Nr 51

PRACE NAUKOWE

Uniwersytetu Ekonomicznego we Wrocławiu

Projektowanie, ocena i wykorzystanie danych rynkowych

Redaktor naukowy
Józef Dziechciarz

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 51 — 2009
Projektowanie, ocena i wykorzystanie danych rynkowych

Iwona Kasprzyk
Akademia Ekonomiczna w Katowicach

WYKORZYSTANIE KONFIGURACYJNEJ ANALIZY CZĘSTOŚCI W ANALIZIE KLAS UKRYTYCH

1. Wstęp

W statystyce jest niewiele metod pozwalających na analizę danych, które mierzone są na najsłabszych skalach pomiaru. Wśród nich można wyróżnić m.in. konfiguracyjną analizę częstości (*configural frequency analysis*, CFA) oraz analizę klas ukrytych (*latent class analysis*, LCA).

CFA została zaproponowana przez G.A. Lienerta w 1968 r. i pozwala na zidentyfikowanie tzw. typów oraz antytypów poprzez badanie poszczególnych komórek tablicy kontyngencji. W wyniku zastosowania CFA w tablicy kontyngencji wskazywane są te komórki zawierające liczebności obserwowane, które wyraźnie odchylają się od liczebności oczekiwanych.

Istotą analizy klas ukrytych jest badanie zależności między kategoriami zmiennych niemierzalnych. Umożliwia ona zidentyfikowanie wzajemnie wyłączających się klas ukrytych, które wyjaśniają rozkład liczebności zawartych w tablicy kontyngencji.

Celem artykułu jest określenie profilu użytkowników Internetu za pomocą dwóch metod: analizy klas ukrytych oraz konfiguracyjnej analizy częstości. Następnie wyniki uzyskane metodą CFA zostaną wykorzystane do przedstawienia poszczególnych konfiguracji na wykresie współrzędnych barycentrycznych.

2. Analiza klas ukrytych

Załóżmy, że mamy daną tablicę kontyngencji z trzema zmiennymi obserwowalnymi: A ($i = 1, 2, ..., I$), B ($j = 1, 2, ..., J$) oraz C ($k = 1, 2, ..., K$). Zmienna ukryta X przyjmuje wartości $w = 1, 2, ..., W$, gdzie W oznacza liczbę klas. Model z jedną zmienną ukrytą X można przedstawić za pomocą poniższego równania:

$$\pi_{ijk} = \sum_{w=1}^{W} \pi_{ijkw}^{ABCX}, \tag{1}$$

gdzie π_{ijkw}^{ABCX} oznacza prawdopodobieństwo warunkowe tego, że *i*-ta, *j*-ta oraz *k*-ta kategoria zmiennych *A*, *B* oraz *C* znajdzie się w opisie klasy ukrytej *w*.

Wykorzystanie wzoru (1) wymaga spełnienia założenia o lokalnej niezależności zmiennych:

$$\pi_{ijkw}^{ABCX} = \pi_w^X \pi_{iw}^{A\backslash X} \pi_{jw}^{B\backslash X} \pi_{kw}^{C\backslash X}, \tag{2}$$

gdzie π_w^X oznacza prawdopodobieństwo przynależności danych obserwacji do klasy *w* zmiennej ukrytej *X*, zaś $\pi_{iw}^{A\backslash X}$ prawdopodobieństwo warunkowe tego, że *i*-ta kategoria zmiennej *A* znajdzie się w opisie klasy ukrytej *w*.

Prawdopodobieństwa określone po prawej stronie równania (2) wymagają spełnienia poniższego założenia:

$$\sum_{w=1}^{W} \pi_w^X = \sum_{i=1}^{I} \pi_{iw}^{A\backslash X} = \sum_{j=1}^{J} \pi_{jw}^{B\backslash X} = \sum_{k=1}^{K} \pi_{kw}^{C\backslash X} = 1. \tag{3}$$

3. Wykres współrzędnych barycentrycznych

Van der Heijden, Gilula i van der Ark [1999] pokazali, że modele klas ukrytych można przedstawić za pomocą wykresu współrzędnych barycentrycznych (*barycentric coordinates*). Tego rodzaju współrzędne na wykresie tworzą trójkąt równoboczny, którego wierzchołki P_1, P_2, P_3 tworzą bazę przestrzeni. Każdemu z tych punktów przyporządkowana jest pewna waga (w tym przypadku zakłada się, że wszystkie wagi są równe 1). Dowolny punkt w przestrzeni trójwymiarowej można wyrazić jako sumę ważoną:

$$P = xP_1 + yP_2 + zP_3.$$

Współczynniki *x*, *y*, *z* są nazywane współrzędnymi barycentrycznymi, które przyjmują wartości z przedziału [0,1], a ich suma jest równa 1 ($x + y + z = 1$). Kategorię danej zmiennej można potraktować jako punkt *P* o współrzędnych *x*, *y*, *z*.

Poprzez odpowiednie przekształcenie poszczególnych prawdopodobieństw warunkowych $\pi_{iw}^{A\backslash X}$, $\pi_{jw}^{B\backslash X}$, $\pi_{kw}^{C\backslash X}$ model klas ukrytych można przedstawić za pomocą równania:

$$\frac{\pi_{ijk}}{\pi_{i\ldots}} = \sum_{w=1}^{W} \pi_{wi}^{X/A} \pi_{jw}^{B/X} \pi_{kw}^{C/X}, \tag{4}$$

gdzie: $\pi_{wi}^{X/A} = \dfrac{\pi_w^X \pi_{iw}^{A/X}}{\sum_{w=1}^{W} \pi_w^X \pi_{iw}^{A/X}} = \dfrac{\pi_w^X \pi_{iw}^{A/X}}{\pi_{i..}}.$ (5)

W ten sposób przekształcone prawdopodobieństwa warunkowe pozwalają na naniesienie punktów na wykresie współrzędnych barycentrycznych i odczytanie zależności między poszczególnymi kategoriami zmiennych obserwowalnych.

4. Konfiguracyjna analiza częstości

Kombinacje kategorii zmiennych zawartych w tablicy kontyngencji będziemy nazywać konfiguracjami. Niech f_{ijk} oznacza liczebności obserwowane, natomiast $\hat{f}_{ijk}$ liczebności oczekiwane dla danych konfiguracji.

Dla określenia typów oraz antytypów w konfiguracyjnej analizie częstości weryfikuje się hipotezę:

$$H_0 : f_{ijk} = \hat{f}_{ijk} ,$$
$$H_1 : f_{ijk} > \hat{f}_{ijk} \text{ lub } f_{ijk} < \hat{f}_{ijk} . \tag{6}$$

Jeśli $f_{ijk} > \hat{f}_{ijk}$, to takie konfiguracje nazywamy typami, natomiast gdy $f_{ijk} < \hat{f}_{ijk}$, to mówmy o tzw. antytypach. Gdy nie ma istotnej różnicy między f_{ijk} a $\hat{f}_{ijk}$, to dana konfiguracja nie może być określona ani typem, ani antytypem.

Do testowania hipotez możemy się posłużyć kilkoma rodzajami testów. Krauth i Lienert (zob. [Haberman 1973]) do identyfikacji typów oraz antytypów zastosowali test dwumianowy, tj.:

$$B(f_{ij}) = \binom{n}{f_{ijk}} p^{f_{ijk}} q^{(n-f_{ijk})} , \tag{7}$$

gdzie: p − prawdopodobieństwo wystąpienia danej konfiguracji, tj. $p = \hat{f}_{ijk} / n$, $q = 1-p$.

Ten test jest stosowany w przypadku małych prób.

Innym testem, którym można się posłużyć do badania rodzaju typów, jest test z, który można zapisać jako:

$$z = \frac{f_{ijk} - np}{\sqrt{npq}} . \tag{8}$$

Statystykę z można zastosować zamiast testu dwumianowego, jeśli $np \geq 10$. Krauth i Lienert wprowadzili poprawkę do testu z, która stosowana jest wówczas, gdy $5 \leq np \leq 10$. Test z określony wzorem (8) po uwzględnieniu korekty można zapisać w poniższy sposób:

$$z = \frac{f_{ijk} - np - 0,5}{\sqrt{npq}} . \tag{9}$$

Innym testem weryfikującym hipotezę o występowaniu typów albo antytypów jest test chi-kwadrat, który można zapisać za pomocą formuły:

$$\chi^2 = \frac{f_{ijk} - np}{\sqrt{np}}. \tag{10}$$

W konfiguracyjnej analizie częstości, przy weryfikacji hipotezy o występowaniu typu albo antytypu, poziom istotności α najczęściej ustalany jest zgodnie z modyfikacją Bonferroniego, tj.:

$$\alpha^* = \alpha/r, \tag{11}$$

gdzie: r − liczba konfiguracji.

Inne testy, które można zastosować do wyróżnienia typów oraz antytypów za pomocą CFA, zostały omówione w pracy [von Eye 1990].

5. Zastosowanie

Z Internetu korzysta coraz większa liczba osób. Ułatwia on komunikowanie, stanowi źródło informacji dostępnych w dowolnym momencie, z dowolnego miejsca. W związku z tym w lutym 2008 r. przeprowadzono badanie na temat korzystania z Internetu. Badanie zostało przeprowadzone pośród 147 użytkowników Internetu za pomocą kwestionariusza ankiety internetowej. Próba badawcza nie jest reprezentatywna dla badanej populacji i wyników badań nie można uogólnić na całą populację. Ostatecznie do dalszych badań przyjęto 139 ankiet. W niniejszym artykule wykorzystano odpowiedzi na 5 spośród 27 pytań, które zostały ujęte w ankiecie, tj.:

1. Gdzie Pan(i) korzysta najczęściej z Internetu?
2. Które konto poczty elektronicznej wykorzystuje Pan(i) najczęściej?
3. Czy rozmawia Pan(i) ze znajomymi przez komunikatory, np. gadu-gadu, tlen?
4. Czy czyta Pan(i) internetowe wersje gazet codziennych?
5. Czy słucha Pan(i) radia przez Internet?

W wyniku zastosowania analizy klas ukrytych okazało się, że modelem najlepiej dopasowanym do danych jest model z dwiema klasami ($L^2 = 31{,}2263$, $df = 56$, $p = 0{,}997$), co również pokazują kryteria informacyjne, takie jak: AIC oraz BIC (tab. 1):

Tabela 1. Kryteria informacyjne

Liczba klas (t)	df	AIC	BIC
1	64	830,6692	851,2105
2	56	793,9930	842,0101
3	48	807,0297	874,5226

Źródło: opracowanie własne z wykorzystanie programu **R**.

W klasie pierwszej znalazła się nieliczna grupa respondentów (17,27%). Osoby należące do niej wykorzystują Internet głównie w pracy (77,27%). Większość z nich posiada służbowe konto poczty elektronicznej zapewnione przez firmę (89,65%). Do celów komunikacyjnych w pracy wykorzystują główne komunikatory: gadu-gadu, tlen (69,46%). W przeciwieństwie do respondentów korzystających z Internetu w domu, ponad połowa badanych czyta internetowe wersje gazet codziennych. Ponad jedna trzecia (37,44%) respondentów słucha radia przez Internet.

Drugą największą grupę badanych stanowili respondenci, którzy najchętniej łączą się z Internetem w domu (82,73%). Większość badanych znajdujących się w tej klasie używa prywatnego konta poczty elektronicznej założonego na jednym z bezpłatnych serwerów. Zdecydowana większość użytkowników Internetu (93,96%) opisujących tę klasę komunikuje się ze znajomymi przez gadu-gadu lub inne tego typu komunikatory. Ponad połowa (52,57%) respondentów słucha radia przez Internet. 56,66% badanych osób nie czyta internetowych wersji gazet codziennych.

Szczegółowe wyniki przedstawiono w tab. 2.

Tabela 2. Wyniki segmentacji dla dwóch klas ukrytych

	Oznaczenia	Klasa I	Klasa II
		0,1727	**0,8273**
1. Gdzie Pan(i) korzysta najczęściej z Internetu?			
a) w domu	P11	0,1896	**0,9271**
b) w pracy	P12	**0,7727**	0,0641
c) na uczelni	P13	0,0377	0,0088
2. Które konto poczty elektronicznej wykorzystuje Pan(i) najczęściej?			
a) służbowe, zapewnione przez firmę	P21	**0,8965**	0,0701
b) prywatne, założone na jednym z serwerów bezpłatnych kont pocztowych	P22	0,0732	**0,9104**
c) prywatne, wykupione lub ewentualnie opłacone przeze mnie	P23	0,0304	0,0195
3. Czy rozmawia Pan(i) ze znajomymi przez komunikatory, np. gadu-gadu, tlen?			
a) tak	P31	**0,6946**	**0,9396**
b) nie	P32	0,3054	0,0604
4. Czy czyta Pan(i) internetowe wersje gazet codziennych?			
a) tak	P41	**0,5730**	0,4334
b) nie	P42	0,4270	**0,5666**
5. Czy słucha Pan(i) radia przez Internet?			
a) tak	P51	0,3744	**0,5257**
b) nie	P52	**0,6256**	0,4743

Źródło: opracowanie własne z wykorzystaniem programu **R**.

W celu wskazania konfiguracji, w których liczebności obserwowane odchylają się znacznie od liczebności oczekiwanych, wykorzystano konfiguracyjną analizę częstości. Analizę przeprowadzono z wykorzystaniem funkcji cfa dostępnej w pakiecie pod tą samą nazwą w programie R. W tym przypadku został zastosowany test z, w którym przyjęto poziom istotności $\alpha = 0{,}05$. Do weryfikacji hipotezy o występowaniu typu lub antytypu najczęściej poziom istotności ustalany był zgodnie z przyjętą modyfikacją Benofferoniego, tj. $\alpha^* = 0{,}05 / 72 = 0{,}000694444$, gdzie 72 oznacza liczbę wszystkich możliwych konfiguracji.

Tabela 3. Wyniki konfiguracyjnej analizy częstości badania użytkowników Internetu

Konfiguracje (s)	f_{ijklm}	f_{ijklm}	$\hat{f}_{ijklm}$	z	p	sig.z	
31211	1	1	0,0114	9,2772	0,0000	TRUE	antytyp
21222	3	3	0,189	6,4656	0,0000	TRUE	antytyp
21112	5	5	1,3334	3,1754	0,0007	FALSE	
23111	1	1	0,1232	2,4978	0,0062	FALSE	
21111	4	4	1,3143	2,3426	0,0096	FALSE	
21212	1	1	0,1613	2,0883	0,0184	FALSE	
21122	4	4	1,5625	1,9499	0,0256	FALSE	
13121	2	2	0,5621	1,9179	0,0276	FALSE	
21221	1	1	0,1863	1,8851	0,0297	FALSE	
22122	1	1	5,0782	1,8097	0,0352	FALSE	
12121	27	27	19,4864	1,7021	0,0444	FALSE	
22111	1	1	4,2715	1,5829	0,0567	FALSE	
11112	2	2	5,1906	1,4004	0,0807	FALSE	
11111	2	2	5,1164	1,3778	0,0841	FALSE	
11122	3	3	6,0827	1,2499	0,1057	FALSE	
11121	3	3	5,9958	1,2235	0,1106	FALSE	
22112	2	2	4,3334	1,1209	0,1312	FALSE	
32121	1	1	0,3575	1,0744	0,1413	FALSE	
12112	21	21	16,8694	1,0057	0,1573	FALSE	
22121	3	3	5,0057	0,8965	0,185	FALSE	
12221	1	1	2,3572	0,884	0,1883	FALSE	
12211	1	1	2,0115	0,7132	0,2379	FALSE	
12111	19	19	16,6284	0,5816	0,2804	FALSE	
12122	22	22	19,7689	0,5018	0,3079	FALSE	
22222	1	1	0,6143	0,4921	0,3113	FALSE	
11211	1	1	0,6189	0,4844	0,3141	FALSE	
11212	1	1	0,6279	0,4696	0,3193	FALSE	
21121	1	1	1,5402	0,4353	0,3317	FALSE	
12222	2	2	2,3914	0,2531	0,4001	FALSE	
12212	2	2	2,0407	0,0285	0,4886	FALSE	

Źródło: opracowanie własne z wykorzystaniem programu **R**.

W wyniku przeprowadzonej analizy wyróżniono dwa antytypy, tj. konfigurację 31211 oraz 21222 (zob. tab. 3). Konfiguracja 31211 sugeruje, że respondent nie korzysta, jak należałoby oczekiwać, tak często z Internetu na uczelni oraz nie wykorzystuje służbowego konta poczty elektronicznej zapewnionego przez firmę.

Podobna sytuacja ma miejsce przy interpretacji konfiguracji 21222, gdzie osoby zadeklarowały, że najczęściej korzystają z Internetu w pracy i wykorzystują prywatne konto poczty elektronicznej założone na jednym z bezpłatnych serwerów.

W tab. 3 przedstawiono tylko te konfiguracje, których liczebności obserwowane w tablicy kontyngencji wynoszą co najmniej 1.

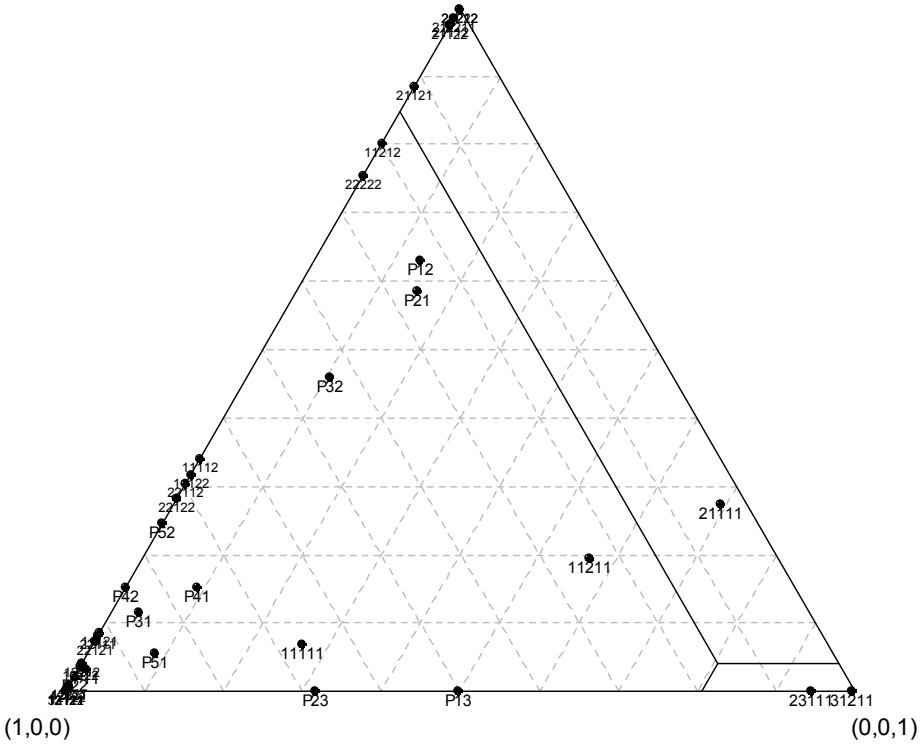
Tabela 4. Przekształcone wartości prawdopodobieństw warunkowych dla trzech klas

Konfiguracje (s)	f_{ijklm}	$p(k_1/s)$	$p(k_2/s)$	$p(k_3/s)$
31211	1	0,0000	0,9955	0,0045
21222	3	0,9966	0,0000	0,0034
21112	5	0,9766	0,0000	0,0234
23111	1	0,0000	0,9443	0,0557
21111	4	0,2744	0,6932	0,0324
21212	1	0,9969	0,0000	0,0031
21122	4	0,9740	0,0000	0,0260
13121	2	0,0000	0,0000	1,0000
21221	1	0,9834	0,0000	0,0166
22122	1	0,2818	0,0000	0,7182
12121	27	0,0010	0,0000	0,9990
22111	1	0,0815	0,0000	0,9185
11112	2	0,3409	0,0000	0,6591
11111	2	0,0699	0,2650	0,6651
11122	3	0,3172	0,0000	0,6828
11121	3	0,0862	0,0000	0,9138
22112	2	0,3040	0,0000	0,6960
32121	1	0,0000	0,0000	1,0000
12112	21	0,0008	0,0000	0,9992
22121	3	0,0738	0,0000	0,9262
12221	1	0,0077	0,0000	0,9923
12211	1	0,0085	0,0000	0,9915
12111	19	0,0011	0,0000	0,9989
12122	22	0,0048	0,0000	0,9952
22222	1	0,7536	0,0000	0,2464
11211	1	0,1959	0,5648	0,2393
11212	1	0,8012	0,0000	0,1988
21121	1	0,8837	0,0000	0,1163
12222	2	0,0366	0,0000	0,9634
12212	2	0,0406	0,0000	0,9594

Źródło: opracowanie własne.

Przedstawienie konfiguracji zawartych w tab. 3 na wykresie współrzędnych barycentrycznych wymaga przekształcenia, zgodnie ze wzorem (5), prawdopodobieństw warunkowych otrzymanych za pomocy analizy klas ukrytych[1]. Przekształcone wartości prawdopodobieństw warunkowych przedstawiono w tab. 4.

Z rys. 1 wynika, że konfiguracje 31222 oraz 23111 odbiegają od pozostałych. Pierwsza z tych konfiguracji to antytyp, który został omówiony wyżej. Konfiguracja 23111 – to respondent, który zadeklarował wykorzystanie Internetu w pracy. Osoba ta korzysta w firmie z płatnego konta wykupionego lub ewentualnie opłaconego przez nią. Będąc w pracy, korzysta z komunikatorów, słucha radia przez Internet oraz czyta codzienne gazety w wersji *on-line.*



Rys. 1. Wykres współrzędnych barycentrycznych[2] dla wyników LCA i CFA

Źródło: opracowanie własne z wykorzystaniem programu **R**.

[1] Przyjęto model z trzema klasami, mimo że model ten nie okazał się dobrze dopasowany do danych. Założenie takie występuje tylko dlatego, iż możliwe jest wówczas przedstawienie poszczególnych konfiguracji na wykresie współrzędnych barycentrycznych.

[2] Wykres współrzędnych barycentrycznych został wykonany za pomocą fukcji `triplot` dostępnej w pakiecie `klaR` w programie **R**.

Z powyżej przeprowadzonej analizy można wnioskować, że badane osoby są zróżnicowane w niewielkim stopniu pod względem miejsca korzystania z Internetu. Dom okazuje się być najbardziej preferowanym miejscem do surfowania po sieci.

Internet zyskuje również szybko na znaczeniu jako miejsce nawiązywania kontaktów. Większość badanych używa komunikatorów (np. gadu-gadu) do kontaktów ze znajomymi. Ponadto respondenci najczęściej korzystają z bezpłatnych kont poczty elektronicznej.

Zastosowana konfiguracyjna analiza częstości wskazała tylko dwie konfiguracje, których liczebności istotnie się odchylały od liczebności oczekiwanych i okazały się one antytypami, tj. $f_{ijk} < \hat{f}_{ijk}$.

Literatura

Haberman S.J., *The analysis of residuals in cross-classiefied tables*, „Bimetrics" 1973, no. 29, s. 205--220.

Heijden P.G.M. van der, Gilula Z., van der Ark L.A., *An extended study into the relationship between correspondence analysis and latent class analysis*, "Sociological Methodology" 1999, no. 29, s. 147-186.

Lazarsfeld, P.F., Henry N.W., *Latent structure analysis*, Houghton Mill, Boston 1968.

Lienert G.A., Krauth J., *Configural frequency analysis as a statistical tool for defining types*, „Educational and Psychological Measurement" 1975, no. 35, s. 231-238.

McCutcheon A.L., *Latent class analysis*, Quantitative Applications in the Social Sciences, Sage, Newbury Park, CA 1987.

Von Eye A., *Introduction to configural frequency analysis*, Cambridge University Press 1990.

APPLICATION OF CONFIGURAL FREQUENCY ANALYSIS IN LATENT CLASS ANALYSIS

Summary

The latent class analysis and configural frequency analysis are a multivariate analysis techniques of the contingency table which is based on discrete data.

The main aim of the configural frequency analysis is to analyze cells in a multidimensional contingency table. We use the CFA for identifying types and antytypes. The latent class model is a method for analyzing relationships in categorical data.

This article presents a result of research which was done among Internet users. Additionally, the results received from the CFA will be used to show configurations on the barycentric coordinates graph.