

Iwona Kasprzyk

GRAFICZNA PREZENTACJA MODELI KLAS UKRYTYCH

1. Wstęp

Analiza klas ukrytych (*Latent Class Analysis*) jest zbiorem metod, która mogą być wykorzystywane do analizy danych marketingowych uzyskanych za pomocą np. badania ankietowego. Pozwala ona na odkrycie pewnych nieznanych cech konsumentów (respondentów), których bezpośrednio nie można zaobserwować. Takimi cechami ukrytymi mogą być np.: nastawienie do danego produktu (marki), osobowość itp.

Istotą modelowania klas ukrytych jest badanie związków między kategoriami zmiennych nominalnych i porządkowych. Metoda ta została wprowadzona przez Lazarsfelda [5] w latach pięćdziesiątych. Wykorzystuje ona dane zawarte w tablicy kontyngencji. W ramach tej analizy wyróżnia się dwa różne podejścia pozwalające na budowę modelu klas ukrytych: podejście probabilistyczne oraz podejście oparte na modelowaniu log-liniowym.

Celem artykułu jest omówienie możliwości graficznej prezentacji wyników analizy za pomocą tzw. wykresu współrzędnych barycentrycznych.

2. Dwa podejścia do modelowania klas ukrytych

Analiza klas ukrytych może zostać przeprowadzona na dwa sposoby. Pierwszy z nich – podejście probabilistyczne – został wprowadzony przez Goodmana [1], drugi zaś opiera się na modelowaniu log-liniowym, zaproponowanym przez Habermana [2], a następnie rozwijanym przez Hagenaarsa [3] oraz Vermunta [8; 9].

2.1. Podejście probabilistyczne

Załóżmy, że mamy daną tablicę kontyngencji N z czterema zmiennymi Z ($i = 1, 2, \dots, I$), S ($j = 1, 2, \dots, J$), D ($k = 1, 2, \dots, K$) oraz P ($l = 1, 2, \dots, L$). Zmienna ukryta X przyjmuje wartości $t = 1, 2, \dots, T$, gdzie T oznacza liczbę klas.

Model z jedną zmienną ukrytą X można przedstawić za pomocą poniższego równania:

$$\pi_{ijkl} = \sum_{t=1}^T \pi_{ijklt}^{ZSDPX}, \quad (1)$$

gdzie: π_{ijklt}^{ZSDPX} – prawdopodobieństwo warunkowe tego, że i -ta, j -ta, k -ta, l -ta kategoria zmiennej Z, S, D oraz P znajdzie się w klasie ukrytej t .

Tabela 1. Przykład czterowymiarowej tablicy kontyngencji wraz z oznaczeniami

			Czy Pan/i pije kawę codziennie?		Zmienna P				
Płeć	Czy Pan/i pali papierosy codziennie?	Czy Pan/i pije herbatę codziennie?	tak	nie	Zmienna Z	Zmienna S	Zmienna D	P_1	P_2
Mężczyzna	tak	tak	12	14	Z_1	S_1	D_1	n_{1111}	n_{1112}
		nie	96	125			D_2	n_{1121}	n_{1122}
	nie	tak	54	12		S_2	D_1	n_{1211}	n_{1212}
		nie	19	21			D_2	n_{1221}	n_{1222}
Kobieta	tak	tak	48	98	Z_2	S_1	D_1	n_{2111}	n_{2112}
		nie	2	9			D_2	n_{2121}	n_{2122}
	nie	tak	71	31		S_2	D_1	n_{2211}	n_{2212}
		nie	11	7			D_2	n_{2221}	n_{2222}

Źródło: opracowanie własne.

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

$$\pi_{ijklt}^{ZSDPX} = \pi_t^X \pi_{it}^{Z \setminus X} \pi_{jt}^{S \setminus X} \pi_{kt}^{D \setminus X} \pi_{lt}^{P \setminus X}, \quad (2)$$

gdzie: π_t^X – prawdopodobieństwo przynależności danych obserwacji do klasy t zmiennej ukrytej X ,

$\pi_{it}^{Z \setminus X}$ – prawdopodobieństwo warunkowe tego, że i -ta kategoria zmiennej Z znajdzie się w opisie klasy ukrytej t .

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

$$\sum_{t=1}^T \pi_t^X = \sum_{i=1}^I \pi_{it}^{Z \setminus X} = \sum_{j=1}^J \pi_{jt}^{S \setminus X} = \sum_{k=1}^K \pi_{kt}^{D \setminus X} = \sum_{l=1}^L \pi_{lt}^{P \setminus X} = 1. \quad (3)$$

2.2. Podejście log-liniowe

Model log-liniowy ze zmienną ukrytą X zaproponowany przez Habermana [2] uwzględnia jedynie efekt drugiego rzędu, tzn. interakcje pomiędzy zmiennymi obserwowalnymi a zmienną ukrytą X . Taki model można zapisać w poniższy sposób:

$$\log(m_{ijklt}) = u + u_t^X + u_i^Z + u_j^S + u_k^D + u_l^P + u_{it}^{ZX} + u_{jt}^{SX} + u_{kt}^{DX} + u_{lt}^{PX}, \quad (4)$$

gdzie: $m_{ijklt} = n \sum_{t=1}^T \pi_{ijklt}^{ZSDPX}$,

u – średnia logarytmów częstości oczekiwanych,

u_t^X – główny efekt zmiennej X ze względu na t klasę,

u_i^Z – główny efekt i -tej kategorii zmiennej Z ,

u_j^S – główny efekt j -tej kategorii zmiennej S ,

u_k^D – główny efekt k -tej kategorii zmiennej D ,

u_l^P – główny efekt l -tej kategorii zmiennej P ,

u_{it}^{ZX} – efekt interakcji i -tej kategorii zmiennej Z oraz klasy t zmiennej ukrytej X .

W analogiczny sposób definiowane są pozostałe efekty interakcji.

Parametry modelu (4) spełniają następujące warunki:

$$\begin{aligned} \sum_{t=1}^T u_t^X &= \sum_{i=1}^I u_i^Z = \sum_{j=1}^J u_j^S = \sum_{k=1}^K u_k^D = \sum_{l=1}^L u_l^P = 0, \\ \sum_{i=1}^I u_{it}^{ZX} &= \sum_{j=1}^J u_{jt}^{SX} = \sum_{k=1}^K u_{kt}^{DX} = \sum_{l=1}^L u_{lt}^{PX} = 0, \\ \sum_{t=1}^T u_{it}^{ZX} &= \sum_{t=1}^T u_{jt}^{SX} = \sum_{t=1}^T u_{kt}^{DX} = \sum_{t=1}^T u_{lt}^{PX} = 0. \end{aligned} \quad (5)$$

Dla $T = 1$ model opisany równaniem (4) sprowadza się do modelu wzajemnie niezależnego (*Mutually Independent Model*). W przypadku czterech zmiennych obserwowalnych będziemy rozpatrywać model następującej postaci:

$$\log(m_{ijk}) = u + u_i^Z + u_j^S + u_k^D + u_l^P. \quad (6)$$

Haberman [2] wykazał, że istnieje związek pomiędzy parametrami dwóch różnych od siebie podejść do modelowania klas ukrytych. Na przykład prawdopodobieństwo warunkowe tego, że i -ta kategoria zmiennej Z znajdzie się w klasie ukrytej t , można zapisać:

$$\pi_{it}^{Z \setminus X} = \frac{\exp(u_i^Z + u_{it}^{ZX})}{\sum_{i=1}^I \exp(u_i^Z + u_{it}^{ZX})}. \quad (7)$$

Za pomocą równania (7) można powrócić do podejścia probabilistycznego zaproponowanego przez Lazarsfelda (zob. [5]) i dzięki temu model staje się łatwiejszy w interpretacji.

W przypadku wyboru modelu najlepiej dopasowanego do danych empirycznych można się kierować statystyką chi-kwadrat (χ^2), statystyką największej wiarygodności (L^2) oraz statystyką CR (Cressie-Reada). Jeżeli n_{ijkl} są obserwowanymi liczebnościami, to oczekiwane liczebności są oznaczone przez $\hat{m}_{ijkl}$ i przybierają następującą postać:

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L \frac{(n_{ijkl} - \hat{m}_{ijkl})^2}{\hat{m}_{ijkl}}, \quad (8)$$

$$L^2 = 2 \left(\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L n_{ijkl} \log \left(\frac{n_{ijkl}}{\hat{m}_{ijkl}} \right) \right), \quad (9)$$

$$CR = \frac{2}{\lambda(\lambda + 1)} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \sum_{l=1}^L n_{ijkl} \left[\left(\frac{n_{ijkl}}{\hat{m}_{ijkl}} \right)^\lambda - 1 \right], \quad (10)$$

gdzie λ przyjmuje wartość $2/3$.

Dodatkowymi statystykami stosowanymi w wyborze modelu są kryteria informacyjne: bayesowskie kryterium informacyjne (BIC) oraz kryterium informacji Akaikego (AIC):

$$BIC = L^2 - \ln(n) \cdot (df), \quad (11)$$

$$AIC = L^2 - 2 \cdot (df), \quad (12)$$

gdzie df oznacza liczbę stopni swobody, a n liczbę obserwacji.

W przypadku obydwu kryteriów bardziej preferowany jest taki model, dla którego wartość statystyki (11) lub (12) jest najmniejsza.

3. Wizualizacja danych

Van der Heijden, Gilula oraz van der Ark [4] pokazali, że modele klas ukrytych można przedstawić za pomocą wykresu współrzędnych barycentrycznych (*Bary-*

centric Coordinates). Wykresem tego rodzaju współrzędnych jest trójkąt równoboczny, którego wierzchołki (trzy niewspółliniowe punkty P_1, P_2, P_3) tworzą bazę przestrzeni. Każdemu z tych punktów przyporządkowana jest pewna waga (w tym przypadku zakładamy, 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, przyjmującymi wartości z przedziału $[0, 1]$, a ich suma jest równa 1 ($x + y + z = 1$). Kategorię danej zmiennej można potraktować jako taki punkt P o współrzędnych x, y, z .

Poprzez odpowiednie przekształcenie poszczególnych prawdopodobieństw warunkowych: $\pi_{it}^{Z \setminus X}$, $\pi_{jt}^{S \setminus X}$, $\pi_{kt}^{D \setminus X}$, $\pi_{lt}^{P \setminus X}$ model klas ukrytych można przedstawić za pomocą poniższego równania:

$$\frac{\pi_{ijkl}}{\pi_{i...}} = \sum_{t=1}^T \pi_{ti}^{X/Z} \pi_{jt}^{S/X} \pi_{kt}^{D/X} \pi_{lt}^{P/X}, \quad (13)$$

gdzie:

$$\pi_{it}^{X/Z} = \frac{\pi_t^X \pi_{it}^{Z/X}}{\sum_{i=1}^T \pi_t^X \pi_{it}^{Z/X}} = \frac{\pi_t^X \pi_{it}^{Z/X}}{\pi_{i...}}. \quad (14)$$

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

4. Zastosowanie

Analizę klas ukrytych oraz jej graficzną prezentację przeprowadzono na podstawie danych uzyskanych z bezpłatnej bazy danych Polskiego Generalnego Sondażu Społecznego (PGSS) 1992-2002¹. W niniejszym artykule pod uwagę zostały wzięte dane za rok 2002. Uwzględniono pytania, które zostały zadane wśród osób pracujących. Analiza została przeprowadzona z uwzględnieniem czterech zmiennych i z pominięciem odpowiedzi „nie wiem”. Badana próba liczyła 461 osób.

¹ <http://pgss.iss.uw.edu.pl/>.

Tabela 2. Modele klas ukrytych dla czterech klas

Model	T	df	χ^2	p	L^2	p	CR	p	AIC
Model I	1	63	254,2838	0,0000	124,5380	0,0000	172,4849	0,0000	-1,4620
Model II	2	54	75,5634	0,0280	61,9370	0,2140	65,6927	0,1322	46,0630
Model III	3	45	39,3478	0,7094	39,5405	0,7018	36,9000	0,7995	-50,4595
Model IV	4	36	36,3189	0,4538	34,1396	0,5573	33,3593	0,5948	-37,8604

Źródło: opracowanie własne z wykorzystaniem programu LEM.

Obliczenia przeprowadzono dla czterech klas ($T = 1, 2, 3, 4$) zgodnie ze wzorem (1). Model pierwszy (tab. 2) dany równaniem (4) jest modelem podstawowym w analizie klas ukrytych. Pozwala na określenie, czy pomiędzy zmiennymi występuje zależność, czy też nie. Jak pokazują statystyki L^2 , χ^2 oraz CR , należy przyjąć, że zależność pomiędzy zmiennymi występuje, co w efekcie daje możliwość przeprowadzenia właściwej analizy.

Następnie sprawdzono model dla $T = 2, 3$ oraz 4. Jak pokazują statystyki, model z trzema klasami ukrytymi jest modelem najlepiej dopasowanym do danych empirycznych ($L^2 = 39,5405$, $df = 45$). Można to dodatkowo stwierdzić na podstawie kryterium informacji Akaikego (AIC). Został wybrany model, dla którego wartość AIC jest najmniejsza.

W wyniku przeprowadzenia analizy klas ukrytych uzyskano błąd klasyfikacji 16,45%, co oznacza, że poprawnie zostało sklasyfikowanych 83,55% badanych respondentów.

Klasa druga jest klasą najliczniejszą. Należy do niej prawie 55% respondentów. W klasie tej znalazły się osoby, których miesięczne zarobki nie przekraczają 750 zł. Prawie 88% badanych z tej grupy nie pracuje dodatkowo. Są oni raczej zadowoleni ze swojej pracy, a zapytani: „czy łatwo mógłby Pan/i znaleźć inną pracę zapewniającą takie same dochody i świadczenia, jakie ma Pan/i z obecnej pracy”, większość z nich odpowiedziała, że to nie jest takie łatwe (93,56%).

Do klasy pierwszej należą respondenci, których dochód wynosi więcej niż 1200 zł na miesiąc. Tutaj znalazły się te osoby, które raczej są zadowolone z wykonywanej obecnie pracy i nie dorabiają do swojej podstawowej pensji. 84,63% z badanych, którzy znaleźli się w tej grupie, uważa, że w ogóle nie jest łatwo znaleźć podobną pracę.

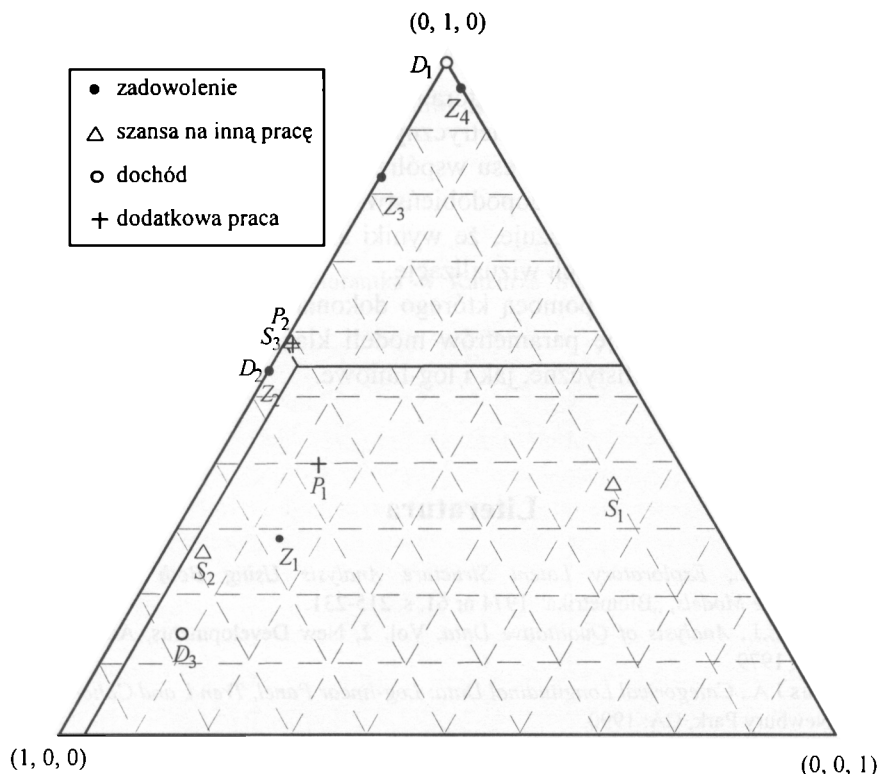
Do klasy trzeciej należy najmniej liczna grupa, ok. 3% badanych. Są to osoby bardzo zadowolone ze swojej pracy. Większość osób w tej grupie (67,27%) podejmuje się różnego rodzaju prac zleconych. Miesięczne wynagrodzenie respondentów, którzy znaleźli się w tej klasie zarabia więcej niż 1200 zł miesięcznie.

W wyniku przekształcenia prawdopodobieństw warunkowych można było przeprowadzić graficzną prezentację danych zawartych w tab. 3. Za pomocą wykresu współrzędnych barycentrycznych można zinterpretować zależności zachodzące pomiędzy kategoriami zmiennych Z , S , D oraz P (rys. 1).

Tabela 3. Wyniki segmentacji dla trzech klas ukrytych

	Pytania	Odszetek respondentów			Przeskalowane wartości zgodnie ze wzorami (13), (14)		
		klasa I	klasa II	klasa III	klasa I	klasa II	klasa III
Zadowolenie	1) Jak bardzo Pan/i jest zadowolony/a z wykonywanej przez siebie pracy?	0,4199	0,5476	0,0325			
	a) bardzo zadowolony/a (Z_1)	0,2704	0,1017	0,9345	0,57	0,28	0,15
	b) raczej zadowolony/a (Z_2)	0,6742	0,5962	0,0000	0,46	0,54	0,00
	c) raczej niezadowolony/a (Z_3)	0,0555	0,2070	0,0000	0,17	0,83	0,00
	d) bardzo niezadowolony/a (Z_4)	0,0000	0,0951	0,0655	0,00	0,96	0,04
Szansa na inną pracę	2) Jak łatwo mógłby Pan/i znaleźć inną pracę zapewniającą takie same dochody i świadczenia, jakie ma Pan/i z obecnej pracy?	0,0087	0,0225	0,5758	0,11	0,36	0,54
	a) bardzo łatwo (S_1)						
	b) raczej łatwo (S_2)	0,1450	0,0418	0,1585	0,68	0,26	0,06
Dochód	c) w ogóle nietatwo (S_3)	0,8463	0,9356	0,2657	0,41	0,58	0,01
	3) Biorąc pod uwagę ostatnie 12 miesięcy, ile wyniosły Pana/i przeciętne miesięczne zarobki (dochody) netto pochodzące z obecnej Pana/i pracy?	0,0000	0,6060	0,0000	0,00	1,00	0,00
	a) do 750 zł (D_1)						
	b) 760-1150 zł (D_2)	0,3252	0,2973	0,0000	0,46	0,54	0,00
Dodatkowa praca	c) powyżej 1200 zł (D_3)	0,6748	0,0967	1,0000	0,77	0,14	0,09
	4) Czy w ciągu ostatnich 12 miesięcy wykonywał(a) Pan/i jakiegś innego stałego, dorywczego lub zleconego prace przynoszące dochód?	0,1907	0,1268	0,6727	0,47	0,41	0,13
	a) tak (P_1)						
	b) nie (P_2)	0,8093	0,8732	0,3273	0,41	0,58	0,01
	Błąd klasyfikacji	0,1645					

Źródło: opracowanie własne z wykorzystaniem programu LEM.



Rys. 1. Wykres współrzędnych barycentrycznych na podstawie danych z tab. 3

Źródło: opracowanie własne z wykorzystaniem programu R.

Na podstawie wykresu można się dowiedzieć dodatkowo, że osoby, które są niezadowolone z pracy, to te, które zarabiają najmniej, tzn. do 750 zł miesięcznie (D_1). Osoby, które twierdzą, że są raczej zadowolone z pracy, zarabiają od 760 zł do 1160 zł netto (D_2). Ich głównym źródłem utrzymania jest tylko pensja (P_2). Mają przeświadczenie, że w przypadku utraty przez nich pracy, nie będzie im łatwo znaleźć podobnej do wykonywanej obecnie (S_3).

Jak można zauważyć na rys. 1, kategoria S_1 znacznie odbiega od innych punktów prezentowanych na wykresie. Prawdopodobnie wynika to z tego, że w klasie trzeciej, która jest najmniej liczna, większość z respondentów uważa, że bardzo łatwo jest znaleźć pracę przynoszącą ten sam dochód co w obecnym miejscu zatrudnienia.

5. Podsumowanie

Analiza klas ukrytych jest metodą, która może być wykorzystywana do badania zależności pomiędzy zmiennymi niemetrycznymi. Analiza ta pozwala na pokazanie tych zależności za pomocą wykresu współrzędnych barycentrycznych poprzez odpowiednie przekształcenie prawdopodobieństw warunkowych.

Przedstawiony przykład pokazuje, że wyniki analizy dostarczają dodatkowej interpretacji poprzez graficzną ich wizualizację.

Program *LEM* [8; 9], za pomocą którego dokonane zostały obliczenia w tej pracy, pozwala na estymację parametrów modeli klas ukrytych ze względu na podejście zarówno probabilistyczne, jak i log-liniowe.

Literatura

- [1] Goodman L.A., *Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models*, „Biometrika” 1974 nr 61, s. 215-231.
- [2] Haberman S.J., *Analysis of Qualitative Data*, Vol. 2, New Developments, Academic Press, New York 1979.
- [3] Hagenaars J.A., *Categorical Longitudinal Data: Log-linear Panel, Trend, and Cohort Analysis*, Sage, Newbury Park, CA: 1990.
- [4] Heijden P.G.M. van der, Gilula Z., Ark L.A. van der, *An Extended Study into the Relationship between Correspondence Analysis and Latent Class Analysis*, „Sociological Methodology” 1999 nr 29, s. 147-186.
- [5] Lazarsfeld P.F., Henry N.W., *Latent Structure Analysis*, Houghton Mill, Boston 1968.
- [6] McCutcheon A.L., *Latent Class Analysis, Quantitative Applications in the Social Sciences*, Sage, Newbury Park, CA 1987.
- [7] Read T.R.C., Cressie N.A.C., *Goodness-of-fit Statistics for Discrete Multivariate Data*, Springer, New York 1988.
- [8] Vermunt J.K., *LEM: A General Program for the Analysis of Categorical Data*, Tilburg University, Tilburg 1997.
- [9] Vermunt J.K., *Log-linear Models for Event Histories*, Sage, Thousand Oaks 1997.

GRAPHICAL PRESENTATION OF RESULTS OF LATENT CLASS MODELS

Summary

Latent class modeling is one of a multivariate analysis techniques of the contingency table. This method was introduced first by Lazarsfeld and Henry [5] and developed by Goodman [1], McCutcheon [6], Habermann [2], Hagenaars [3], and Vermunt [8; 9].

The main purpose of this article is to show some results of latent class modeling by means of barycentric coordinates graph. This kind of presentation permits to describe the relations between categories variable.

The main software used here is *LEM* that allows us to estimate parameters of latent class model.

Iwona Kasprzyk – mgr. doktorantka w Katedrze Statystyki Akademii Ekonomicznej w Katowicach.