzie pusty. Oznacza to, że dodanie dowolnego artykułu do zbioru niepustego nie spowoduje, iż zbiór ten stanie się pusty. Kolejnym wnioskiem jest ten, iż żaden nadzbiór niepusty zbioru nie będzie pusty. Oznacza to, że poszukując zbiorów częstych, algorytm najpierw przeanalizuje wszystkie jednoelementowe podzbiory i dopiero wśród tych częstych będzie szukał kandydatów na częste zbiory dwuelementowe i tak dalej. Po odnalezieniu wszystkich zbiorów częstych (k) algorytm wyszuka wszystkie podzbiory (l) znalezionych zbiorów częstych. Następnie zbada występowanie reguły, jeżeli l to $(k-l)$. Dla podanych reguł algorytm wylicza poziom wsparcia i ufności. Od badacza zależy, jaki poziom wsparcia i ufności uzna za minimalny dla danego badania [Rauch 2005]. W przeprowadzonym badaniu za pomocą algorytmu *a priori* przedmiotem analizy będą następujące miary asocjacji:

– następnik − w przypadku zależności, jeżeli *A* to *B*, jest to szukane *B*,
– poprzednik − w przypadku zależności, jeżeli *A* to *B*, jest to szukane *A*,
– wsparcie (%) − jest to prawdopodobieństwo wystąpienia zdarzenia *A*,
– poziom ufności (%) − jest to stosunek rekordów z *A* i *B* do wszystkich rekordów z *A*.

2. Baza danych transakcyjnych

Przed przystąpieniem do analizy koszykowej bazy danych za pomocą programu SPSS Clementine niezbędna była częściowa obróbka posiadanych danych. Przede wszystkim do badania nie była brana pod uwagę ilość zakupionego towaru. Niezależnie, czy było to 10 czy 25 dag szynki, w badaniu jest to zakup jednej sztuki produktu. Ponadto pominięte zostały wszystkie szczegóły dotyczące sprzedaży, np. numer kasy, z której pochodził paragon – całość sprzedaży badana jest łącznie. Do tego celu wykorzystywane są informacje z nazwą (kodem) produktu oraz unikalnym numerem paragonu. W badanym sklepie występowała jednakże zmienność nazw tych samych produktów. Spowodowane było to na przykład akcjami promocyjnymi, wyprzedażami lub niestandardowymi opakowaniami – wtedy produkt zapisywany był pod innym kodem. Podobnie produkty charakteryzujące się drobnymi różnicami skatalogowane zostały jako oddzielne produkty. To wszystko wymusiło pogrupowanie ich w większe grupy produktowe. Dzięki temu nie zaistniały zniekształcenia spowodowane zmianą kodu w przypadku sprzedaży promocyjnej lub w innym opakowaniu, a ponadto można zaobserwować generalny trend dotyczący grupy produktów, a nie pojedynczych sztuk, przy których zależności byłyby bardzo trudne do wychwycenia, jednakże takie badanie może być w przyszłości rozważone. Oczywiste jest, że pominięcie części danych i pewna generalizacja materiału po obróbce powoduje utratę części informacji i uniemożliwia pełniejszą

analizę problemu. Jednakże zbyt duża drobiazgowość może doprowadzić do niezauważenia ogólnych zmian, trendów, co z kolei może nie do końca pozwala wytłumaczyć przeszłe zachowania konsumentów, ale umożliwia prognozowanie, a taka wiedza jest przydatna osobom zajmującym się marketingiem danego ośrodka.

W celu przeprowadzenia analizy zmienności w czasie baza została pogrupowana ze względu na dzień tygodnia oraz na miesiąc. Następnie przeanalizowana została każda grupa oddzielnie. Zastosowano kryterium minimalnego poziomu wsparcia i ufności wynoszącego 10%. Zarówno w przypadku miesięcy, jak i dni tygodnia dominującym następnikiem była grupa „wędliny". Kolejnym następnikiem z dużą częstością była „kiełbasa". Odnośnie do poprzedników przekrój był znacznie większy, jednakże analizując te, które najczęściej występowały w grupie z najwyższym poziomem ufności, udało się wyselekcjonować jako przykładowe poprzedniki następujące grupy produktowe: „ser", „parówki" oraz „boczek".

3. Roczna zmienność reguł asocjacji

W tabeli 1 przedstawione zostały wyniki wsparcia oraz poziomu ufności w kolejnych miesiącach dla podanych poprzedników przy następniku „wędliny".

Tabela 1. Wyniki miesięczne dla wybranych poprzedników przy następniku „wędliny"

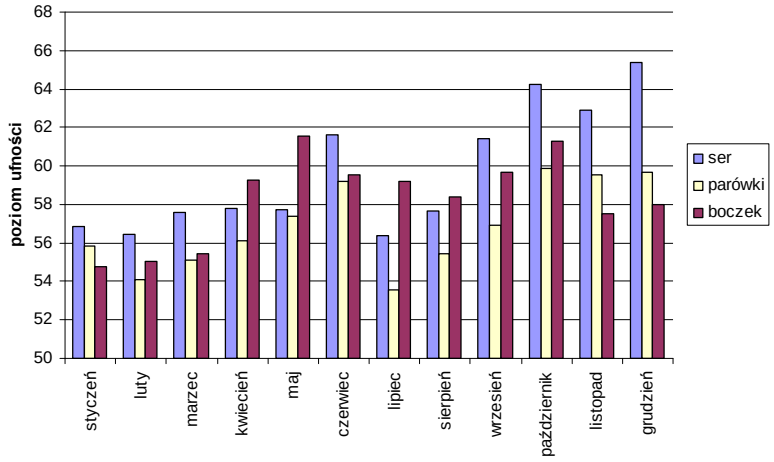
Miesiąc	Wsparcie (%)			Poziom ufności (%)		
	ser	boczek	parówki	ser	boczek	parówki
Styczeń	16,41	9,43	20,55	56,86	54,79	55,88
Luty	13,79	12,00	24,26	56,45	55,02	54,07
Marzec	14,30	11,01	22,65	57,62	55,43	55,10
Kwiecień	14,83	11,40	23,59	57,77	59,26	56,14
Maj	18,92	12,19	24,26	57,73	61,57	57,40
Czerwiec	18,54	12,21	25,17	61,60	59,52	59,23
Lipiec	19,85	10,38	24,74	56,36	59,19	53,58
Sierpień	17,26	11,86	24,80	57,64	58,42	55,42
Wrzesień	15,74	11,80	24,15	61,39	59,66	56,90
Październik	16,00	12,49	23,34	64,21	61,30	59,89
Listopad	17,31	10,57	22,01	62,91	57,55	59,53
Grudzień	15,20	11,01	21,98	65,38	57,99	59,67

Źródło: obliczenia własne.

Dla każdego miesiąca przy wspomnianych kryteriach znalezionych zostało od ok. 130 do 200 reguł (najwięcej od lipca do października). Liczebność zbiorów bazowych wahała się około 46 000 rekordów, przy czym najmniejszą liczbę transakcji zarejestrowano w miesiącach styczeń-luty, a największą (54 878) w październiku. W przypadku stycznia niezbędne było obniżenie kryteriów minimalnego poziomu wsparcia do 5%, ponieważ w przeciwnym razie program nie znajdował re-

guły dla grupy produktowej „boczek". Obniżenie kryteriów minimalnych wpłynęło na zwiększenie liczby odnalezionych reguł nawet do ok. 700.

Do lepszego zobrazowania zmienności niezbędna jest prezentacja graficzna. Na rysunku 1 można zaobserwować zmienność poziomu ufności, czyli w przypadku niebieskich słupków jest to procent paragonów, na których wystąpiła pozycja „wędliny" w grupie paragonów z pozycją „ser".



Rys. 1. Poziom ufności w miesiącach dla wybranych poprzedników przy następniku „wędliny"

Źródło: opracowanie własne.

Z powyższych wyliczeń wynika, że poziom ufności w kolejnych miesiącach jest zmienny. Charakterystyczne jest to, że największe wartości poziomu ufności znajdują się w październiku. Oznacza to, że jeżeli wystąpi którykolwiek z podanych poprzedników, to wystąpienie produktów z grupy „wędliny" można prognozować z pewnością przekraczającą 60%. Co prawda poziom wsparcia w tym okresie nie jest wysoki, ale równocześnie jest to miesiąc z największą liczbą pozycji na paragonach i liczbą paragonów. Ogólnie od początku roku do czerwca wartości poziomu ufności wzrastają. Następnie ma miejsce spadek w okresie letnim i znaczny wzrost we wrześniu i w październiku. Co ciekawe, duże wartości obserwujemy również w listopadzie i grudniu, jednakże dla grupy „ser", „parówki" i „boczek" pozostają bez większych zmian. Oznacza to, że w miesiącach, w których poziom ufności jest znaczny, można położyć większy nacisk na reklamę grup produktów poprzedników („ser", „parówki", „boczek"), ponieważ ich zwiększona sprzedaż powinna pociągnąć za sobą zwiększenie sprzedaży z grupy „wędliny". Natomiast w grudniu nacisk ten powinien zostać położony jedynie na grupę „sery", ponieważ ona, pomimo nie aż tak dużego poziomu wsparcia, generuje znaczny poziom ufności dla następnika „wędliny".

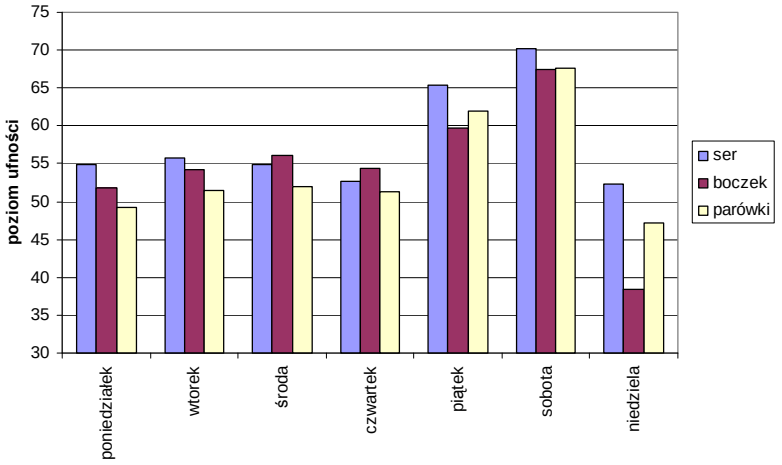
4. Tygodniowa zmienność reguł asocjacji

Nieco inna sytuacja jest w przypadku kolejnych dni tygodnia, ale wyniki również świadczą o niewielkiej zmienności. W tabeli 2 znajdują się wartości dla kolejnych dni tygodnia.

Tabela 2. Wyniki w poszczególnych dniach tygodnia dla wybranych poprzedników przy następniku „wędliny"

Miesiąc	Wsparcie (%)			Poziom ufności (%)		
	ser	boczek	parówki	ser	boczek	parówki
Poniedziałek	16,08	10,92	20,89	54,84	51,82	49,19
Wtorek	15,79	9,96	21,79	55,78	54,27	51,49
Środa	15,96	10,82	21,11	54,91	56,06	51,99
Czwartek	16,19	10,39	21,61	52,65	54,35	51,27
Piątek	16,80	12,23	24,78	65,31	59,65	61,93
Sobota	17,47	13,32	28,87	70,14	67,48	67,68
Niedziela	10,10	6,25	17,31	52,38	38,46	47,22

Źródło: obliczenia własne.



Rys. 2. Poziom ufności w dniach tygodnia dla wybranych poprzedników przy następniku „wędliny"

Źródło: opracowanie własne.

W przypadku dni tygodnia istotną sprawą jest to, że liczba pozycji paragonowych od poniedziałku do czwartku waha się od 70 000 do 84 000, a w piątek i sobotę osiąga średnio 117 000. Większa liczba transakcji ujawnia się również w wyższych wartościach wsparcia w tych dniach. Natomiast w niedzielę sprzedaż odbywała się jedynie okazjonalnie i z zasady sklep był zamknięty, stąd też tylko 670

pozycji, dlatego porównywanie wyników dla tego dnia z innymi byłoby z góry skazane na zniekształcenia.

Na rysunku 2 można zauważyć, że od poniedziałku do czwartku zmienność pomiędzy poszczególnymi dniami jest niewielka i z dość zbliżonym prawdopodobieństwem występuje współkupowanie grup towarowych poprzedników z następnikiem. Zdecydowany skok następuje w piątek i sobotę, kiedy zakup któregokolwiek z poprzedników daje ponadsześćdziesięcioprocentowe prawdopodobieństwo zakupu następnika. Co interesujące, wbrew obiegowej opinii, że w poniedziałki nie powinno się kupować produktów mięsnych ze względu na ich potencjalną nieświeżość, zarówno dane dotyczące poziomu ufności, jak i wsparcia nie sugerują znacznej różnicy pomiędzy tym dniem a następnymi. Co prawda liczba transakcji w poniedziałki jest najniższa spośród całego tygodnia, ale jedynie o 15%-17% od liczby transakcji w trzy kolejne dni.

4. Podsumowanie

Analiza koszykowa umożliwia uzyskanie reguł asocjacji w pogrupowanym zbiorze empirycznym, jednakże kryteria minimalne muszą być przyjęte na dość niskim pułapie. Obniżenie kryteriów minimalnych pozwala odnaleźć nawet trzy razy więcej reguł, ale ich przydatność może być wątpliwa ze względu na znikomy udział. Badanie poszczególnych miesięcy oraz dni tygodnia pozwoliło zaobserwować zmienność zachowań konsumenckich, a szczególnie ich preferencji wyboru produktów. Największym poziomem ufności charakteryzują się reguły zaistniałe w październiku oraz w soboty. Oznacza to, że kampania reklamowa powinna być uzależniona od poszczególnych dni tygodnia oraz od miesięcy – nacisk powinien być zmienny oraz położony na inne grupy produktowe. Ostatnim wnioskiem jest ten, że mimo mniejszej nieznacznie ilości sprzedaży w poniedziałki struktura zależności jest zbliżona.

Literatura

Hansen F., *Consumer Choice Behavior. A Cognitive Theory*, The Free Press, New York 1972.

Larose D.T., *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, Wyd. Naukowe PWN, Warszawa 2006.

Rauch J., *Logic of association rules*, „Applied Intelligence" 2005, vol. 22.

Rudnicki L., *Zachowanie konsumentów na rynku*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2000.

THE USE OF BASKET ANALYSIS IN THE RESEARCH OF CYCLICAL NATURE OF ASSOCIATION RULES IN RETAIL TRADE

Summary: This paper assumes that consumer preferences about the product choice vary over time. Knowledge on the rules of choosing consequents and probability of choosing predecessors on individual days and months may enable better preparation and targeting of an advertisement to advertising addressees. Basket analysis enables empirical analysis of data and it shows the changes in confidence and support level over time. The study shows the existence of the most significant association rules in October and on Saturdays, combined with the highest number of transactions. Moreover, it presents that minimal criteria of basket analysis in empirical tests must be low.