PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1107 — 2006
Badania marketingowe w przestrzeni europejskiej

Mariusz Łapczyński
Akademia Ekonomiczna w Krakowie

ANALIZA KOSZYKOWA JAKO KLASYCZNY PRZYKŁAD WYKORZYSTANIA REGUŁ SKOJARZENIOWYCH

1. Wprowadzenie w zagadnienie reguł skojarzeniowych

Skojarzenia (*associations*) to jeden z 6 modeli *data mining*, zaliczany do grupy modeli rozpoznawanych bez nauczyciela lub inaczej do taksonomii bezwzorcowej. Modele skojarzeniowe przedstawiają współwystępowanie wartości różnych zmiennych w danym przypadku. Modele skojarzeniowe (asocjacyjne) mają postać zdań warunkowych, w których pojawia się spójnik międzyzdaniowy: *jeżeli zdanie* Z_1, *to zdanie* Z_2. Używając tego spójnika w mowie potocznej przyjmuje się, że między zdaniami składowymi istnieje powiązanie rzeczowe lub formalne, tzn. pierwsze zdanie Z_1 implikuje drugie Z_2. Z punktu widzenia logiki, związki między poprzednikiem Z_1 (*antecedent*) a następnikiem Z_2 (*consequent*) mogą mieć różnoraki charakter, jednak w przypadku badań rynkowych i marketingowych mowa o związkach przyczynowo-skutkowych i strukturalnych (tj. takich, które wynikają z rozmieszczenia przedmiotów w przestrzeni albo zdarzeń w czasie). Choć podobną postać zdania warunkowego mają reguły indukcyjne, to te należą jednak do narzędzi umożliwiających budowę modeli dyskryminacyjnych, czyli rozpoznawanych z nauczycielem (wzorcowych).

Reguła skojarzeniowa przyjmuje formę $A \Rightarrow B$, gdzie A i B to zbiory atrybutów. Zapis ten oznacza, że jeśli w danym przypadku wartość zmiennej A przyjmuje 1 (prawda), to wartość zmiennej B również przyjmuje z określonym prawdopodobieństwem wartość 1 (prawda). Przykład takiej reguły może być następujący: „40% klientów, którzy kupili mrożoną pizzę, nabyło również keczup. 15% wszystkich klientów nabyło oba te produkty jednocześnie." W tym przypadku zmienna A = klient kupił mrożoną pizzę, a zmienna B = klient kupił keczup. 40% to współ-

czynnik *confidence*, a 15% to współczynnik *support*, czyli udział tego typu transakcji we wszystkich transakcjach w danej placówce handlowej.

Pisząc o tych miarach, warto posłużyć się terminologią rachunku prawdopodobieństwa. Współczynnik *confidence* to inaczej prawdopodobieństwa warunkowe $P(B \mid A)$. Interpretując formalnie, trzeba by powiedzieć: „prawdopodobieństwo zdarzenia A obliczone przy założeniu, że zaszło zdarzenie B". Adaptując to na potrzeby analizy koszykowej, można powiedzieć: „prawdopodobieństwo zakupu produktu A obliczone przy założeniu, że zakupiono również produkt B". Współczynnik *support* to z kolei prawdopodobieństwo koniunkcji zdarzeń P(AB), czyli prawdopodobieństwo łącznego zajścia dwu zdarzeń. Interpretacja tej miary jest znacznie łatwiejsza, chodzi bowiem o udział transakcji, w których kupiono A i B jednocześnie, w zbiorze wszystkich transakcji. Problemem analityka jest znalezienie wszystkich reguł skojarzeniowych spełniających minimalną (przyjętą przez badacza lub menedżera) wartość *support* i minimalną wartość *confidence*.

2. Analiza koszykowa (*market basket analysis*)

Typowe problemy decyzyjne menedżerów super- i hipermarketów związane są z wyborem produktów, które powinny znaleźć się na sali, projektowaniem gazetek reklamowych czy rozmieszczeniem asortymentu na półkach. Analiza transakcji z przeszłości jest powszechnie stosowanym podejściem, poprawiającym jakość tych decyzji. Postęp w technologii kodów kreskowych umożliwił przechowywanie tzw. danych koszykowych – opartych na pojedynczych transakcjach. Dane koszykowe nie zawsze zawierają pozycje (produkty) nabyte razem w danej chwili. Mogą również zawierać produkty nabyte przez konsumenta w określonym przedziale czasowym. Przykładem mogą być miesięczne zakupy członków klubu książki lub sprzedaż odzieży przez firmy wysyłkowe[1].

Najpopularniejszą aplikacją reguł skojarzeniowych jest analiza koszykowa (*market basket analysis*). Zmienna (atrybut) nosi tutaj nazwę pozycji (*item*), a przypadek – transakcji. Tabela relacyjna to macierz, w której kolumnach znajdują się pozycje (produkty), a w wierszach transakcje. Celem analizy koszykowej jest [1, s. 208]:

1) znalezienie wszystkich reguł, które w następniku zawierają konkretny produkt (np. wędzonego łososia) – ustalenie takich wzorców zakupów pozwoliłoby na optymalizację oferty w celu zwiększenia sprzedaży produktu z następnika;

2) znalezienie wszystkich reguł, które w poprzedniku zawierają konkretny produkt (np. sos tatarski) – reguły te pozwoliłyby określić grupę produktów, których wielkość sprzedaży mogłaby ulec obniżeniu, gdyby sklep zaprzestał sprzedaży produktu z poprzednika;

[1] Mamy wtedy do czynienia z innym bezwzorcowym modelem *data mining* – z sekwencjami.

3) znalezienie wszystkich reguł wiążących pozycje (produkty) umieszczone na półce A i pozycje umieszczone na półce B – identyfikacja takich wzorców ułatwiłaby rozmieszczenie asortymentu w danej placówce handlowej oraz skuteczną promocję produktów.

Mimo że reguły skojarzeniowe są często utożsamiane z analizą koszykową, to jednak znajdują również zastosowanie w innych obszarach badań rynkowych i marketingowych:

- w analizie zakupów dokonywanych za pomocą karty kredytowej,
- w badaniach zakupów usług telekomunikacyjnych,
- w analizie produktów bankowych nabywanych przez klientów detalicznych,
- w identyfikacji oszustw klientów firm ubezpieczeniowych [2, s. 124].

3. Rodzaje reguł skojarzeniowych w analizie koszykowej

Mimo swej przejrzystości i łatwości w interpretacji reguły skojarzeniowe nie zawsze są przydatne do usprawnienia działań marketingowych firmy. Wyróżnia się trzy typy reguł skojarzeniowych:

1) reguły użyteczne (*useful rules*) – najbardziej wartościowe z praktycznego punktu widzenia, odkrywają nieznane wcześniej wzorce zakupów, np. „jeżeli kupi 2 kg cukru, to kupi 4 butelki piwa";

2) reguły trywialne (*trivial rules*) – to klasyczne wzorce transakcji, znane każdemu pracownikowi z danej branży, np. „jeżeli kupi węgiel drzewny, to kupi podpałkę do grilla", tu należy dodać, że reguł nie uznaje się za trywialne, jeśli analiza koszykowa wykorzystana została do oceny skuteczności akcji promocyjnej, sprawdza się wtedy, czy oczekiwania dotyczące łącznego nabywania pewnych produktów, potwierdziły się;

3) reguły niewytłumaczalne (*inexplicable rules*) – to faktyczne i nieoczekiwane wzorce zakupów, które nie przekładają się na działalność marketingową, np. „jeżeli klient przychodzi w dzień otwarcia sklepu, to kupuje odświeżacz powietrza".

Czasami do analizy włącza się tzw. wirtualne pozycje (*virtual items*), które zwykle są tożsame z okresem, w którym kupiono dany produkt, np. „jeżeli kupił ziemię do kwiatów i jest maj, to kupi również rękawice ogrodowe".

Ogólnie rzecz ujmując, można wyróżnić dwa rodzaje reguł skojarzeniowych: jakościowe i ilościowe. W przypadku jakościowych reguł skojarzeniowych (zwanych też boolowskimi lub określanymi angielskim zwrotem *Boolean association rules*) tabela relacyjna zawiera zmienne binarne, czyli wartości 1 lub 0; gdzie 1 oznacza, że dana pozycja występuje w transakcji, a 0 oznacza, że dana pozycja nie występuje w transakcji. Wyszukiwanie reguł skojarzeniowych polega na poszukiwaniu związków między jedynkami w tabeli relacyjnej. Proces ten przebiega w dwóch etapach: przeszukiwania całej bazy danych i generowania reguł skojarzeniowych. Przykład reguły boolowskiej może wyglądać następująco: *pieczywo = tak* $\Rightarrow$ *masło = tak*. Zdarza się, że tabele relacyjne zawierają wielokategorialne zmien-

ne jakościowe i zmienne ilościowe. Otrzymane reguły noszą wtedy nazwę ilościowych reguł skojarzeniowych (*quantitative association rules*). W celu zredukowania liczby wartości zmiennej ilościowej łączy się je w przedziały, co niestety wiąże się z problemem minimalnej wartości współczynnika *confidence* (*MinConf*) i problemem minimalnej wartości współczynnika *support* (*MinSup*)[2].

Inną klasyfikacją reguł jest ich podział na jednowymiarowe reguły skojarzeniowe (*single-dimensional association rules*) i wielowymiarowe reguły skojarzeniowe (*multi-dimensional association rules*). Wymiarowość wiąże się z liczbą pozycji w poprzedniku lub następniku reguły. Przykład reguły jednowymiarowej może być następujący: *chleb* $\Rightarrow$ *mleko*, a przykład reguły wielowymiarowej następujący: *chleb i ser żółty* $\Rightarrow$ *mleko*. O dwuwymiarowości tej drugiej zdecydowało pojawienie się w poprzedniku, obok chleba, żółtego sera. W tym miejscu należy wspomnieć o pewnej ciekawostce, jaką są reguły dysocjacyjne (*dissociation rules*). One również mają kilka wymiarów, jednak inny jest spójnik łączący wymiary poprzednika (następnika): *jeżeli A i nie B, to C*.

W 1995 r. opracowano algorytm do eksploracji wielopoziomowych reguł skojarzeniowych. Stąd kolejny [4] podział reguł na jednopoziomowe reguły skojarzeniowe (*single-level association rules*) i wielopoziomowe reguły skojarzeniowe (*multi-level association rules*). Różnica między pierwszymi i drugimi sprowadza się do stopnia szczegółowości opisu poprzednika i następnika. Przykład jednopoziomowej reguły skojarzeniowej może wyglądać następująco: „jeśli kupi mleko, to kupi serek homogenizowany" (*mleko* $\Rightarrow$ *serek homogenizowany*). Natomiast przykład wielopoziomowej reguły skojarzeniowej wygląda tak: „jeśli kupi 2-procentowe mleko łaciate w półlitrowym kartonie, to kupi serek homogenizowany o smaku waniliowym firmy Danone" (*2% mleko „Łaciate" w półlitrowym kartonie* $\Rightarrow$ *serek homogenizowany o smaku waniliowym firmy Danone*). Autorzy algorytmu wykorzystali fakt, że kody kreskowe umieszczone na opakowaniach produktów zawierają szereg informacji, np. rodzaj towaru, objętość opakowania, nazwę producenta.

4. Podsumowanie

Pierwszy algorytm do generowania reguł skojarzeniowych powstał w 1994 r. pod nazwą Apriori. Wyszukiwał jedynie boolowskie reguły skojarzeniowe, ale jednocześnie nie wymagał dużej mocy obliczeniowej komputerów. Większość późniejszych prac dotyczących algorytmów do reguł skojarzeniowych bazowała właśnie na Apriori. Kolejnym krokiem usprawniającym wynajdywanie asocjacji był algorytm z 1996 r. o nazwie „Rozszerzony Apriori" (*Extended Apriori Algorithm*). Modyfikacja była znaczna, pozwalała bowiem na generowanie reguł ilościowych. W tym samym roku opracowano algorytmy do generowania reguł wielo-

[2] Niestety, ograniczona objętość pracy i złożoność zagadnienia nie pozwalają, aby w tym miejscu wyjaśnić ten mechanizm. Szczegóły można znaleźć w pracy [7].

wymiarowych oraz zaczęto prace nad algorytmami równoległymi, znacznie przyspieszającymi obliczenia.

W chwili obecnej reguły skojarzeniowe doczekały się kilkunastu implementacji. Oprogramowanie dostępne na rynku to m.in.: Apriori, ARtool, Azmy SuperQuery, CBA, IBM Intelligent Miner, IREX, Magnum Opus, SPSS Clementine, SRA KDD Explorer Suite, STATISTICA Data Miner, Wiz Rule, XAffinity, Xpertrule Miner.

Literatura

[1] Agrawal R., Imielinski T., Swami A., *Mining Association Rules Between Sets of Items in Large Databases. In Proceedings of the ACM SIGMOD* Conference on Management of Data, Washington D.C., maj 1993.

[2] Berry M.J.A., Linoff G., *Data Mining Techniques for Mar eting Sales and Customer Support*, John Wiley & Sons, New York 1997.

[3] Cengiz I., *Mining Association Rules*, Bilkent University, Department of Computer Engineering and Information Sciences, Ankara, 1997.

[4] Han J., Fu Y., *Discovery of Multiple-Level Association Rules from Large Databases*, Proceedings of 21th International Conference on Very Large Data Bases, Zurich, wrzesień 1995.

[5] Han J., Kamber M., *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Francisco 2001.

[6] Łapczyński M., *Wprowadzenie do data mining*, Zeszyty Naukowe AE nr 640, Kraków 2003.

[7] Łapczyński M., *Wy orzystanie analizy reguł s ojarzeniowych w badaniach ryn owych i mar etingowych.* [w:] Badania statutowe nr 45/KARiBM/2/2003/S pod kier. Prof. dr hab. S. Mynarskiego pt. *„Analiza współzależności i intera cji w badaniach ryn owych i mar etingowych"*, Kraków 2003.

[8] Srikant R., Agrawal R., *Mining Quantitative Association Rules in Large Relational Tables. In Proceedings of the ACM SIGMOD*, Conference on Management of Data, czerwiec 1996.

MARKET BASKET ANALYSIS AS A CLASSICAL IMPLEMENTATION OF ASSOCIATION RULES

Summary

This article presents popular data mining tool – association rules. The author demonstrates common measures of quality i.e. support and confidence. He focuses on the main research area – market basket analysis and describes all types of rules (qualitative and quantitative, single-level and multi-level, single-dimensional and multi-dimensional). Furthermore, problems of minimum support and minimum confidence are mentioned. The article includes a list of several computer programmes – implementations of association rules.