PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Michał Kukliński

Uniwersytet Mikołaja Kopernika w Toruniu

STUDIUM PORÓWNAWCZE WYBRANYCH ALGORYTMÓW ANALIZY KOSZYKOWEJ

Streszczenie: Artykuł zawiera porównanie metod analizy koszykowej na przykładzie transakcyjnej bazy danych. W publikacji przedstawione zostały poszczególne etapy przygotowania danych oraz analizy za pomocą oprogramowania Statistica i SPSS Clementine. Zestawienie podstawowych charakterystyk metod *a priori* oraz GRI pozwala na wybór odpowiedniego algorytmu w zależności od typu danych oraz ilości danych wejściowych.

Słowa kluczowe: analiza koszykowa, reguły asocjacyjne, GRI, *a priori.*

1. Wstęp

Celem artykułu jest ocena efektywności metod badania zjawiska asocjacji w analizie koszykowej. W publikacji zaprezentowano etapy badania asocjacji za pomocą różnych technik − algorytmów, z użyciem oprogramowania SPSS Clementine oraz Statistica. Odkrywanie reguł asocjacyjnych, w zależności od sposobu przedstawienia danych, wymaga specjalnych zabiegów mających na celu przygotowanie posiadanych informacji do obliczeń oraz wybranie odpowiednio efektywnej i skutecznej metody. Badanie zastosowania metod GRI i *a priori* oraz dokładności otrzymanych wyników w możliwie najkrótszym czasie pozwoli na uzyskanie informacji niezbędnych do analiz marketingowych. Umożliwi sformułowanie wskazówek, którą metodę należy zastosować w konkretnych badaniach asocjacji w celu stworzenia skutecznych strategii marketingowych [Larose 2008].

2. Procedura przygotowania danych empirycznych

Obliczenia reguł asocjacyjnych zostały przeprowadzone na danych empirycznych pochodzących ze sklepu z obuwiem, obejmujących dwa pełne lata sprzedaży. Baza danych transakcyjnych zawiera 32 245 dokumentów sprzedaży, a w swoim asortymencie posiada 37 grup towarowych. Pierwotna hurtownia danych zawierała

78 564 rekordów z informacjami dotyczącymi sprzedaży, tj. numer dokumentu sprzedaży, datę i godzinę sprzedaży, kod produktu (zawierający grupę asortymentową), cenę jednostkową, wysokość udzielonego rabatu oraz stawkę podatku VAT. Proces obliczeń poprzedzony musi być etapami wstępnej obróbki danych, na które składają się: czyszczenie danych, obsługa brakujących danych, identyfikacja błędnych klasyfikacji, graficzne metody identyfikacji punktów oddalonych, numeryczne metody identyfikacji punktów oddalonych oraz przekształcenia danych.

ID Transakcji	Aktualizacja	ID Sprzedaży	ID Grupy	ID Rozmiar	ID Produkt	Nazwa produktu	Ilość	Wartość	Stawka VAT	Obniżka
11	02.01.2009 11:34:47	1	6	410	8966802	MENS WINTER	1.000000	159.000000	22.000000	0.000000
11	02.01.2009 11:34:47	2	25	320	3924305	KIDS WINTER	1.000000	79.000000	7.000000	0.000000
11	02.01.2009 11:34:47	3	66	35	9030123	GLOVES	1.000000	9.000000	22.000000	0.000000
11	02.01.2009 11:34:47	4	62	320	9022361	SOCKS CHILDS	1.000000	19.000000	22.000000	0.000000
12	02.01.2009 11:36:39	1	63	50	9954850	STOCKINGS	1.000000	3.000000	22.000000	0.000000
12	02.01.2009 11:36:39	2	63	80	9954830	STOCKINGS	1.000000	3.000000	22.000000	0.000000
12	02.01.2009 11:36:39	3	63	30	9954850	STOCKINGS	1.000000	3.000000	22.000000	0.000000
13	02.01.2009 11:40:10	1	2	430	8364313	MENS DRESS	1.000000	159.000000	22.000000	0.000000
13	02.01.2009 11:40:10	2	69	10	9014904	ACCESSORIES	1.000000	2.000000	22.000000	0.000000
14	02.01.2009 11:42:38	1	15	400	7164330	LADIES CITY HEELS	1.000000	159.000000	22.000000	0.000000
14	02.01.2009 11:42:38	2	69	10	9022104	ACCESSORIES	1.000000	9.000000	22.000000	0.000000
16	02.01.2009 11:50:00	1	2	390	8119039	MENS DRESS	1.000000	59.000000	22.000000	0.000000
17	02.01.2009 12:04:50	1	6	440	8567336	MENS WINTER	1.000000	199.000000	22.000000	0.000000
17	02.01.2009 12:04:50	2	62	210	9022108	SOCKS CHILDS	1.000000	9.000000	22.000000	0.000000
17	02.01.2009 12:04:50	3	62	210	9022802	SOCKS CHILDS	1.000000	13.000000	22.000000	0.000000
18	02.01.2009 12:06:34	1	2	420	8263304	MENS DRESS	1.000000	99.000000	22.000000	0.000000
18	02.01.2009 12:06:34	2	69	10	9022131	ACCESSORIES	1.000000	9.000000	22.000000	0.000000
19	02.01.2009 12:43:44	1	11	390	7662312	LADIES CITY HEELS	1.000000	99.000000	22.000000	0.000000
19	02.01.2009 12:43:44	2	60	410	9022108	SOCKS MENS	1.000000	9.000000	22.000000	0.000000
19	02.01.2009 12:43:44	3	69	10	9026023	ACCESSORIES	1.000000	13.000000	22.000000	0.000000
20	02.01.2009 12:46:57	1	13	390	5946307	LADIES FREE TIME	1.000000	359.000000	22.000000	0.000000
20	02.01.2009 12:46:57	2	69	10	9022104	ACCESSORIES	1.000000	9.000000	22.000000	0.000000
20	02.01.2009 12:46:57	3	61	390	9936709	SOCKS LADIES	1.000000	7.000000	22.000000	0.000000
20	02.01.2009 12:46:57	4	61	390	9936709	SOCKS LADIES	1.000000	7.000000	22.000000	0.000000
21	02.01.2009 12:50:21	1	12	360	6622334	LADIES CITY CASUAL	1.000000	99.000000	22.000000	0.000000
22	02.01.2009 12:51:40	1	66	10	9010582	CLOTH LADIES	1.000000	1.000000	22.000000	0.000000
22	02.01.2009 12:51:40	2	69	10	9014803	ACCESSORIES	1.000000	4.000000	22.000000	0.000000
23	02.01.2009 12:54:50	1	12	380	6622334	LADIES CITY CASUAL	1.000000	99.000000	22.000000	0.000000
23	02.01.2009 12:54:50	2	63	70	9954850	STOCKINGS	1.000000	3.000000	22.000000	0.000000
23	02.01.2009 12:54:50	3	63	70	9954850	STOCKINGS	1.000000	3.000000	22.000000	0.000000
24	02.01.2009 13:03:43	1	60	390	9904750	SOCKS MENS	1.000000	7.000000	22.000000	0.000000

Rys. 1. Baza transakcyjna (wersja pierwotna)

Źródło: baza transakcyjna sklepu z obuwiem za lata 2008 i 2009.

Na potrzeby badania reguł asocjacyjnych pominięto część informacji, które opisywały sprzedaż. Zostało stworzone dodatkowe pole zawierające kolejny numer

transakcji, tak aby numery z poszczególnych lat, nadawane co roku od początku, nie pokrywały się. W wyniku przekształceń do właściwych obliczeń reguł asocjacyjnych zostały użyte 2 pola: numer transakcji oraz towar. W związku z tym, że występowanie kilku produktów zawierających się w jednej grupie asortymentowej w ramach jednego dokumentu sprzedaży powodowało powtórzenia, należało dokonać grupowania. Grupowanie miało na celu doprowadzenie danych do postaci, w której dana grupa asortymentowa występowała na każdym dokumencie sprzedaży najwyżej jeden raz.

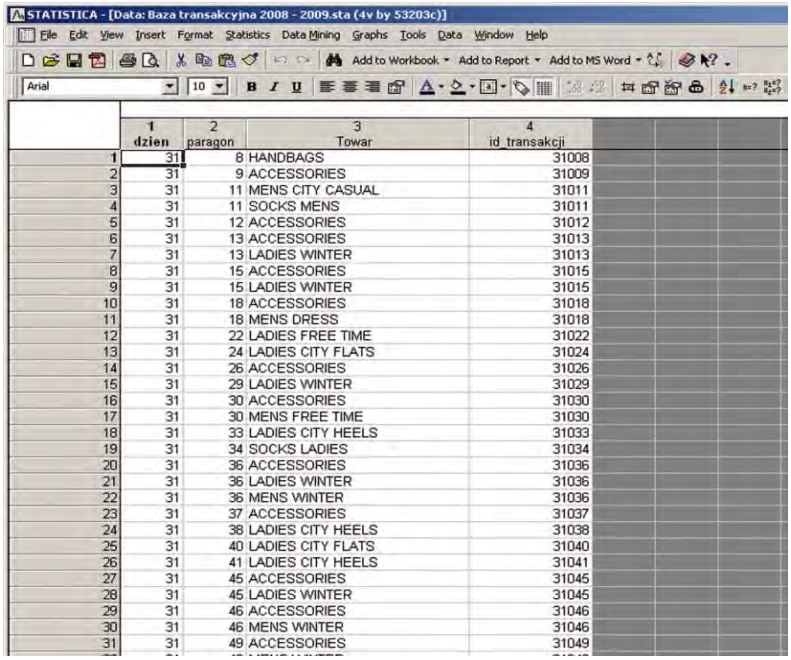
Następnie dokonano obliczeń metodą *a priori* i GRI, poprzedzając obliczenia dodatkowymi zabiegami, mającymi na celu dopasowanie danych do wymagań poszczególnych metod.

3. Metoda *a priori*

Zastosowanie metody *a priori* wymaga wprowadzenia do posiadanej bazy transakcyjnej dodatkowego pola o nazwie *id_transakcji*, aby nadać transakcjom z dwóch lat unikatowe indeksy. Jeżeli pozostawilibyśmy tylko pola *dzień* i *paragon*, to zarówno w roku 2008, jak i w 2009 istnieje rekord zawierający na przykład dzień o numerze 45 oraz paragon o numerze 1. W pierwszym kroku dni z dwóch lat zostały ponumerowane narastająco, następnie powstało nowe pole łączące *dzień* i *paragon* według następującego algorytmu:

$$\text{id\_transakcji} = \text{dzień} \cdot 1000 + \text{paragon}.$$

Algorytm *a priori* odkrywa zestaw reguł z danych, wybierając reguły, które mają najwyższą zawartość informacji. Metoda *a priori* oferuje 5 różnych metod selekcji reguł i wykorzystuje model indeksowania do wydajnego przetwarzania dużych baz danych. Metoda *a priori* wymaga, aby wszystkie dane wejściowe i wyjściowe były jakościowe, ale oferuje lepszą wydajność, ponieważ algorytm został zaimplementowany do tego rodzaju danych [Rauch 2005]. Algorytm 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 towaru do zbioru niepustego nie spowoduje, że zbiór ten stanie się pusty. Kolejnym wnioskiem jest, 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. Mając już wszystkie zbiory częste (k) algorytm wyszuka wszystkie podzbiory (l) znalezionych zbiorów częstych. Następnie zbada występowanie reguły, jeżeli l, to $(k-1)$. Dla podanych reguł algorytm wylicza poziom wsparcia i ufności. Od zadanego minimalnego poziomu wsparcia i ufności zależy ilość reguł asocjacyjnych, jakie zaproponuje metoda, zakładając, że dane wejściowe są identyczne [Larose 2006].



	1 dzien	2 paragon	3 Towar	4 id_transakcji
1	31	8	HANDBAGS	31008
2	31	9	ACCESSORIES	31009
3	31	11	MENS CITY CASUAL	31011
4	31	11	SOCKS MENS	31011
5	31	12	ACCESSORIES	31012
6	31	13	ACCESSORIES	31013
7	31	13	LADIES WINTER	31013
8	31	15	ACCESSORIES	31015
9	31	15	LADIES WINTER	31015
10	31	18	ACCESSORIES	31018
11	31	18	MENS DRESS	31018
12	31	22	LADIES FREE TIME	31022
13	31	24	LADIES CITY FLATS	31024
14	31	26	ACCESSORIES	31026
15	31	29	LADIES WINTER	31029
16	31	30	ACCESSORIES	31030
17	31	30	MENS FREE TIME	31030
18	31	33	LADIES CITY HEELS	31033
19	31	34	SOCKS LADIES	31034
20	31	36	ACCESSORIES	31036
21	31	36	LADIES WINTER	31036
22	31	36	MENS WINTER	31036
23	31	37	ACCESSORIES	31037
24	31	38	LADIES CITY HEELS	31038
25	31	40	LADIES CITY FLATS	31040
26	31	41	LADIES CITY HEELS	31041
27	31	45	ACCESSORIES	31045
28	31	45	LADIES WINTER	31045
29	31	46	ACCESSORIES	31046
30	31	46	MENS WINTER	31046
31	31	49	ACCESSORIES	31049

Rys. 2. Baza danych przygotowana do obliczeń metodą *a priori*

Źródło: obliczenia własne.

Tabela 1. Reguły asocjacyjne – metoda *a priori*

Podsumowanie reguł asocjacyjnych (Baza transakcyjna 2008 - 2009.sta)
Min. Wsparcie = 3,0%, Poziom ufności = 5,0%
Max. rozmiar zestawu = 5

	Poprzednik	==>	Następnik	Wsparcie (%)	Poziom ufności (%)	Wdrażalność
1	LADIES CITY HEELS	==>	ACCESSORIES	6,263003	49,82708	1,102495
2	ACCESSORIES	==>	LADIES CITY HEELS	6,263003	13,85778	1,102495
3	STOCKINGS	==>	ACCESSORIES	4,775656	37,77014	0,835718
4	ACCESSORIES	==>	STOCKINGS	4,775656	10,56682	0,835718
5	MENS DRESS	==>	ACCESSORIES	4,368887	59,34205	1,313027
6	ACCESSORIES	==>	MENS DRESS	4,368887	9,66678	1,313027
7	SOCKS MENS	==>	ACCESSORIES	4,325415	45,13934	0,998772
8	ACCESSORIES	==>	SOCKS MENS	4,325415	9,57059	0,998772
9	LADIES CITY CASUAL	==>	ACCESSORIES	4,191896	46,64824	1,032158
10	ACCESSORIES	==>	LADIES CITY CASUAL	4,191896	9,27516	1,032158
11	LADIES SUMMER	==>	ACCESSORIES	3,778916	38,65947	0,855395
12	ACCESSORIES	==>	LADIES SUMMER	3,778916	8,36139	0,855395
13	LADIES WINTER	==>	ACCESSORIES	3,428039	56,81935	1,257209
14	ACCESSORIES	==>	LADIES WINTER	3,428039	7,58502	1,257209
15	MENS CITY CASUAL	==>	ACCESSORIES	3,238628	49,92820	1,104732
16	ACCESSORIES	==>	MENS CITY CASUAL	3,238628	7,16592	1,104732

Źródło: obliczenia własne za pomocą Statistica 8.0.

Zawarte w tab. 1 pojęcia oznaczają:
Poprzednik − w przypadku zależności: jeżeli *A*, to *B*, jest to szukane *A*.

Następnik − w przypadku zależności: jeżeli *A*, to *B*, jest to szukane *B*.

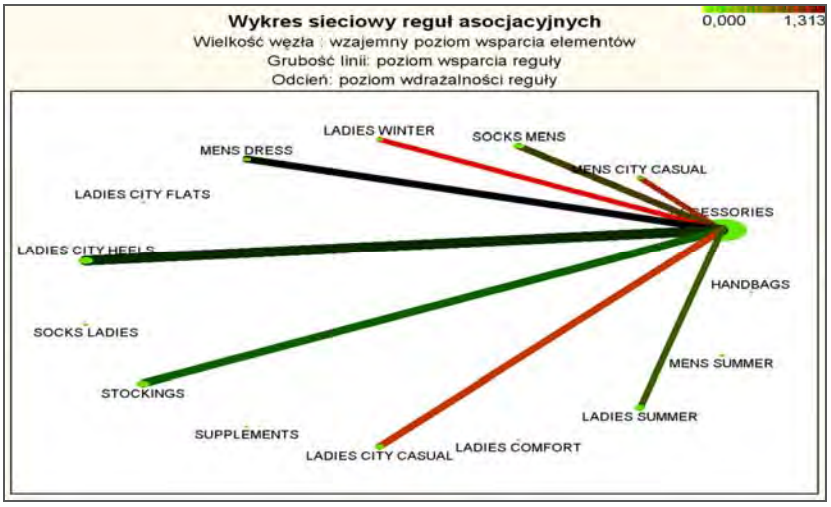
Wsparcie pokazuje udział rekordów zarówno z odpowiednim poprzednikiem, jak i następnikiem w stosunku do ogólnej liczby rekordów. Jest to stosunek rekordów z A i B do wszystkich rekordów wyrażony w procentach.

Poziom ufności wskazuje, w postaci procentowej, proporcje liczby rekordów zarówno z odpowiednim poprzednikiem, jak i jego następnikiem do liczby rekordów z jedynie odpowiednim poprzednikiem. Oznacza to, że poziom ufności jest to stosunek rekordów z A i B do wszystkich rekordów z A.

Wdrażalność jest procentową miarą wystąpienia danych spełniających warunki poprzednika, ale niespełniających warunków następnika. W odniesieniu do zakupu produktów oznacza to, jaki odsetek klientów posiada (lub zakupiło) poprzednik, ale nie zakupiło jeszcze następnika. Statystyka wyrażalności jest określona jako:

(*liczba rekordów spełniająca warunek poprzednika – liczba rekordów spełniająca regułę*)/*liczba rekordów*,

gdzie: *spełniająca warunek poprzednika* oznacza liczbę rekordów, dla których poprzednik jest prawdziwy, liczba rekordów *spełniająca regułę* oznacza liczbę rekordów, dla których zarówno poprzednik, jak i następnik są prawdziwe. Innymi słowy wdrażalność jest to iloraz różnicy liczby *A* i liczby rekordów z zarówno *A*, jak i *B* oraz sumy wszystkich rekordów.



Rys. 3. Wykres sieciowy reguł asocjacyjnych – metoda *a priori*

Źródło: obliczenia własne za pomocą Statistica 8.0.

Otrzymane rezultaty poszukiwania reguł asocjacyjnych możemy w oprogramowaniu Statistica 8.0 przedstawić za pomocą wykresów reguł oraz wykresów sie-

ciowych reguł asocjacyjnych. Zarówno z wykresu reguł, jak i z wykresu sieciowego możemy odczytać wszystkie najistotniejsze wskaźniki, jednakże asocjacje są przedstawione w sposób bardziej obrazowy na wykresie sieciowym. Z wykresu sieciowego możemy z łatwością odczytać zarówno towary wzajemnie powiązane, jak i stwierdzić poziom wsparcia reguł.

4. Metoda uogólnionej indukcji reguł − GRI

W celu zastosowania metody GRI na danych transakcyjnych wymagane jest przekształcenie bazy danych do formatu macierzowego. Metodę przekształcenia transakcyjnej bazy danych w macierzowy zapis transakcji udostępnia pakiet oprogramowania SPSS Clementine. W wyniku przekształcenia baza danych przyjmuje format macierzowy w postaci: jeden rekord przedstawia zapis jednej transakcji sprzedaży oraz zawiera pole identyfikujące transakcję (*id_transakcji*) oraz kolejne pola odpowiadające nazwom poszczególnych produktów. W naszym przypadku każdy rekord zawiera $p + 1$ pól, gdzie p oznacza ilość unikalnych produktów w bazie danych. Jeżeli dany produkt wystąpił w koszyku, pole identyfikujące przyjmuje wartość *PRAWDA* (*T = TRUE*), natomiast jeżeli produkt nie został kupiony, wartość pozostaje *FAŁSZ* (*F = FALSE*).

Table (37 fields, 32 217 records)

File Edit Generate

	ID TRANSAKCJI	Nazwa_ACCESSORIES	Nazwa_BRIEFCASES	Nazwa_CHILDS SUMMER	Nazwa_LADIES CITY FLATS	Nazwa_LADIES SUMMER	Nazwa_LADIES WINTER	Nazwa_GLOVES
1	31008.000	F	F	F	F	F	F	F
2	31009.000	T	F	F	F	F	F	F
3	31011.000	F	F	F	F	F	F	F
4	31012.000	T	F	F	F	F	F	F
5	31013.000	T	F	F	F	F	T	F
6	31015.000	T	F	F	F	F	T	F
7	31018.000	T	F	F	F	F	F	F
8	31022.000	F	F	F	F	F	F	F
9	31024.000	F	F	T	F	F	F	F
10	31026.000	T	F	F	F	F	F	F
11	31029.000	F	F	F	F	F	T	F
12	31030.000	T	F	F	F	F	F	F
13	31033.000	F	F	F	F	F	F	F
14	31034.000	F	F	F	F	F	F	F
15	31036.000	T	F	F	F	F	T	F
16	31037.000	T	F	F	F	F	F	F
17	31038.000	F	F	F	F	F	F	F
18	31040.000	F	F	T	F	F	F	F
19	31041.000	F	F	F	F	F	F	F
20	31045.000	T	F	F	F	F	T	F
21	31046.000	T	F	F	F	F	F	F
22	31049.000	T	F	F	F	F	F	F
23	31051.000	F	F	T	F	F	F	F
24	31052.000	T	F	F	F	F	T	F
25	31053.000	T	F	F	F	F	F	F
26	31054.000	T	F	F	F	F	F	F
27	31055.000	T	F	F	F	F	F	F
28	31056.000	T	F	F	F	F	F	F
29	31057.000	F	F	F	F	F	F	F
30	31065.000	T	F	F	F	F	F	F
31	31066.000	F	F	F	F	F	F	T
32	31067.000	T	F	F	F	F	F	F

Rys. 4. Baza danych przygotowana do obliczeń metodą GRI

Źródło: obliczenia własne za pomocą SPSS Clementine.

Metoda GRI wskazuje reguły o najwyższej zawartości informacji oparte na indeksie biorącym pod uwagę zarówno uogólnienie, jak i dokładność. GRI może wykorzystywać dane ilościowe oraz jakościowe, ale cel musi być jakościowy. Aby określić, czy kandydująca reguła jest interesująca, „GRI wykorzystuje *J-miarę*.

$$J = p(A)\left[p(A \mid B) \ln \frac{p(B \mid A)}{p(B)} + \left[1 - p(B \mid A)\right] \ln \frac{1 - p(B \mid A)}{1 - p(B)} \right],$$

gdzie: $p(A)$ – reprezentuje prawdopodobieństwo lub ufność obserwowanej wartości *A*, jest to miara zakresu poprzednika;

$p(B)$ – reprezentuje prawdopodobieństwo lub ufność obserwowanej wartości *B*, jest to miara zakresu następnika;

$p(B \mid A)$ – reprezentuje prawdopodobieństwo warunkowe lub późniejszą ufność zmiennej *B* dla danego *A*, które następuje; jest to miara prawdopodobieństwa zaobserwowania wartości *B* pod warunkiem, że występuje *A*.

Zatem $p(B \mid A)$ reprezentuje uaktualnione prawdopodobieństwo obserwowania wartości B po zyskaniu dodatkowej wiedzy o wartości *A*. W terminologii reguł asocjacyjnych $p(B \mid A)$ jest mierzone bezpośrednio jako ufność reguły. [...] Dla reguł z więcej niż jednym poprzednikiem $p(A)$ jest obliczane jako prawdopodobieństwo koniunkcji wartości zmiennych w poprzedniku" [Larose 2006].

GRI

File Generate

Sort by Wsparcie (%) 16 of 16

Poprzednik	Następnik	Poziom ufności (%)	Wsparcie (%)	Wdrażalność
Nazwa_ACCESSORIES	Nazwa_MENS DRESS	59,34	4,367	1,313
Nazwa_MENS DRESS	Nazwa_SOCKS MENS	17,66	1,692	2,4
Nazwa_SOCKS MENS	Nazwa_MENS DRESS	22,99	1,692	2,4
Nazwa_MENS CITY CASUAL	Nazwa_SOCKS MENS	15,1	1,446	2,329
Nazwa_SOCKS MENS	Nazwa_MENS CITY CASUAL	22,31	1,446	2,329
Nazwa_SOCKS MENS	Nazwa_ACCESSORIES Nazwa_MENS DRESS	28,5	1,245	2,975
Nazwa_MENS DRESS	Nazwa_ACCESSORIES Nazwa_SOCKS MENS	28,79	1,245	3,912
Nazwa_LADIES CITY HEELS	Nazwa_ACCESSORIES Nazwa_STOCKINGS	25,88	1,235	2,06
Nazwa_MENS SUMMER	Nazwa_SOCKS MENS	12,31	1,18	2,513
Nazwa_SOCKS LADIES	Nazwa_SOCKS MENS	11,24	1,077	2,362
Nazwa_STOCKINGS	Nazwa_LADIES COMFORT	26,3	0,959	2,081
Nazwa_MENS CITY CASUAL	Nazwa_ACCESSORIES Nazwa_SOCKS MENS	21,46	0,928	3,31
Nazwa_SOCKS MENS	Nazwa_ACCESSORIES Nazwa_MENS CITY CASUAL	28,67	0,928	2,993
Nazwa_ACCESSORIES	Nazwa_SUPPLEMENTS	24,34	0,826	0,539
Nazwa_MENS SUMMER	Nazwa_ACCESSORIES Nazwa_SOCKS MENS	14,72	0,636	3,005
Nazwa_LADIES COMFORT	Nazwa_ACCESSORIES Nazwa_STOCKINGS	12,68	0,605	3,477

Rys. 5. Baza danych przygotowana do obliczeń metodą GRI

Źródło: obliczenia własne za pomocą SPSS Clementine.

W przypadku tego algorytmu badacz określa maksymalną liczbę reguł, które chce osiągnąć, ponieważ znajdowanie kolejnych reguł polega na obliczaniu *J-miary* dla kolejnych przypadków i porównywaniu wartości z najniższą dostępną wartością z tabeli. Jeżeli nowa wartość jest większa, to zostaje nadpisana na poprzednią.

5. Podsumowanie

Negatywnym zjawiskiem występującym podczas dokonywania obliczeń jest problem dużego zestawu danych oraz ilości teoretycznie możliwych reguł asocjacyjnych [Taniar 2008]. Przykładowo, jeżeli posiadamy *k*-atrybutowy zbiór, to liczba reguł asocjacyjnych jest rzędu $k \cdot 2^{k-1}$, zatem jeżeli dysponujemy zbiorem danych transakcyjnych małego sklepu, który ma w swojej ofercie tylko 100 artykułów, to liczbę reguł asocjacyjnych szacujemy na $100 \cdot 2^{99} \cong 6,4 \cdot 10^{31}$. Należy zwrócić uwagę, że trudno znaleźć sklep, który oferuje tylko 100 produktów, obecnie w hipermarketach mamy tysiące unikalnych produktów. O ile w metodzie *a priori* dane nie wymagają specjalnie skomplikowanych przekształceń, to w przypadku metody GRI, gdzie bazę należy przekształcić na format macierzowy, powstaje problem dużego zbioru danych. Własne doświadczenia badawcze nad inną bazą transakcyjną, zawierającą 1245 unikalnych produktów, metoda GRI nie podołała obliczeniom na tak dużej macierzy.

Na przedstawionym przykładzie widać, że dokładniejsza okazała się metoda *a priori*, która wskazała znacznie więcej reguł, szczególnie tych z wysokimi wartościami współczynnika wsparcia. Uwagę należy zwrócić również na szybkość obliczeń, które w metodzie GRI trwały siedem razy dłużej niż w metodzie *a priori*. W przypadku dużych problemów badawczych algorytm *a priori* jest szybszy niż GRI i ponadto nie ma on odgórnego limitu liczby reguł, które można uzyskać, a także może obsługiwać reguły aż do 32 założeń.

Metoda *GRI* może wykorzystywać dane ilościowe oraz jakościowe, ale cel musi być jakościowy. W przypadku algorytmu *a priori* wszystkie dane muszą mieć charakter jakościowy, natomiast dane ilościowe można dyskretyzować lub zastosować wspomniany algorytm GRI [Agrawal i in. 1993].

Literatura

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

Larose D.T., *Metody i modele eksploracji danych*, Wyd. Naukowe PWN, Warszawa 2008.

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, 22, Springer Science 2005.

Taniar D., *Data Mining and Knowledge Discovery Technologies*, IGI Publishing, Hershey 2008.

COMPARATIVE STUDY OF SELECTED METHODS OF MARKET BASKET ANALYSIS

Summary: The article presents the comparison of methods of market basket analysis on the example of transactional database. The publication presents various stages of data preparation and analysis using the Statistica software and SPSS Clementine. The combination of basic characteristics of *a'priori* and *GRI* methods allows to choose the appropriate algorithm, depending on the type and amount of input data.