PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 47 — 2009
TAKSONOMIA 16
Klasyfikacja i analiza danych – teoria i zastosowania

Iwona Kasprzyk
Akademia Ekonomiczna w Katowicach

PROBLEM WYBORU LICZBY KLAS W ANALIZIE KLAS UKRYTYCH

1. Wstęp

Analiza klas ukrytych jest jedną z wielowymiarowych technik analizy tablic kontyngencji umożliwiającą zidentyfikowanie wzajemnie rozłącznych klas. Metoda została wprowadzona przez Lazarsfelda [1968] i pozwala na analizę danych mierzonych na słabych skalach pomiaru.

W przypadku tej metody pojawia się problem wyboru liczby klas. Jest wiele kryteriów pozwalających na określenie, który z modeli jest lepiej dopasowany do danych, mianowicie można kierować się takimi miarami, jak np.: AIC, BIC, NEC.

W artykule zostaną przedstawione badania symulacyjne określające poprawność wyboru liczby klas ze względu na założoną liczebność próby, wielkość klas, liczbę zmiennych. Dane zostały wygenerowane za pomocą funkcji `poLCA.simdata` znajdującej się w pakiecie `poLCA` w programie R.

2. Analiza klas ukrytych

Załóżmy, że mamy daną tablicę kontyngencji z pięcioma zmiennymi obserwowalnymi: A ($i = 1,2,...,I$), B ($j = 1,2,...,J$), C ($k = 1,2,...,K$), D ($l = 1,2,...,L$), E ($m = 1,2,...,M$). Zmienna ukryta X przyjmuje wartości $s = 1,2,...,S$, gdzie S oznacza liczbę klas. Model z jedną zmienną ukrytą X można przedstawić za pomocą poniższego równania:

$$\pi_{ijklm} = \sum_{s=1}^{S} \pi_{ijklms} , \tag{1}$$

gdzie π_{ijklms} oznacza prawdopodobieństwo warunkowe tego, że *i*-ta, *j*-ta, *k*-ta, *l*-ta, *m*-ta kategoria zmiennej *A*, *B* , *C*, *D* oraz *E* znajdzie się w opisie klasy ukrytej *s*.

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

$$\pi_{ijklms} = \pi_s^X \pi_{is}^{A\backslash X} \pi_{js}^{B\backslash X} \pi_{ks}^{C\backslash X} \pi_{ls}^{D\backslash X} \pi_{ms}^{E\backslash X}, \tag{2}$$

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

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

$$\sum_{s=1}^{S} \pi_s^X = \sum_{i=1}^{I} \pi_{is}^{A\backslash X} = \sum_{j=1}^{J} \pi_{js}^{B\backslash X} = \sum_{k=1}^{K} \pi_{ks}^{C\backslash X} = \sum_{l=1}^{L} \pi_{ks}^{D\backslash X} = \sum_{m=1}^{M} \pi_{ks}^{E\backslash X} = 1. \tag{3}$$

Jeżeli założymy, że zmienne reprezentują pytania zawarte w kwestionariuszu, a kategorie zmiennych − możliwe odpowiedzi, to w wyniku przeprowadzonej analizy klas ukrytej otrzymujemy tablicę, która przedstawia rozkład procentowy liczebności zawartych w początkowej tablicy kontyngencji z podziałem na klasy.

3. Kryteria wyboru liczby klas

W celu wyznaczenia optymalnej liczby klas można się kierować różnymi kryteriami. W badaniu wykorzystano trzy grypy kryteriów: kryteria opierające się na mierze informacji Kullbacka-Leiblera [Kullback, Leibler 1951], kryteria opierające się na podejściu bayesowskim oraz kryteria klasyfikacyjne. W niniejszym artykule przedstawiono kilka najbardziej popularnych kryteriów stosowanych w analizie klas ukrytych.

Najczęściej wykorzystywanym kryterium informacyjnym jest kryterium AIC [Akaike 1974], które można zapisać za pomocą formuły:

$$AIC = -2\log L + 2k, \tag{4}$$

gdzie *L* oznacza funkcję wiarygodności, a *k* − liczbę estymowanych parametrów.

Liczbę parametrów *k* wyznacza się zgodnie z następującą formułą:

$$k = S\sum_{j=1}^{J}(K_j - 1) + (S - 1). \tag{5}$$

gdzie K_j oznacza liczbę kategorii zmiennych , J – zmienne obserwowalne ($j = 1,...,J$).

Jednym z największych zarzutów przypisywanych kryterium AIC jest tendencja do wyboru modeli nazbyt rozbudowanych.

Bozdogan [1994] zaproponował modyfikację kryterium AIC znaną jako AIC_3:

$$AIC_3 = -2\log L + 3k. \tag{6}$$

Bozdogan [1992] zaproponował pewną modyfikację kryterium informacyjnego AIC (CAIC − *constant AIC*), które jest definiowane za pomocą formuły:

$$CAIC = -2\log L + k[(\log N) + 1]. \tag{7}$$

Kryterium informacyjne BIC (*Bayesian Information Criteria*) jest często stosowane w analizie klas ukrytych. W odróżnieniu od kryterium AIC kryterium zaproponowane przez Schwarza [1978] uwzględnia wielkość próby *N*:

$$BIC = -2\log L + k\log N. \tag{8}$$

Scolve [1987] skorygował próbę w kryterium informacyjnym *BIC,* przyjmując za liczbę obserwacji $N = (N+2)/24$. W dalszej części to kryterium będzie oznaczane jako *ABIC*.

Kolejnymi kryteriami, które można wykorzystać do oceny dopasowania modelu, są kryteria klasyfikacyjne. W celu wyboru odpowiedniej liczby klas Celeux i Soromenho [1996] zaproponowali znormalizowane kryterium entropii (NEC):

$$NEC = \frac{E(s)}{\log L_s - \log L_1}, \tag{9}$$

gdzie L_1 oznacza funkcję wiarygodności modelu z jedną klasą, a L_s – funkcję wiarygodności weryfikowanego modelu o liczbie klas równej s, *E*(s) jest miarą entropii dla *s* klasy wskazującą stopień separacji między klasami i definiowaną jako:

$$E(s) = -\sum_{i=1}^{n}\sum_{s=1}^{S}\pi_{is}\log\pi_{is}, \tag{10}$$

gdzie π_{is} oznacza prawdopodobieństwo *a posteriori* tego, że *i*-ta obserwacja należy do klasy *s*.

Prawdopodobieństwo *a posteriori* wyznaczane jest zgodnie z regułą Bayesa:

$$\pi_{is} = \frac{\pi_{ijklms}}{\sum_{v=1}^{S}\pi_{ijklmv}}, \tag{11}$$

gdzie: π_{ijklms} − prawdopodobieństwo tego, że respondent wybierze *i*-ty, *j*-ty, *k*-ty, *l*-ty, *m*-ty wariant odpowiednio zmiennej *A, B, C, D* oraz *E* pod warunkiem znalezienia się w klasie *s*, $\sum_{v=1}^{S}\pi_{ijklmv}$ jest sumą prawdopodobieństw warunkowych tego, że respondent wybiera *i*-ty, *j*-ty, *k*-ty, *l*-ty, *m*-ty wariant odpowiednio zmiennej *A, B, C, D* oraz *E* dla *S* klas ukrytych.

Dzięki prawdopodobieństwu *a posteriori* można odpowiedzieć na pytanie, jakie są szanse respondenta znalezienia się w klasie ukrytej *s*, gdy zaobserwowano rozkład udzielonych przez niego odpowiedzi.

Kryterium NEC nie daje jednoznacznego rozstrzygnięcia, gdy należy się zdecydować na wybór modelu z jedną klasą bądź na model z dwoma klasami.

Biernacki, Celoux, Govaert [2000] zaproponowali kryterium, które jest połączeniem kryterium bayesowskiego BIC oraz klasyfikacyjnego:

$$ICL\ BIC = -2\log L + 2E(s) + k\log N. \tag{12}$$

Za pomocą tych kryteriów wybierany jest ten model, dla którego dane kryterium przyjmuje wartość najmniejszą.

4. Eksperyment

Do estymacji parametrów modelu klas ukrytych wykorzystano pakiet `poLCA` [Linzer, Lewis 2006] pracujący w programie **R**. Parametry modelu zostały oszacowane metodą największej wiarygodności z wykorzystaniem algorytmu maksymalizacji funkcji wiarygodności EM.

Celem przeprowadzonego badania symulacyjnego było wyodrębnienie tych kryteriów, które najczęściej wskazywały poprawny wybór odpowiedniego modelu. Jeśli kryterium przyjmowało wartość najmniejszą, to tym samym wskazywało odpowiedni model, co jest jednoznaczne z poprawnym wyborem modelu.

Charakterystyka badania:

1) dla każdego modelu dane zostały wygenerowane za pomocą funkcji `poLCA.simdata` z pakietu `poLCA`,

2) dane: zmienne dychotomiczne, nominalne wielostanowe o takiej samej liczbie kategorii, nominalne wielostanowe o różnej liczbie kategorii,

3) liczba klas zmiennej ukrytej: 2 i 3,

4) liczba zmiennych: 5 i 8,

5) liczba obserwacji: 300, 600,1200 i 2400,

6) proporcje klas:

- dwie klasy: 20% i 80%; 30% i 70%; 40% i 60%; 50% i 50%,
- trzy klasy: 60%,30% i 10%; 40%,40% i 20%; 70%,20% i 10%,

7) maksymalna liczba iteracji: 3000.

W wyniku przeprowadzonego eksperymentu wygenerowano w sumie 16 800 modeli.

Dla każdego modelu oszacowano parametry modelu właściwego oraz modelu z jedną klasą mniej i jedną klasą więcej od modelu właściwego.

Tabela 1 przedstawia ogólne wyniki przeprowadzonego badania symulacyjnego ze względu na przyjęty rodzaj zmiennej z uwzględnieniem czynników, tj. liczby klas zmiennej ukrytej, liczby zmiennych obserwowalnych, liczby obserwacji oraz proporcji klas. Okazuje się, że dla zmiennych o charakterze dychotomicznym naj-

lepszym kryterium informacyjnym najczęściej wskazującym właściwy model było kryterium AIC, dla zmiennych nominalnych wielostanowych o tej samej liczbie kategorii (tj. 5) – kryterium BIC, natomiast dla zmiennych nominalnych wielostanowych o różnej liczbie kategorii – kryterium ABIC wskazujące poprawność wyboru właściwego modelu w 90,5%.

Tabela 1. Średni procent wskazań właściwego modelu z podziałem na rodzaj zmiennej

Zmienne	AIC	AIC_3	BIC	CAIC	ABIC	NEC	ICL BIC
Dychotomiczne	**72,09**	67,89	67,34	61,61	66,68	50,34	54,70
Nominalne wielostanowe o takiej samej liczbie kategorii	60,30	62,59	**64,79**	52,68	61,45	46,48	58,02
Nominalne wielostanowe o różnej liczbie kategorii	79,45	88,41	89,91	84,14	**90,54**	55,73	57,54

Źródło: opracowanie własne.

Tabela 2. Procent wskazań właściwego modelu z podziałem na rodzaj zmiennych oraz czynniki uwzględnione w badaniu

	AIC	AIC_3	BIC	CAIC	ABIC	NEC	ICL BIC
Liczba klas	Dychotomiczne						
2	90,13	97,34	97,66	97,13	**97,69**	74,09	88,28
3	**48,13**	28,63	26,92	14,25	25,33	18,67	9,92
	Nominalne wielostanowe o takiej samej liczbie kategorii						
2	70,53	92,94	96,34	86,25	93,34	71,66	**99,38**
3	**46,67**	22,13	22,71	7,92	18,92	12,92	2,88
	Nominalne wielostanowe o różnej liczbie kategorii						
2	82,47	99,38	98,28	**100,0**	97,69	94,63	99,38
3	75,42	73,79	78,75	63,00	**81,00**	3,88	1,75
Liczebność próby	Dychotomiczne						
300	**60,71**	54,79	58,36	52,64	59,93	43,07	55,07
600	**67,79**	62,07	59,57	57,64	63,14	47,14	52,57
1200	**76,31**	71,29	70,14	63,36	67,57	54,07	54,43
2400	**83,50**	83,43	81,29	72,79	76,07	57,07	56,71
	Nominalne wielostanowe o takiej samej liczbie kategorii						
300	54,07	42,79	53,57	31,14	52,21	25,14	**59,29**
600	**59,71**	58,21	58,14	51,71	58,71	45,00	58,21
1200	59,14	67,07	**68,29**	58,93	62,93	57,79	57,43
2400	68,29	**82,29**	79,14	68,93	71,93	58,00	57,14
	Nominalne wielostanowe o różnej liczbie kategorii						
300	73,36	70,57	75,00	65,14	**77,00**	54,29	57,93
600	78,50	85,93	87,07	76,86	**87,43**	55,43	57,36
1200	83,00	**97,79**	**97,79**	94,86	**97,79**	56,43	57,71
2400	82,93	99,36	99,79	99,71	**99,93**	56,79	57,14

Tabela 2, cd.

Liczba zmiennych			Dychotomiczne						
5			**72,42**	67,93	66,57	60,75	66,71	50,21	54,36
8			**71,75**	67,86	68,11	62,46	66,64	50,46	55,04
			Nominalne wielostanowe o takiej samej liczbie kategorii						
5			59,86	62,86	**65,96**	52,75	61,36	45,86	57,71
8			60,75	62,32	**63,61**	52,61	61,54	47,11	58,32
			Nominalne wielostanowe o różnej liczbie kategorii						
5			79,57	88,79	90,18	84,43	**90,61**	55,61	57,50
8			79,32	88,04	89,64	83,86	**90,46**	55,86	57,57
Liczebność klas – 2 klasy			Dychotomiczne						
0,2	0,8		90,25	94,50	96,13	92,63	**96,88**	79,38	91,88
0,3	0,7		89,88	97,50	**98,00**	97,63	97,75	72,88	85,88
0,4	0,6		89,75	98,88	98,38	**99,13**	98,00	72,00	87,75
0,5	0,5		90,63	98,50	98,13	**99,13**	98,13	72,13	87,63
			Nominalne wielostanowe o takiej samej liczbie kategorii						
0,2	0,8		68,25	82,00	90,63	70,25	92,25	76,13	**99,38**
0,3	0,7		74,38	93,75	97,13	87,38	88,25	73,25	**98,25**
0,4	0,6		69,50	97,88	98,63	93,13	97,25	70,38	**100,0**
0,5	0,5		70,00	98,13	99,0	94,25	95,63	66,88	**99,88**
			Nominalne wielostanowe o różnej liczbie kategorii						
0,2	0,8		80,38	99,00	98,13	**100,0**	96,38	95,13	99,75
0,3	0,7		83,00	99,75	98,50	**100,0**	**100,0**	93,75	97,88
0,4	0,6		83,50	99,38	98,50	**100,0**	97,63	95,63	**100,0**
0,5	0,5		83,00	99,38	98,00	**100,0**	96,75	94,00	98,88
Liczebność klas – 3 klasy			Dychotomiczne						
0,6	0,3	0,1	**46,38**	22,63	20,50	7,50	17,33	3,63	1,63
0,7	0,2	0,1	**40,00**	18,75	16,88	5,75	13,88	3,38	1,38
0,4	0,4	0,2	**57,90**	44,50	43,38	29,50	45,00	4,63	2,25
			Nominalne wielostanowe o takiej samej liczbie kategorii						
0,6	0,3	0,1	**44,25**	10,75	12,38	0,25	6,25	18,18	10,00
0,7	0,2	0,1	**41,88**	11,63	12,25	0,88	6,00	16,88	9,25
0,4	0,4	0,2	**53,88**	44,00	43,50	22,63	44,50	20,25	10,50
			Nominalne wielostanowe o różnej liczbie kategorii						
0,6	0,3	0,1	71,50	67,13	72,00	53,38	**75,00**	3,63	1,63
0,7	0,2	0,1	72,88	65,88	72,88	54,13	**73,75**	3,38	1,38
0,4	0,4	0,2	81,88	88,38	91,38	81,50	**94,25**	4,63	2,25

Źródło: opracowanie własne.

Pośrednim celem badania było procentowe wskazanie właściwego modelu w podziale na kryteria, rodzaj zmiennej oraz czynniki uwzględnione w badaniu (tab. 3). Okazuje się, że wraz ze wzrostem liczby klas redukuje się poprawność wyboru właściwego modelu wskazywanego przez kryteria. Wzrost liczebności próby prowadzi do poprawy wyboru właściwego modelu. Przy liczebności próby 1200, w przypadku zmiennych nominalnych wielostanowych o różnej liczbie kategorii, kryteria AIC_3, BIC i ABIC wskazują najczęściej właściwy model (97,79%).

Zwiększenie liczby zmiennych nie ma istotnego wpływu na poprawność wskazań właściwego modelu przez analizowane kryteria.

Analizując poprawność wyboru dwóch klas, można stwierdzić, że im mniejsza różnica liczebności klas, tym wyższa poprawność wskazań prawdziwego modelu przez kryteria wyboru liczby klas. W przypadku zmiennych nominalnych wielostanowych o różnej liczbie kategorii można sądzić, iż różnica wynikająca z liczebności klas nie ma wpływu na poprawność wskazań przez kryteria. W tym przypadku kryterium CAIC w 100% poprawnie wskazało właściwy model.

Biorąc pod uwagę poprawność wyboru trzech klas, można zauważyć, że analizowane kryteria najczęściej wskazywały właściwy model, gdy dwie klasy były równoliczne.

5. Podsumowanie

Na podstawie wyników uzyskanych z opisanego eksperymentu można stwierdzić, iż kryteria wzięte pod uwagę w niniejszym artykule wskazują wyższą poprawność wyboru właściwego modelu w przypadku analizy zmiennych wielostanowych o różnej licznie kategorii niż w przypadku zmiennych dychotomicznych oraz nominalnych wielostanowych o tej samej liczbie kategorii.

Liczba zawartych pytań w kwestionariuszu nie ma praktycznie żadnego wpływu na wybór właściwego modelu przez analizowane kryteria wyboru liczby klas.

Z badań symulacyjnych wynika, iż najlepiej budować kwestionariusz o różnej liczbie kategorii, gdyż zwiększa się wówczas odsetek wskazań właściwego modelu przez kryteria informacyjne. W przypadku tego typu zmiennych najlepszym kryterium oceny dopasowania modelu do danych okazało się kryterium ABIC.

Literatura

Akaike H. (1974), *A new look at the statistical model identification*, IEEE Transactions on Automatic Control 19 (6), s. 716-723.

Biernacki C., Celeux G., Govaert, G. (2000), *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*, IEEE Transactions on Pattern Analy sis and Machine Intelligence, 22 (7), s. 719-725.

Bozdogan H. (1987), *Model selection and Akaike's information criterion* (*AIC*)*: the general theory and its analytical extensions,* "Psychometrica" 52, s. 345-370.

Bozdogan H. (1992), *Choosing the number of component clusters in the mixture- model using a new informational complexity criterion of the inverse-fisher information matrix*" [w:] *Information and classification: concepts, methods and applications*, red. O. Opitz, B. Lausen, R. Klar, Springer--Verlag, New York, s. 44-54.

Bozdogan H. (1994*), Mixture-model cluster analysis using model selection criteria and a new information measure of complexity*, [w:] *Proceedings of the Fist US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approac*h, red. H. Bozdogan, vol. 2, Kluwer Academic Publishers, Boston, s. 69-113.

Celeux G., Soromenho G. (1996), *An entropy criterion for assessing the number of clusters in a mixture model,* "Classifcation Journal" 13, s. 195-212.

Dias J.G. (2006), *Data science and classification. model selection for the binary latent class model: a monte carlo simulation,* Heidelberg-Springer-Verlag.
Kullback S., Leibler R.A. (1951), *On information and sufficiency*, "The Annals of Mathematics Statistics" 22, 76-86.
Lazarsfeld P.F., Henry N.W. (1968), *Latent structure analysis*, Houghton Mil, Boston.
Linker D.A., Lewis J. (2006), poLca: Polytomous Variable Latent Lass Analysis, http://userwww.service.emory.edu/~dlinzer/poLCA/.
McLachlan G., Peel D. (2000), *Finite mixture models*, Wiley.
Schwartz G. (1978), *Estimating the dimension of a model,* "The Annals of Statistics", nr 6(2), s. 461-464.
Scolve L.S. (1987), Application of model-selection criteria to some problems in multivariate Analysis, „Psychometrika" 52, s. 333-343.

THE PROBLEM OF CHOOSING THE NUMBER OF CLASSES

Summary

The latent class analysis is one of multivariate techniques of the contingency table which is based on discrete data. This method was introduced by Lazarsfeld [1968].

The main aim of this article is to show simulation research results. The results describe choosing the correct number of classes with respect to number of sample size, number of latent classes and number of variables. In this article, various criteria of choosing the correct model were used.