Artur GRAMACKI, Jarosław GRAMACKI**

Nowa metoda grupowania danych koszyka sklepowego

Eksploracja danych (*data mining*) jest obecnie bardzo intensywnie rozwijającą się dziedziną wiedzy. Głównym motorem napędowym jest tutaj gromadzenie przez ludzi coraz to większych ilości danych (np. typowy supermarket rejestruje dziennie dziesiątki tysięcy operacji sprzedaży), które coraz trudniej jest analizować za pomocą metod znanych z klasycznych baz danych (zapytanie, podsumowanie, zestawienie itp.). Można zaryzykować twierdzenie, że gdy ilość danych zaczyna przekraczać pewną wartość krytyczną, stają się one praktycznie bezwartościowe (szukanie igły w stogu siana). Użytkownicy zaczynają „tonąć" w tym ogromie i aby jakoś rozwiązać problem, należy opracować zupełnie inne metody analizowania zgromadzonych danych. Ponieważ współczesne systemy bazodanowe są bardzo wydajne i pojemne, stąd rzeczywistym problemem jest nie to, jak je gromadzić, ale jak z nich efektywnie korzystać.

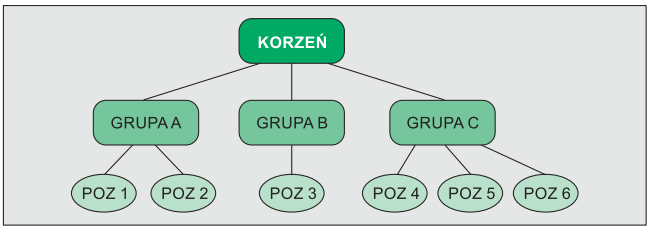
Literatura na temat eksploracji danych jest bardzo bogata. Oprócz tradycyjnie anglojęzycznych pozycji, np. [1], [3], [6], zaczynają również pojawiać się książki w języku polskim [4], [5]. Pierwsza pozycja [4] jest dość zaawansowana, jednak bardzo obszerna i dokładna, druga natomiast [5] stanowi bardzo przystępne wprowadzenie w zagadnienie eksploracji danych. Pozycja [3] może być uznana za klasyczną w dziedzinie eksploracji danych, omawia praktycznie wszystkie obszary, znajdujące się w polu zainteresowania naukowców. Na uwagę zasługują również bardzo obszerne i doskonale opracowane prezentacje w *PowerPoint*, dotyczące poszczególnych rozdziałów tej książki.

W artykule skupiono się na jednym wybranym aspekcie eksploracji danych, a mianowicie ich grupowaniu (czasami, choć chyba niezbyt poprawnie językowo, nazywanym też klastrowaniem danych). Samo zagadnienie grupowania znane jest od bardzo dawna i stanowi jeden z ważniejszych działów statystyki, jednak w dziedzinie eksploracji danych nabiera zupełnie nowego znaczenia. Najogólniej mówiąc, celem grupowania jest automatyczne (lub prawie automatyczne) pogrupowanie zgromadzonych danych w klasy zgodnie z przyjętą miarą podobieństwa. Do tej pory opracowano bardzo wiele technik i metod grupowania [2], w niniejszym artykule skupiono się tylko na jednej z nich (metoda hierarchiczna aglomeracyjna) i zastosowano ją do grupowania specyficznego rodzaju danych, zwanego w literaturze koszykiem sklepowym (*market-basket*).

ANALIZA KOSZYKA SKLEPOWEGO

Analiza koszyka sklepowego jest jednym z ważniejszych problemów rozważanych w eksploracji danych. Polega ona na analizie danych, zawierających informacje o zakupach dokonywanych przez klientów supermarketu (lub innego tego typu sklepu). Najczęściej takie dane są przedstawiane w tabeli, gdzie w wierszach umieszcza się identyfikatory poszczególnych transakcji, a w kolumnach identyfikatory oferowanych towarów. Ponadto poszczególne towary są zwykle pogrupowane i tworzą pewną taksonomię. Na przykład dwa produkty: „mleko" oraz „jajka" należą do grupy „nabiał", a ta z kolei do grupy „artykuły spożywcze". Przykładową taksonomię w postaci drzewa pokazano na rys. 1.

*** Uniwersytet Zielonogórski, Instytut Informatyki i Elektroniki, e-mail: A.Gramacki@iie.uz.zgora.pl, J.Gramacki@iie.uz.zgora.pl**



Rys. 1. Prosty przykład drzewa taksonomii

Prosty przykład koszyka sklepowego pokazano w tabeli 1a. W ostatniej kolumnie zaznaczono dodatkowo wartości netto poszczególnych transakcji. Znaczenie tej kolumny zostanie wyjaśnione w kolejnym rozdziale. Tam też będzie omówiona tabela 1b. Występowanie jedynki w wierszu *t* oraz kolumnie *p* oznacza, że transakcja *t* zawiera produkt *p*. Warto zwrócić uwagę, że w przypadku rzeczywistych danych tabela (macierz) będzie bardzo rzadka (jedynek bardzo mało w stosunku do pustych pól; zwykle ułamek procenta). Jest to oczywiście spowodowane tym, że oferta towarowa typowego marketu (lub innego wielkiego sklepu) jest bardzo duża (wiele tysięcy pozycji), podczas gdy w typowym koszyku klienta znajduje się zwykle nie więcej, niż kilkadziesiąt pozycji.

Tabela 1. Przykład koszyka sklepowego

(a)

	p_1	p_2	p_3	p_4	p_5	p_6	netto
t_1	1					1	200
t_2					1		100
t_3				1			300
t_4			1				300
t_5		1					100
t_6	1						200

(b)

	p_{12}	p_3	p_{456}	netto
t_1	1		1	200
t_2			1	100
t_3			1	300
t_4		1		300
t_5	1			100
t_6	1			200

Zwykle poszukuje się tzw. reguł asocjacyjnych (*association rules*), obrazujących, które produkty są najczęściej kupowane razem. Tym samym odkrywa się tzw. wzorce zachowań klientów. Typowy przykład odkrytych reguł asocjacyjnych może wyglądać następująco: klienci, którzy kupują komputer, decydują się również na zakup programu antywirusowego. Literatura na temat odkrywania reguł asocjacyjnych jest bardzo obszerna. Więcej informacji można np. znaleźć w [3], [5]. Problem odkrywania reguł asocjacyjnych nie jest jednak głównym tematem tej publikacji, więc ograniczono się tu jedynie do wspomnienia o nim. W artykule zajęto się innym problemem. Chodzi mianowicie o grupowanie danych koszyka sklepowego. Motywacja jest następująca. Właściciel sklepu często jest zainteresowany nie szczegółami określonych transakcji, ale tym, czy nie można ich w jakiś sposób pogrupować. Każda transakcja jest w gruncie rzeczy obrazem preferencji klienta, który ją wygenerował. Istnieje wówczas *de facto* możliwość pogrupowania klientów, a nie tylko transakcji. Stwarza to możliwość precyzyjniejszego kierowania różnych akcji promocyjnych do dość dokładnie określonych grup klientów. Oczywiście trzeba pamiętać, że typowy supermarket nie gromadzi danych osobowych o nich i w takiej sytuacji wspomniana wyżej akcja promocyjna może być trudna do przeprowadzenia. Pozostaje jednak pewna furtka w postaci coraz powszechniej rozwijanych tzw. programów lojalnościowych. Zwykle polega to na

tym, że klient dobrowolnie godzi się na ujawnienie pewnych danych osobowych w zamian ze pewne korzyści zapewniane przez sklep (np. promocyjne ceny, ale nie tylko). Wobec tego założonym celem będzie automatyczne identyfikowanie *podobnych transakcji* w celu identyfikacji *podobnych klientów*. Oprócz samych danych z koszyka sklepowego należy wykorzystywać również dane na temat taksonomii produktów oraz dane o *wartościach netto* poszczególnych transakcji.

ALGORYTM

Jak wspomniano wcześniej, celem jest poszukiwanie podobnych transakcji (a tym samym podobnych klientów). Trzeba więc na wstępie zdefiniować, według jakiej miary będą porównywane poszczególne transakcje. Ponieważ model koszyka sklepowego operuje wyłącznie na danych binarnych (rys. 1), naturalną miarą podobieństwa transakcji będzie np. tzw. współczynnik Jaccarda [3]. Mierzy on stopień, w jakim dwie transakcje, *A* i *B*, pokrywają się. Jest on zdefiniowany jako:

$$s_{AB} = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Dane z koszyka sklepowego mogą być traktowane jako tzw. dane asymetryczne (fakt występowania w określonej transakcji pewnego produktu jest dużo istotniejszy, niż fakt, że jakiś inny produkt w niej nie występuje). Dlatego też *współczynnik niepodobieństwa* transakcji t_i oraz t_j jest zdefiniowany jako:

$$dsim(t_i, t_j) = \frac{b+c}{a+b+c} \tag{2}$$

gdzie: *a* – całkowita liczba produktów, które występują zarówno w transakcji t_i, jak i t_j; *b* – całkowita liczba produktów, które występują w transakcji t_i, a nie występują w transakcji t_j; *c* – całkowita liczba produktów, które nie występują w transakcji t_i, a występują w transakcji t_j. Wobec powyższego *współczynnik podobieństwa* transakcji t_i oraz t_j może być zdefiniowany jak poniżej:

$$sim(t_i, t_j) = 1 - dsim(t_i, t_j) = \frac{a}{a+b+c} \tag{3}$$

Do właściwego grupowania danych będzie zastosowana klasyczna hierarchiczna metoda aglomeracyjna. Wówczas typowo stosowanymi miarami podobieństwa tworzonych grup (klastrów) są: minimalna, maksymalna oraz średnia odległość między grupami (*single-linkage, complete-linkage and average-linkage algorithms*). W miarach tych oblicza się minimalną/maksymalną/średnią odległość między dwoma dowolnymi punktami p_1 oraz p_2, gdzie p_1 należy do grupy G_1, a p_2 do grupy G_2.

W tym miejscu należy zauważyć, że próba grupowania danych koszyka sklepowego tylko z zastosowaniem metody hierarchicznej nie da dobrych wyników. Spowodowane jest to tym, że dane są bardzo rzadkie i w praktyce nieczęsto zdarza się natrafić na dwie transakcje, które będą podobne do siebie w sensie współczynnika Jaccarda. Jeżeli jednak połączy się grupowanie hierarchiczne z dodatkowymi posiadanymi danymi, tj. drzewem taksonomii oraz wartościami netto poszczególnych transakcji, można spodziewać się lepszych wyników grupowania. Mówiąc o lepszym grupowaniu, rozumie się lepszą identyfikację podobnych klientów (podobnych transakcji).

Założono, że dwaj klienci będą podobni, jeżeli ich koszyki zawierają podobne towary i wartości netto tych koszyków są też podobne. Biorąc pod uwagę drzewo taksonomii, można ponadto przyjąć, że dwa produkty są podobne, jeżeli należą do tego samego węzła w drzewie taksonomii (przykładowo pozycja 1 oraz 2 na rys. 1 są podobne). Dla uproszczenia bierze się pod uwagę tylko najbliższy węzeł, choć oczywiście w ogólności drzewo taksonomii może mieć więcej niż jeden poziom. W efekcie uzyskuje się znaczną redukcję wymiarowości macierzy koszyka sklepowego. W tabeli 1b pokazano zredukowaną macierz z tabeli 1a po uwzględnieniu drzewa taksonomii z rys. 1. Ponadto będą również wykorzystywane informacje o wartościach netto poszczególnych transakcji – widać je w tabeli 1a oraz 1b (ostatnia kolumna). Można więc zaproponować następujący wzór do obliczania podobieństwa transakcji:

$$sim(t_i, t_j) = \begin{cases} \alpha\, sim_j(t_i, t_j) + \beta \dfrac{n_i}{n_j}, gdy\ n_i \le n_j \\ \alpha\, sim_j(t_i, t_j) + \beta \dfrac{n_j}{n_i}, gdy\ n_i > n_j \end{cases} \tag{4}$$

gdzie α oraz β są eksperymentalnie dobranymi współczynnikami, tak aby $\alpha+\beta=1$, $\alpha>0$, $\beta>0$. $sim(t_i, t_j)$ jest współczynnikiem podobieństwa pomiędzy transakcjami t_i oraz tj, $sim_j(t_i, t_j)$ jest współczynnikiem Jaccarda dla transakcji t_i oraz t_j. Ponadto wprowadza się współczynnik n_i/n_j (lub n_j/n_i, patrz równanie (4)), który obrazuje stopień podobieństwa wartości netto transakcji t_i oraz t_j. Dąży on do 0, gdy wartości netto są bardzo różne oraz do 1, gdy wartości te są identyczne. Czynnik $sim_j(t_i, t_j)$ można więc nazwać *podobieństwem produktów*, podczas gdy n_i/n_j (lub n_j/n_i) *podobieństwem wartości netto*. Wybierając różne wartości dla α oraz β można kontrolować, jak wielki jest wpływ obu podobieństw cząstkowych na wynikowe podobieństwo $sim(t_i, t_j)$. W tym miejscu należy również zwrócić uwagę, że ze względu na rzadkość macierzy koszyka sklepowego, czynnik $sim_j(t_i, t_j)$, zwykle będzie miał małe wartości i jego wkład w całkowity współczynnik podobieństwa $sim(t_i, t_j)$ nie będzie zbyt wielki.

PROSTY PRZYKŁAD

Aby zademonstrować zaproponowaną metodę grupowania, wykonano prosty eksperyment. Dokonano pogrupowania danych pokazanych w tabeli 1a oraz 1b. Tabela 1a pokazuje koszyk sklepowy w oryginalnej postaci, podczas gdy tabela 1b – ten sam ko-

Tabela 2. Macierze podobieństwa. (a) – bez uwzględniania drzewa taksonomii oraz wartości netto poszczególnych transakcji (α=1, β=0). Przykładowe obliczenia: *sim* (t_1, t_6)=1/ (1+1) =0,5, (b) – z uwzględnieniem drzewa taksonomii oraz ignorując wartości netto poszczególnych transakcji (α=1, β=0). Przykładowe obliczenia: *sim* (t_1, t_6) =1/ (1+1)=0,5, (c) – bez uwzględniania drzewa taksonomii, ale uwzględniając wartości netto poszczególnych transakcji. α oraz β są arbitralnie wybrane i wynoszą odpowiednio 0,7 oraz 0,3. Przykładowe obliczenia: *sim* (t_1, t_5) =0,7 * 0 + 0,3 * 0,5 = 0,15, (d) – z uwzględnieniem zarówno drzewa taksonomii, jak i wartości netto poszczególnych transakcji. α oraz β są arbitralnie wybrane i wynoszą odpowiednio 0,7 oraz 0,3. Przykładowe obliczenia: *sim* (t_1, t_5) =0,7 * 0,5 + 0,3 * 0,5 = 0,5

(a)

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	–	–	–	–	–	–
t_2	0	–	–	–	–	–
t_3	0	0	–	–	–	–
t_4	0	0	0	–	–	–
t_5	0	0	0	0	–	–
t_6	0,50	0	0	0	0	–

(b)

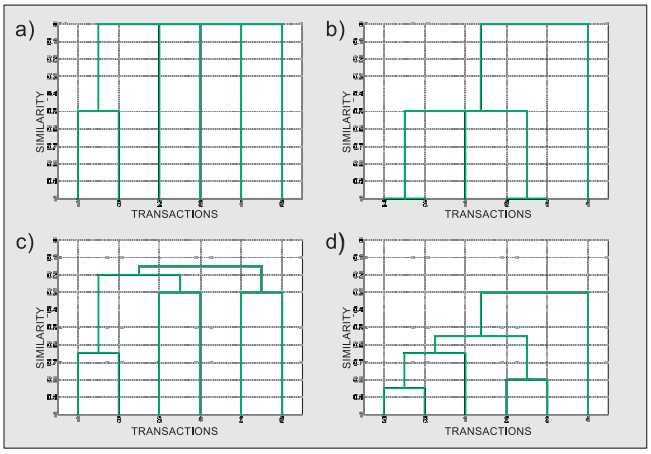
	t_1	t_2	t_3	t_4	t_5	t_6
t_1	–	–	–	–	–	–
t_2	0,50	–	–	–	–	–
t_3	0,50	1	–	–	–	–
t_4	0	0	0	–	–	–
t_5	0,50	0	0	0	–	–
t_6	0,50	0	0	0	1	–

(c)

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	–	–	–	–	–	–
t_2	0,15	–	–	–	–	–
t_3	0,20	0,10	–	–	–	–
t_4	0,20	0,10	0,30	–	–	–
t_5	0,15	0,30	0,10	0,10	–	–
t_6	0,65	0,15	0,20	0,20	0,15	–

(d)

	t_1	t_2	t_3	t_4	t_5	t_6
t_1	–	–	–	–	–	–
t_2	0,50	–	–	–	–	–
t_3	0,55	0,80	–	–	–	–
t_4	0,20	0,10	0,30	–	–	–
t_5	0,50	0,30	0,10	0,10	–	–
t_6	0,65	0,15	0,20	0,20	0,85	–



Rys. 2. Wyniki grupowania koszyka sklepowego z tabeli 1. Rysunki a-d odpowiadają danym z tabeli 2 (a-d)

szyk, jednak po jego zredukowaniu wynikającym z uwzględnienia drzewa taksonomii pokazanego na rys. 1. Przykładowo kolumna p_{456} oznacza, że oryginalne kolumny od 4 do 6 zostały połączone w jedną, gdyż należą do tego samego węzła taksonomii (grupa C na rys. 1). Po wyliczeniu współczynnika podobieństwa zgodnie ze wzorem (4) otrzymuje się różne macierze podobieństwa pokazane w tabeli 2 (dla różnych wartości współczynnika α oraz β). Gdy nie będą brane pod uwagę ani informacje o taksonomii, ani też o wartościach netto poszczególnych transakcji, otrzyma się grupowanie pokazane na rys. 2a[1)] (odpowiadającą mu macierz podobieństwa pokazano w tabeli 2a). Gdy uwzględni się dane o taksonomii, jednak pominie dane o wartościach netto, otrzyma się grupowanie pokazane na rys. 2b. Gdy nie będzie się brać pod uwagę informacji o taksonomii, ale uwzględni informacje o wartościach netto, otrzyma się grupowanie pokazane na rys. 2c. Gdy wreszcie weźmie się pod uwagę zarówno informacje o taksonomii, jak i o wartościach netto, otrzyma się grupowanie pokazane na rys. 2d (rysunkom 2b, 2c, 2d odpowiadają macierze podobieństwa pokazane odpowiednio w tabeli 2b, 2c, 2d).

Z analizy otrzymanych wyników widać, że najlepsze grupowanie (najlepiej rozróżnialne grupy) otrzymuje się uwzględniając w obliczeniach zarówno informacje o taksonomii, jak i o wartościach netto transakcji (tj. gdy $\alpha \neq 0$ oraz $\beta \neq 0$, rys. 2d). Z kolei najgorszy wynik otrzymuje się nie uwzględniając taksonomii ani wartości netto transakcji (tj. gdy $\alpha = 1$ oraz $\beta = 0$, rys. 2a). Z rys. 2a można odczytać, że transakcje 2, 3, 4 oraz 5 są do siebie całkowicie niepodobne (współczynnik podobieństwa = 0), podczas gdy w praktyce tak nie jest. Mówiąc o praktycznym podobieństwie, ma się na myśli podobieństwo w tym sensie, o którym była mowa w pierwszym rozdziale, gdzie celem nadrzędnym jest identyfikacja podobnych transakcji np. do celów prowadzenia skuteczniejszej akcji marketingowej.

[1)] Wyniki grupowania przedstawiono w postaci dendrogramu. Na osi *X* pokazano numery poszczególnych transakcji, a na osi *Y* wartość współczynnika podobieństwa wyliczonego według wzoru (4).

UWAGI KOŃCOWE

Na koniec należy zauważyć, że zaproponowaną metodę zademonstrowano na niezwykle prostym przykładzie, podczas gdy rzeczywiste koszyki sklepowe mają ogromną wymiarowość i są bardzo rzadkie. Oczywiście może się okazać, że wyniki grupowania mogą nie być za każdym razem takie „wzorcowe", jak w pokazanym przykładzie. Grupowanie jako takie nie jest bowiem procesem automatycznym i dającym zawsze optymalne rozwiązanie (bardzo często po prostu nie jest znane rozwiązanie optymalne w matematycznym rozumieniu tego słowa). Ponadto w przykładzie pokazano tylko zastosowanie współczynnika podobieństwa Jaccarda, podczas gdy można by wykonać również badania dla współczynnika niepodobieństwa (patrz wzory (2) oraz (3)). Ponadto jako miarę podobieństwa tworzonych grup użyto miary minimalnej. Można by również wykonać eksperymenty dla innych popularnych miar.

* * *

W artykule zaproponowano nową metodę grupowania danych z tzw. koszyka sklepowego. Użyto klasycznej metody grupowania hierarchicznego aglomeracyjnego. Jednak – inaczej niż w znanych pracach innych autorów – wykorzystano dodatkowo informacje o taksonomii produktów oraz wartości netto poszczególnych transakcji. Zaproponowano powiązanie tych informacji, służące do wyliczania podobieństwa poszczególnych transakcji. Na prostym przykładzie pokazano zaletę opisanej metody. Wymagane są jednak dalsze prace, które pokażą, czy dla rzeczywistych danych wyniki będą równie obiecujące.

LITERATURA

[1] Bramer M.: *Principles of Data Mining, Springer-Verlag*, 2007, ISBN: 978-1-84628-765-7
[2] Gan G., Ma Ch., Wu J.: *Data Clustering. Theory, Algorithms and Applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007, ISBN: 978-0-898716-23-8
[3] Han J., Kamber, M.: *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann Publishers, 2006 (prezentacje Power Point do książki dostępne pod adresem: http: //www-faculty. cs. uiuc. edu/~hanj/bk2/slidesindex. html)
[4] Hand D., Mannila H., Smyth P.: *Eksploracja danych*, WNT Warszawa, 2005, ISBN: 83-204-3053-4
[5] Larose D. T.: *Odkrywanie wiedzy z danych*. *Wprowadzenie do eksploracji danych*, PWN Warszawa, 2006, ISBN: 83-01-14836-5
[6] Witten I. H., Frank E.: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005, ISBN 0-12-088407-0

Artykuł recenzowany