



Analiza koszykowa to cenne rozszerzenie narzędzi raportujących, wzbogacające standardowy zakres wyników o nowe spojrzenie na dane historyczne

Agnieszka Paszyła

Analiza koszykowa danych transakcyjnych – cele i metody

Celem analizy koszykowej jest wykrywanie zależności ukrytych w bazach danych i przedstawienie ich w postaci prostych reguł. Reguły te mogą dotyczyć na przykład zwyczajów zakupowych klientów lub prawidłowości w korzystaniu z usług wybranego typu. Otrzymane reguły noszą nazwę reguł asocjacyjnych.

Rozwinięciem analizy koszykowej jest analiza sekwencji, która pozwala na uwzględnienie powiązań zdarzeń w czasie. Kolejnym rozszerzeniem jest analiza połączeń, która zapewnia narzędzia do przedstawienia dużej ilości reguł w postaci czytelnych schematów.

Jak już wspomniano, analiza koszykowa służy do znajdowania w dużym zestawie danych ukrytych zależności w postaci prostych reguł. Pierwotnym zastosowaniem analizy koszykowej była analiza danych transakcyjnych pochodzących z supermarketów. Problemem, który postawili kierownicy hal przed analitykami było znalezienie i sformalizowanie prawidłowości, które z dużym prawdopodobieństwem opisują zależności między kupowanymi produktami. Taka wiedza pozwoliłaby pracownikom m. in. tak rozmieścić produkty w sklepie, aby uzyskać największe wyniki sprzedaży lub zaplanować promocje nie zmniejszając w sposób nieprzewidziany potencjalnego zysku. W odpowiedzi, do analizy danych historycznych zgromadzonych w bazach transakcyjnych, zastosowano reguły asocjacyjne wsparte odpowiednio szybkimi algorytmami przeszukiwania baz. Stąd analiza koszykowa pozwala m. in. odpowiedzieć na pytania:

- jakie produkty kupowane są najczęściej razem?
- jakie jest prawdopodobieństwo, że klienci, którzy kupili produkt A, kupią również produkt B?
- co kupują klienci, którzy uczestniczą w programach lojalnościowych?
- co wybrali klienci, którzy skorzystali z wybranej promocji?

Część z otrzymanych wyników potwierdza zwykle zależności, które znają pracownicy, jednak celem analizy koszykowej jest znalezienie „ukrytych” reguł, które nie są oczywiste i wzbogacają wiedzę specjalistów. Choć niewątpliwie, potwierdzenie

na podstawie danych historycznych hipotez dotyczących prawidłowości obserwowanych w danej branży i sformułowanych przez pracowników na podstawie doświadczenia, stanowi również istotną wartość.

Analiza koszykowa ma oczywiście szersze zastosowanie niż badanie koszyków klientów hipermarketów. Mianowicie, gdy przestajemy ograniczać się do postrzegania produktów w sensie fizycznym, na przykład kawy, pizzy, piwa, okazuje się, że nie ma ograniczeń, co do przedmiotu analizy koszykowej. W szczególności możemy rozszerzyć obszar zainteresowań badawczych o usługi. Przykładem może być:

- analiza usług pod kątem zastosowania marketingowych metod zwiększania sprzedaży (up-selling) i sprzedaży krzyżowej (cross-selling),
- optymalizacja pakietów usług, taryf i opłat,
- planowanie kampanii promocyjnych na podstawie uzyskanych wyników,
- weryfikacja efektywności i skuteczności kampanii marketingowych poprzez porównanie wyników analiz z kilku okresów,
- przeciwdziałanie rezygnacji klientów z usług wybranego dostawcy.

Obszary w których ma zastosowanie analiza koszykowa to m. in. analiza transakcji finansowych i ubezpieczeniowych, telekomunikacja, logistyka i farmaceutyka. We wszystkich tych dziedzinach celem analizy może być zarówno wykrywanie prawidłowości, jak i nadużyć.

Rozszerzeniem analizy koszykowej jest analiza sekwencji, która umożliwia uwzględnienie czasu w wykrywanych prawidłowościach. W przypadku danych transakcyjnych

pochodzących ze sklepów, analiza sekwencyjna jest możliwa tylko wtedy, gdy dostępne są dane osobowe klientów, na przykład na karcie umożliwiającej korzystanie z usług lub karcie lojalnościowej. Przykład zastosowania analizy koszykowej do zbadania zachowań użytkowników portalu internetowego można znaleźć w [2]. Połączenie numeru identyfikacyjnego klienta i przeprowadzonych przez niego transakcji pozwala na pogłębioną analizę zachowań i zwyczajów klientów.

Rozszerzając obszar zastosowań reguł asocjacyjnych należy również zwrócić uwagę na dane transakcyjne dotyczące działań klientów nie tylko w aspekcie zysku lub strat z ewentualnego zakupu. Dane, w szczególności takie, które pozwalają na śledzenie zachowań w czasie, pozwalają na przeprowadzenie analiz o większym zakresie merytorycznym. Przykładem może być tutaj wykrywanie wszelkich nieprawidłowości. Analiza koszykowa i sekwencji umożliwia poznawanie mechanizmów nadużyć ma przykład w bankowości, gdzie z danych transakcyjnych wyodrębniane są reguły opisujące pranie brudnych pieniędzy. Znalezione reguły opisują w sposób ilościowy kolejne etapy mechanizmu i mogą być wykorzystane w celu zapobiegania podobnym zjawiskom.

Z kolei analiza połączeń dostarcza narzędzi do prezentacji graficznej wykrytych prawidłowości, zarówno reguł statycznych, jak i reguł opisujących sekwencje zdarzeń.

Podsumowując zastosowania analizy koszykowej i sekwencji, wyniki uzyskane w postaci reguł asocjacyjnych, czyli wyodrębnionych prawidłowości i odpowiadających im prawdopodobieństw pomagają w poznaniu prawidłowości działań klientów i zdobyciu przewagi konkurencyjnej przedsiębiorstwa. Analiza koszykowa, sekwencji i połączeń może stanowić również cenne rozszerzenie narzędzi raportujących, wzbogacając standardowy zakres wyników o nowe spojrzenie na dane historyczne, a także może

wpłynąć na poziom zabezpieczeń przed nadużyciami.

Reguły asocjacyjne – ujęcie teoretyczne

Wynikiem analizy koszykowej są reguły asocjacji postaci:

JEŻELI [poprzednik] TO [następnik]

zapisywane za pomocą warunków:

[warunki poprzednika] => [warunki następnika]

na przykład:

[kawa, śmietanka] => [ciastka]

Przykładowa reguła dotyczy osób, które kupując kawę i śmietankę, kupią również ciastka. Przecinki stosowane w zapisie warunków *poprzednika (body)* lub *następnika (head)* odpowiadają spójnikowi „i”. Jeżeli transakcja (rekord), czyli pojedynczy przypadek ze zbioru danych „pasuje” do reguły, czyli spełnia wszystkie warunki poprzednika i następnika będziemy mówić, że reguła zawiera tę transakcję lub, że *transakcja wspiera regułę asocjacji*. Zazwyczaj pożądane jest znalezienie tych reguł, w których *następnik* wyraża się jednym warunkiem (lub niewielką ich liczbą). W literaturze anglojęzycznej stosowany jest termin *itemset* lub *k-itemset* (rzadziej *item set*) – nie oznacza on zbioru przypadków w próbie uczącej, lecz zbiór elementów reguły lub transakcji. Przykładowo: zestaw {kawa, śmietanka} to *2-itemset* (zestaw dwuelementowy) i w tym przypadku opisuje on zawartość koszyka.

Rodzaje reguł asocjacyjnych

Wśród reguł asocjacyjnych wyróżniamy takie, które oparte są na zmiennych jakościowych lub ilościowych, które dotyczą jednego lub więcej wymiarów (cech, atrybutów) danych oraz reguły, które dotyczą jednego lub więcej poziomów agregacji zmiennych (chodzi tu na przykład o kategorie produktów, branże).

Reguły oparte na zmiennych jakościowych i ilościowych

W zależności od postaci danych wejściowych, możemy mieć do czynienia z różnymi zmiennymi odzwierciedlającymi te same dane. Najczęściej dla pojedynczej transakcji dysponujemy zbiorem wartości opisujących tę transakcję. Możemy również zawartość koszyka przedstawić za pomocą tabeli, w której kolejne zmienne oznaczają towary, które można nabyć. Zmienne te będą miały charakter binarny, na przykład, osoba kupiła chleb (1) lub nie kupiła (0). Te same informacje możemy zapisać w postaci tekstowej *tak/nie*. Tego typu zmienne określamy mianem jakościowych. Oczywiście, mogą one przyjmować więcej niż dwie wartości – na przykład dla wielkości opakowania lub wersji produktu.

Zmienne jakościowe mają w analizie koszykowej dwa zastosowania. Przede wszystkim, mogą dotyczyć cech niemierzalnych, które chcemy uwzględnić w regułach np. płci klienta, wykształcenia, nazwy nabytego produktu, koloru, itp. Zmienne jakościowe mogą też być utworzone sztucznie, w celu zakodowania informacji o tym, czy klient nabył dany produkt (zmienna binarna), w sytuacji, gdy nie są nam potrzebne ilości kupionych towarów.

Zmienne ilościowe opisują ilość lub wartość towaru (usługi) i są wykorzystywane wtedy, gdy chcemy w sposób precyzyjniejszy sformułować zależności dotyczące atrybutów.

Oczywiście, wyszukiwanie reguł dla wszystkich realizacji zmiennej ilościowej jest nieefektywne. W szczególności, gdy zmienna posiada duży obszar zmienności (np. masa produktu mierzona z dokładnością do trzech miejsc po przecinku) możemy mieć do czynienia z bardzo dużą liczbą jej realizacji i w efekcie z przytłaczającą liczbą potencjalnych reguł asocjacyjnych. Dlatego obszar zmienności zmiennej dzielony jest na przedziały i w oparciu o nie tworzone są reguły asocjacyjne. W celu utworzenia optymalnych przedziałów wartości zmiennych ilościowych potrzebnych do uzyskania

reguł asocjacyjnych, które najlepiej będą odzwierciedlać zależności między kupowanymi produktami i czynnikami zewnętrznymi, stosowane są metody klasyfikacyjne, takie jak analiza skupień i inne. Są to jednak zadania trudne do oprogramowania. Klasyczna analiza koszykowa, zapoczątkowana w latach dziewięćdziesiątych m.in. przez Agrawala, Srikanta i Imielińskiego [1] dotyczy jedynie zmiennych jakościowych. Narzędziem, które pozwala na przeprowadzenie analizy koszykowej i sekwencji z wykorzystaniem zmiennych jakościowych i ilościowych jest moduł STATISTICA Sequence, Association and Link Analysis (www.statsoft.pl/products/sal.html).

Liczba wymiarów uwzględnianych w regułach asocjacyjnych

Przez wymiar reguły asocjacyjnej rozumiemy liczbę uwzględnionych zmiennych (atrybutów). Reguła, która dotyczy jednej cechy jest określana jako jednowymiarowa. Przykładem reguły jednowymiarowej może być reguła postaci: [komputer, drukarka] => [oprogramowanie]. Przykładowa reguła opisuje prawidłowość, że klienci, którzy kupują komputer i drukarkę, z obliczonym prawdopodobieństwem kupują również oprogramowanie. Zarówno warunki poprzednika jak i następnika dotyczą tego samego atrybutu czyli tego, co zawiera transakcja lub inaczej, co kupił klient.

Warunki reguły wielowymiarowej uwzględniają więcej niż jeden atrybut. Reguła wielowymiarowa może określać zależności między, na przykład, wiekiem klienta, godziną zakupów i tym, co klient nabył: [wiek_klienta ∈ <14; 19>, godzina_zakupów ∈ <10; 15>] => [produkt = napoje, słodycze]. Na podstawie powyższej reguły dowiadujemy się, że klienci w wieku od 14 do 19 lat, którzy przychodzą do sklepu w godzinach od 10 do 15, z określonym prawdopodobieństwem kupują napoje i słodycze.

Reguły dotyczące danych zagregowanych i surowych

Jak już wspomniano, analiza koszykowa może być prowadzona na dowolnym

poziomie ogólności. To znaczy, że reguły asocjacyjne mogą dotyczyć zależności między kupowanymi produktami, kategoriami produktów lub między branżami. Ogólnie, w celu wyszukiwania prawidłowości w zbiorze danych, powinniśmy szukać najpierw silnych reguł, czyli takich prawidłowości, które występują z dużym prawdopodobieństwem w odniesieniu do danych zagregowanych. Przykładowo, w pierwszej kolejności badamy jak często kupowane są razem pieczywo i nabiał, i jeśli okaże się, że jest to silna reguła, to warto zbadać szczegółowe zależności w tych kategoriach produktów, na przykład pomiędzy niskotłuszczowym mlekiem i jogurtami, a ciemnym pieczywem, itp.

Miary jakości uzyskanych reguł

Wybierając postać danych wejściowych musimy pamiętać, że potencjalnych reguł jest bardzo dużo. Na przykład w przypadku trzech zmiennych dichotomicznych (odpowiedzi typu TAK-NIE na trzy pytania) możemy otrzymać maksymalnie 650 reguł. Jest to liczba wariacji bez powtórzeń dla trzech zmiennych i dwóch możliwych wartości, i nie ma w niej bezużytecznych powtórzeń typu [ODP1=NIE => ODP1=NIE], czyli na

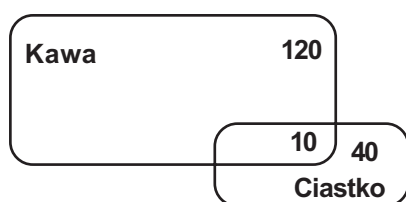
przykład, jeśli klient nie kupił produktu A, to nie kupił również produktu B. Dla supermarketu, gdzie mamy do czynienia z tysiącami produktów w ofercie, trudno wyrazić słowami liczbę możliwych reguł, ponieważ wzrost tej liczby jest typu n!. Interesujące są oczywiście tylko te reguły, które występują często w danych historycznych (próbie uczącej), czyli opisują często występujące zachowania, a nie są jedynie pustymi sformułowaniami. Dlatego, aby wyodrębnić te reguły, które niosą dla nas istotną informację wykorzystujemy trzy parametry służące do oceny „ważności” reguł. Są to:

- wsparcie reguły (support) - odsetek transakcji w danych historycznych, które zawierają wybraną regułę, jest to prawdopodobieństwo kupienia danego produktu przez losowo wybranego klienta,
- pewność reguły (confidence) - odsetek transakcji zawierających analizowaną regułę w zbiorze tych, które zawierają poprzednik (dla reguły A=>B odpowiada to prawdopodobieństwu warunkowemu P(B|A)), jest to prawdopodobieństwo, że losowo wybrany klient, który nabył produkt A, kupi również produkt B,
- korelacja (correlation) - jest to mia-

wsparcie (A) =	$\frac{\text{częstość występowania A}}{\text{liczba transakcji w zbiorze}}$
wsparcie (C) =	$\frac{\text{częstość występowania C}}{\text{liczba transakcji w zbiorze}}$
pewność (jeśli A to C) =	$\frac{\text{wsparcie (jeśli A to C)}}{\text{wsparcie (A)}}$
pewność (jeśli C to A) =	$\frac{\text{wsparcie (jeśli C to A)}}{\text{wsparcie (C)}}$
korelacja (jeśli A to C) =	$\frac{\text{wsparcie (jeśli A to C)}}{\text{pewność (jeśli A to C)}}$
przyrost (jeśli A to C) =	$\frac{\text{pewność (jeśli A to C)}}{\text{wsparcie (C)}}$

Tabela 1. Miary jakości reguł asocjacyjnych

BUSINESS INTELLIGENCE



Rysunek 1. Ilustracja graficzna miar jakości reguł asocjacyjnych

niez produkt B,

- przyrost (lift) - jest to modyfikacja korelacji reguły; informuje o tym, jaki jest wpływ produktu A na sprzedaż produktu B (lub występowanie zjawiska B).

Parametry te objaśnimy na przykładzie dwóch prostych reguł. Załóżmy dla uproszczenia, że zbiór zawiera 300 transakcji i koncentrujemy się nad zależnościami między zakupem kawy i ciastek. Liczba transakcji zawierających jeden lub drugi produkt wynosi 180 i są one przedstawione na Rysunku 1.

Wsparcie (support) dla reguł Kawa => Ciastko i Ciastko => Kawa wynosi $180/300 = 0,6$ czyli 60%. Pewność reguły Kawa => Ciastko obliczamy następująco: $10/130 \approx 0,077$, i wynosi 7,7 %. Pewność reguły Ciastko => Kawa wynosi: $10/50 \approx 0,2$, czyli 20 %. Oznacza to, że osoby, które kupowały ciastka częściej kupowały kawę, niż osoby kupujące kawę sięgały po ciastka. Inaczej, reguła Ciastko => Kawa jest częściej spotykana niż Kawa => Ciastko. Korelacja reguły Kawa => Ciastko wynosi

0,6/(0,433*0,166) $\approx 8,33$. Miara ta jest w ten sposób skonstruowana, że jeśli przyjmuje wartości większe od 1, to wskazuje na pozytywną korelację między produktami. W przeciwnym wypadku możemy mówić o korelacji negatywnej. W naszym przykładzie, możemy powiedzieć, że zakup kawy zwiększa szansę zakupu ciastek. Natomiast przyrost reguły Kawa => Ciastko określa jak zakup kawy zwiększa prawdopodobieństwo zakupu ciastka. Przyrost reguły obliczamy dzieląc pewność reguły Ciastko => Kawa przez prawdopodobieństwo zakupu ciastka (wsparcie produktu): $0,2/0,277 = 0,72$. Interpretując otrzymany wynik możemy powiedzieć, że jeśli klient kupuje kawę, to szansa, że kupi ciastko wzrasta o 72%.

0,6/(0,433*0,166) $\approx 8,33$. Miara ta jest w ten sposób skonstruowana, że jeśli przyjmuje wartości większe od 1, to wskazuje na pozytywną korelację między produktami. W przeciwnym wypadku możemy mówić o korelacji negatywnej. W naszym przykładzie, możemy powiedzieć, że zakup kawy zwiększa szansę zakupu ciastek. Natomiast przyrost reguły Kawa => Ciastko określa jak zakup kawy zwiększa prawdopodobieństwo zakupu ciastka. Przyrost reguły obliczamy dzieląc pewność reguły Ciastko => Kawa przez prawdopodobieństwo zakupu ciastka (wsparcie produktu): $0,2/0,277 = 0,72$. Interpretując otrzymany wynik możemy powiedzieć, że jeśli klient kupuje kawę, to szansa, że kupi ciastko wzrasta o 72%.

Literatura

- [1] R. Agrawal, T. Imieliński, A. Swami „Mining association rules between sets of items in large databases”, Proceedings of ACM SIGMOD International Conference on Management of Data, Washington DC , May 1993
- [2] R. Kita „Analiza sposobu poruszania się użytkowników po portalu internetowym”, w „Data mining – metody i przykłady”, StatSoft Polska 2002 (artykuł dostępny na stronie www.statsoft.pl/czytelnia/dm/wstepdm.html) ■

Agnieszka Pasztyła pełni funkcję konsultanta w firmie Statsoft Polska i pracuje na stanowisku asystenta w Katedrze Statystyki Akademii Ekonomicznej w Krakowie. Zajmuje się modelowaniem ilościowym i prognozowaniem procesów biznesowych. Kontakt: a.pasztyla@statsoft.pl.