PRZEGLĄD STATYSTYCZNY
R. LVII – ZESZYT 1 – 2010

ROBERT KAPŁON

KRYTERIA WYBORU LICZBY SKUPIEŃ W BINARNYM MODELU KLAS UKRYTYCH – ANALIZA SYMULACYJNA

1. ZARYS PROBLEMU

Analiza klas ukrytych (*LCA*) pozwala na zidentyfikowanie wzajemnie wyłączających się klas (skupień), które wyjaśniają rozkład przypadków pojawiających się w wielowymiarowej tabeli kontyngencji [23], [15], [16]. Jeśli $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iJ})^T$ będzie wektorem losowym, natomiast $\mathbf{x}_i^* = \left(x_{i1}^*, x_{i2}^*, \ldots, x_{iJ}^*\right)^T$ jego realizacją, wtedy w binarnym modelu *LCA*, x_{ij}^* przyjmuje wartość 0 lub 1 dla obserwacji i oraz zmiennej j.

Z podziałem na skupienia związana jest zmienna ukryta Y_i, której obecność sprawia, że zbiór zmiennych binarnych jest już niezależny, a ona wyjaśnia istotę ewentualnych związków (por. [16]). Przy takich założeniach rozkład warunkowy ma postać:

$$\Pr\left(\mathbf{X}_i = \mathbf{x}_i^* \,\middle|\, Y_i = s\right) = \prod_{j=1}^{J} \Pr\left(X_{ij} = x_{ij}^* \,\middle|\, Y_i = s\right)^{x_{ij}^*} \left[1 - \Pr\left(X_{ij} = x_{ij}^* \,\middle|\, Y_i = s\right)\right]^{1 - x_{ij}^*},$$

natomiast rozkład brzegowy, przy uwzględnieniu prawdopodobieństwa przynależności do klasy s ($s = 1, \ldots, S$), można zapisać:

$$\Pr\left(\mathbf{X}_i = \mathbf{x}_i^*\right) = \sum_{s=1}^{S} \Pr(Y_i = s) \Pr\left(\mathbf{X}_i = \mathbf{x}_i^* \,\middle|\, Y_i = s\right).$$

Dla tak sformułowanego modelu, na etapie estymacji parametrów, należy przyjąć liczbę klas. Zazwyczaj nie wiadomo, ile ma ich być. Z tego też względu szacuje się parametry dla kilku modeli i spośród nich dokonuje wyboru. Pojawia się tutaj pokusa wykorzystania testu opartego na ilorazie wiarygodności. Niestety, wobec niespełnienia warunków regularności, co w konsekwencji prowadzi do nieznajomości rozkładu statystyki testowej, takie podejście jest niewłaściwe (por. [24]). Można próbować aproksymować ten rozkład wykorzystując podejście bootstrapowe [25], lecz czasochłonność obliczeń sprawia, że poszukuje się innych rozwiązań.

Kryteria informacyjne są najczęściej wykorzystywane w modelach opartych na mieszankach rozkładów. Pozwalają one wybrać ten model spośród rozważanych, dla którego wartość danego kryterium jest najmniejsza. Trudność pojawia się w momencie, gdy różne kryteria wskazują na różną liczbę klas. Z tego też względu prowadzi się badania symulacyjne mające rozstrzygnąć, które kryteria są najbardziej wiarygodne.

Takie badania są potrzebne, gdyż po pierwsze – podstawa koncepcyjna wielu kryteriów jest różna. Przykładowo, punktem wyjścia może być informacja Kulbacka – Leiblera lub czynnik bayesowski. Po drugie – w praktyce wykorzystuje się aproksymacje tych kryteriów, co dodatkowo może być obarczone błędem. Po trzecie wreszcie – poprawność kryteriów we wskazaniu właściwego modelu może się zmieniać w zależności od wykorzystywanej klasy modeli. Przykładowo, w wielomianowym modelu logitowym opartym na mieszkankach rozkładów kryterium informacyjne Akaike jest bardziej wiarygodne od kryterium bayesowskiego BIC [3], natomiast dla mieszkanek rozkładów normalnych jest odwrotnie [24].

Mając na względzie fakt, że badania symulacyjne nad wyborem liczby klas skupiają się głównie na mieszankach rozkładów normalnych, niniejsze opracowanie jest próbą rozszerzenia tych badań o analizę klas ukrytych. Dlatego w punkcie drugim przedstawiono typologię kryteriów informacyjnych oraz ich koncepcję. Zwrócono również uwagę na ich własności. W punkcie trzecim opisano eksperyment oraz czynniki, które mogą mieć decydujący wpływ na wartości rozważanych kryteriów. Ostatni punkt odnosi się do wniosków płynących z eksperymentu oraz rekomendacji dotyczących wyboru kryteriów, które najtrafniej wskazują właściwą liczbę klas ukrytych.

2. KRYTERIA WYBORU SKUPIEŃ

KRYTERIA ZE SZKOŁY AKAIKE

Pierwsza grupa kryteriów wywodzi się z tzw. szkoły Akaike. Ich podstawą jest informacja Kulbacka – Leiblera, która mówi o przeciętnym stopniu rozbieżności między prawdziwą gęstością $g(\mathbf{x})$ a gęstością $f(\mathbf{x}|\boldsymbol{\Theta})$ będącą jej aproksymacją ([22], [12]):

$$I\big(g(\mathbf{x}), f\big(\mathbf{x}\,|\,\boldsymbol{\Theta}\big)\big) = E_G\big[\log g(\mathbf{x}) - \log f\big(\mathbf{x}\,|\,\boldsymbol{\Theta}\big)\big].$$

Stosując zasadę minimalizacji tej informacji, z grupy konkurencyjnych modeli wybiera się ten, dla którego obliczona informacja K-L przyjmuje najmniejszą wartość. Ponieważ wartość oczekiwana z $\log g(\mathbf{x})$ nie zależy od żadnego modelu, więc przy porównaniach nie bierze jej się pod uwagę.

Pominięcie wspomnianej wartości oczekiwanej nie rozwiązuje problemu, bo i tak informacja K-L nie jest bezpośrednio obserwowana, gdyż zależy od prawdziwego rozkładu oraz nieznanych parametrów. Z tego też względu nieznaną, prawdziwą dystrybuantę $G(\mathbf{x})$ zastępuje się jej dystrybuantą empiryczną otrzymując następujące oszacowanie (por. [9], [20]).

$$E_{\hat{G}}\big[\log f\big(\mathbf{x}\,|\,\hat{\boldsymbol{\Theta}}\big)\big] = \int \log f\big(\mathbf{x}\,|\,\hat{\boldsymbol{\Theta}}\big)\,d\hat{G}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \log f\big(x_i\,|\,\hat{\boldsymbol{\Theta}}\big).$$

Należy zauważyć, że ten sam zbiór danych został wykorzystany do oszacowania wektora nieznanych parametrów oraz wyznaczenia dystrybuanty empirycznej, co skutkuje obciążeniem estymatora. Okazuje się jednak, że w wielu wypadkach wyznaczenie dokładnej wartości tego obciążenia jest praktycznie niemożliwe i stąd pojawiają się

różne jego aproksymacje prowadzące do różnych kryteriów. Kryterium informacyjne, w którym wykorzystuje się asymptotyczne obciążenie b ma postać (por. [19], [20]):

$$IC = -2N\left[\frac{1}{N}\sum_{i=1}^{N}\log f\left(x_i \mid \hat{\boldsymbol{\Theta}}\right) - \frac{1}{N}b\right] = -2\log L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + 2b.$$

Takeuchi (1976) zaproponował, aby za obciążenie przyjąć $b = \operatorname{tr}(\mathbf{J}^{-1}\mathbf{I})$ gdzie

$$\mathbf{J} = -E\left[\frac{\partial^2 \log L\left(\boldsymbol{\Theta} \mid \mathbf{x}\right)}{\partial \boldsymbol{\Theta}\, \partial \boldsymbol{\Theta}^T}\right], \quad \mathbf{I} = E\left[\frac{\partial \log L\left(\boldsymbol{\Theta} \mid \mathbf{x}\right)}{\partial \boldsymbol{\Theta}} \frac{\partial \log L\left(\boldsymbol{\Theta} \mid \mathbf{x}\right)}{\partial \boldsymbol{\Theta}^T}\right],$$

a odpowiednie pochodne cząstkowe obliczane są w punkcie mody funkcji wiarygodności. Choć obliczenie wartości oczekiwanych może być kłopotliwe, to jednak dla dużej próby – co wynika z mocnego prawa wielkich liczb – można je zastąpić średnią z próby. Biorąc to pod uwagę kryterium informacyjne Takeuchi można zapisać (por. [27]):

$$TIC = -2L\left(\hat{\boldsymbol{\Theta}} \mid x\right) + 2\operatorname{tr}\left(\hat{\mathbf{J}}^{-1}\hat{\mathbf{I}}\right). \tag{1}$$

Ważną cechą tego kryterium jest to, że przy jego wyprowadzeniu nie zakłada się przynależności nieznanego rozkładu $g(\mathbf{x})$ do rodziny rozważanych modeli $\{f(\mathbf{x}|\boldsymbol{\Theta}); \boldsymbol{\Theta} \in \mathrm{P}\}$. Dlatego może ono również posłużyć do rozstrzygnięcia, czy właściwie zdefiniowano model, bo jak wiadomo, gdy $g(\mathbf{x}) \in \{f(\mathbf{x}|\boldsymbol{\Theta}); \boldsymbol{\Theta} \in P\}$ wtedy dwie macierze informacji są sobie równe ($\mathbf{J} = \mathbf{I}$) i tym samym obciążenie równe jest liczbie parametrów r: $b = \operatorname{tr}(\mathbf{J}^{-1}\mathbf{I}) = r$ (por. [21]).

Właśnie takie założenie odnośnie nieznanej gęstości przyjął Akaike [2] otrzymując następujące kryterium:

$$AIC = -2L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + 2r. \tag{2}$$

W przeciwieństwie do *TIC* jest to jedno z najczęściej wykorzystywanych kryteriów porównawczych. Pojawiają się jednak zarzuty, że ma ono tendencję do wskazywania na modele nazbyt rozbudowane nawet wtedy, gdy próba jest duża. Jest to wynik braku zależności kary ($2r$) od liczebności próby. To powoduje również, że estymator ten nie jest zgodny. Choć nie należy tego przeceniać – gdyż przy porównaniach bazujemy na skończonych próbach, a zgodność jest asymptotyczną własnością [9] – to w literaturze proponuje się pewną modyfikację *AIC*. Bozdogan [7] wprowadza zgodne kryterium informacyjne Akaike (*CAIC – consistent AIC*):

$$CAIC = -2L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + r\left(\log N + 1\right). \tag{3}$$

Pewną modyfikacją kryterium *AIC* jest zastąpienie występującej w karze dwójki – trójką. Koncepcja ta jest konsekwencją propozycji aproksymacji statystyki opartej na ilorazie wiarygodności ([7], [30]). Choć jest ona kwestionowana, to jednak symulacje pokazują, że daje dość dobre wyniki przy wyborze właściwego modelu ([11]). Z tego też względu warto ją włączyć do eksperymentu:

$$AIC3 = -2L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + 3r. \tag{4}$$

Rozważane do tej pory kryteria są do siebie podobne w tym sensie, że nakładają tym większą karę na logarytm funkcji wiarygodności, im większa jest liczba estymowanych parametrów. Można więc powiedzieć: za zwiększoną złożoność modelu płaci się większą karą. O złożoności można również mówić w kontekście trudności związanych z estymacją odwrotnej macierzy Fishera. Propozycję takiego kryterium przedstawił Bozdogan ([8], [10]):

$$ICOMP = -2L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + r\log\left[\frac{\mathrm{tr}\left(\hat{\mathbf{J}}^{-1}\right)}{r}\right] - \log\left|\hat{\mathbf{J}}^{-1}\right|. \tag{5}$$

Owa trudność pojawi się wtedy, gdy wystąpią silne zależności miedzy estymowanymi parametrami modelu będące konsekwencją tego, że jest ich zbyt wiele. Wtedy wyznacznik odwrotnej macierzy Fishera będzie bliski zeru, co w konsekwencji nałoży bardzo dużą karę na logarytm funkcji wiarygodności. Warto nadmienić, że kara będzie również wzrastać, wraz ze wzrostem błędów szacunku parametrów – odpowiada za to drugi składnik prawej strony wzoru (5).

KRYTERIUM BAYESOWSKIE

W podejściu bayesowskim wybiera się ten model, który jest najbardziej prawdopodobny, biorąc pod uwagę rozkład *a posteriori*. Porównując dwa modele m_0 i m_1 – mając dane rozkłady warunkowe $f(\mathbf{x}|m_0)$ i $f(\mathbf{x}|m_1)$ oraz rozkłady *a priori* $\phi(m_0)$ i $\phi(m_1) = 1 - \phi(m_0)$ – wygodnie jest wyznaczyć iloraz szans *a posteriori*:

$$\frac{f\left(m_0 \mid \mathbf{x}\right)}{f\left(m_1 \mid \mathbf{x}\right)} = \frac{f\left(\mathbf{x} \mid m_0\right)}{f\left(\mathbf{x} \mid m_1\right)} \frac{\phi(m_0)}{\phi(m_1)} = BF \times \left(\textit{iloraz szans}\right).$$

Pierwszy składnik prawej strony wzoru nazywa się czynnikiem bayesowskim (*BF*), drugi natomiast to iloraz szans *a priori*. Gdy nie ma żadnych preferencji co do wyboru modelu, wtedy $\phi(m_0) = \phi(m_1)$ i czynnik bayesowski równy jest ilorazowi *a posteriori* (por. [18]).

Kluczową kwestią w obliczeniu czynnika bayesowskiego jest znalezienie rozkładu *a posteriori* każdego modelu:

$$f\left(\mathbf{x} \mid m_0\right) = \int f\left(\mathbf{x} \mid \boldsymbol{\Theta}_0, m_0\right) p\left(\boldsymbol{\Theta}_0 \mid m_0\right) d\boldsymbol{\Theta}_0,$$

w którym parametry reprezentowane są przez wektor $\boldsymbol{\Theta}_0$; natomiast $p(\boldsymbol{\Theta}_0|m_0)$ jest rozkładem *a priori* odzwierciedlającym opinię lub wiedzę badacza o parametrach modelu, zanim próba zostanie pobrana. Z rozkładem tym wiążą się dwa problemy. Pierwszy dotyczy przyjęcia jego postaci. Szczególnie uciążliwe jest to wtedy, gdy nie ma żadnych informacji. Drugi problem polega na trudności w obliczeniu całki. W tej sytuacji, jako remedium, proponuje się wykorzystać przybliżenie rozkładu *a posteriori*.

Takim przybliżeniem jest aproksymacja Laplace'a w punkcie mody. Jednak trudności w znalezieniu wartości modalnej powodują, że zostaje ona zastąpiona estymatorami największej wiarygodności. W konsekwencji przybliżenie rozkładu *a posteriori* ma postać [25]:

$$\log f\left(\mathbf{x} \mid m_0\right) \approx L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + \log p\left(\hat{\boldsymbol{\Theta}}\right) - \frac{1}{2}\log\left|\mathbf{H}\left(\hat{\boldsymbol{\Theta}}\right)\right| + \frac{r}{2}\log\left(2\pi\right).$$

Pewną aproksymacją powyższego przybliżenia jest tzw. kryterium Szwarza, które częściej pojawia się jako bayesowskie kryterium informacyjne (*BIC*):

$$BIC = -2L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + r\log N. \tag{6}$$

Kryterium to zakłada pewien szczególny rozkład *a priori* parametrów – rozkład, który zawiera tyle samo informacji co pojedyncza obserwacja. Prowadzi to do płaskiego rozkładu (*spread out*), co niektórzy uznają to za wadę [28]. Z drugiej jednak strony, jeśli brak jest informacji (wiedzy) o rozkładach *a priori*, to przyjęcie takiego rozkładu wydaje się być rozsądnym rozwiązaniem (por. [25]).

KRYTERIA KLASYFIKACYJNE

Uzupełnieniem powyższych kryteriów są kryteria klasyfikacyjne. Ich podstawą jest entropia, której estymator ma postać:

$$EN(\hat{\boldsymbol{\tau}}) = -\sum_i \sum_c \hat{\tau}_{ic} \log \hat{\tau}_{ic}.$$

Jeśli podział na skupienia jest jednoznaczny i oszacowane prawdopodobieństwo *a posteriori* przynależności do danej klasy dąży do 1, wtedy entropia będzie zbliżała się do 0. Im wartość entropii mniejsza, tym większa trudność w wydzieleniu skupień. Należy odnotować, że sama entropia nie może być podstawą porównań modeli o różnej liczbie klas, gdyż zawsze przyjmie wartość większą dla tego z modeli, który ma ich więcej. Dlatego Celeux i Soromenho [13] proponują normalizację (*NEC*):

$$NEC = \frac{EN(\hat{\boldsymbol{\tau}})}{\log L_s\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) - \log L_1\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right)}, \quad s = 2, \ldots, S. \tag{7}$$

Jako punkt odniesienia przyjęto różnicę między logarytmami funkcji wiarygodności dla modelu z s i jedną klasą. Jeśli w porównaniach bierze udział model z jedną klasą, to wtedy kryterium (7) nie można wykorzystać. Biernacki i in. [4] proponują, aby w tym wypadku przyjąć następującą zasadę: jeśli *NEC* dla każdego modelu ($s > 1$) jest większe od 1, wtedy model z jedną klasą należy wybrać.

Kolejnym kryterium, które zostanie wykorzystane to *CLC* (*classification likelihood information*). Entropia pełni tutaj rolę kary nałożonej na logarytm wiarygodności ([24]):

$$CLC = -2\log L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + 2EN(\hat{\boldsymbol{\tau}}). \tag{8}$$

Kryterium *CLC* daje dobre wyniki, jeśli wielkości klas są takie same. Ma jednak tendencję do przeszacowywania prawdziwej liczby klas, jeśli nie nałoży się ograniczenia odnośnie ich wielkości ([5], [6]).

Modele nazbyt rozbudowane, a więc posiadające dużą liczbę parametrów, nie są za to karane, tak jak w przypadku kryterium *BIC*. Dlatego proponuje się kryterium *ICL BIC*, które łączy kryterium bayesowskie i klasyfikacyjne ([24]):

$$ICL\ BIC = -2\log L\left(\hat{\boldsymbol{\Theta}} \mid \mathbf{x}\right) + 2EN(\hat{\boldsymbol{\tau}}) + r\log N.$$

3. OPIS EKSPERYMENTU

W eksperymencie wyróżniono pięć czynników, które mogą mieć istotny wpływ na wartości rozważanych kryteriów informacyjnych: wielkość próby (mała, średnia), liczba zmiennych (mała, średnia), podobieństwo klas (małe, średnie, duże), liczba klas (2, 3) oraz ich wielkość (równa, różna). Specyfika analizy klas ukrytych powoduje, że poszczególne składowe są zależne. W największym stopniu to liczba klas determinuje wartości, jakie przyjmują pozostałe składowe. Wielkość próby dla modelu z dwoma klasami rozumiana jako mała to 500 obserwacji, duża natomiast to 1000. Inaczej jest w wypadku trzech klas, gdyż próby wynoszą odpowiednio 1000 i 2000 obserwacji. Ta różnica spowodowana jest przyjętą strategią eksperymentu, zgodnie z którą szacowane są parametry modelu wzorcowego (2 lub 3 klasy) oraz modeli z jedną klasą więcej i jedna klasą mniej w odniesieniu do tego modelu. Przy zbyt małej liczebności próby (500 obserwacji) oraz średniej liczbie zmiennych (8) pojawiłoby się wiele zerowych liczebności w tabeli kontyngencji, co by utrudniło oszacowanie parametrów modelu z czterema klasami (tab. 1).

Podobnie jest z liczbą zmiennych. Dla modelu z czterema klasami liczba stopni swobody byłaby relatywnie niska przy pięciu zmiennych. Dlatego dla modelu wzorcowego z trzema klasami za małą liczbę zmiennych przyjęto sześć, natomiast w wypadku modelu z dwoma klasami ustalono liczbę zmiennych na poziomie pięciu. Średnia liczba zmiennych jest taka sama i wynosi osiem (tab. 1).

Ważną składową eksperymentu jest podobieństwo między klasami. Jeśli prawdopodobieństwa warunkowe w dwóch klasach są takie same, wtedy jedna z klas jest zbędna. Ogólnie, im różnica w prawdopodobieństwach warunkowych θ dla odpowiednich zmiennych powiększa się, tym podobieństwo miedzy klasami się zmniejsza. Jako miarę podobieństwa przyjęto uśrednioną sumę odległości euklidesowej między klasami c i s. Dla ustalonej liczby zmiennych J miara podobieństwa ma postać:

$$PK_j = \frac{1}{\#\mathcal{Z} \times J}\sum_{\mathcal{Z}} \left\| \hat{\theta}_s - \hat{\theta}_c \right\|, \quad \mathcal{Z} = \{(s,c): s > c \wedge s, c = 1,\ldots,S\}.$$

Jeśli średnia wartość podobieństwa między klasami wynosi 0.1, wtedy podobieństwo traktowane jest jako duże. Przy wartościach 0.15 i 0.2 podobieństwo jest odpowiednio średnie i małe (tab. 1).

Tabela 1

Składowe eksperymentu

2 klasy	3 klasy
Próba (500, 1000)	Próba (1000, 2000)
Liczba zmiennych (5, 8)	Liczba zmiennych (6, 8)
Podobieństwo (0.1, 0.15, 0.2)	Podobieństwo (0.1, 0.15, 0.2)
Wielkość klas (0.5, 0.5), (0.8, 0.2)	Wielkość klas (0.35, 0.35, 0.3), (0.7, 0.15, 0.15), (0.45, 0.4, 0.15)

Źródło: opracowanie własne.

Ostatnim, wyszczególnionym czynnikiem jest wielkość klas. Jeśli jest ona różna, to przy trzech klasach przyjęto dwie możliwości – jedna klasa większa i dwie klasy mniejsze (0.7, 0.15, 0.15) lub dwie klasy większe i jedna mniejsza (0.45, 0.4, 0.15). W wypadku modelu z dwoma klasami wielkości są następujące: 0.8 i 0.2.

Dla tak ustalonych składowych eksperymentu, na podstawie każdego z 60 modeli wzorcowych (24 i 36 dla dwóch i trzech klas odpowiednio), wygenerowano prawdopodobieństwa warunkowe z rozkładu jednostajnego na przedziale (0,1). Procedurę powtórzono 50 razy. Aby przybliżyć ten etap eksperymentu, dla przykładu, niech jego składowe będą następujące (por. tab. 1): liczebność próby – 500 obserwacji, liczba zmiennych – 5, podobieństwo między klasami – 0.1, liczba klas – 2 oraz wielkość każdej klasy – 0.5. Model jaki powstanie po wygenerowaniu prawdopodobieństw warunkowych na podstawie powyższych informacji zostanie nazwany modelem wzorcowym – znana jest jego liczba klas ukrytych.

Następny krok polega na oszacowaniu 2^J (w przykładzie 2^5) liczebności w tabeli kontyngencji przy znanych prawdopodobieństwach warunkowych, prawdopodobieństwach przynależności do klas oraz liczebności próby. Ostatecznie tak otrzymane liczebności stają się podstawą estymacji parametrów – modelu wzorcowego (w przykładzie 2 klasy) oraz modeli o jedną klasę więcej i jedną klasę mniej w stosunku do tego modelu – algorytmem EM (por. [14]).

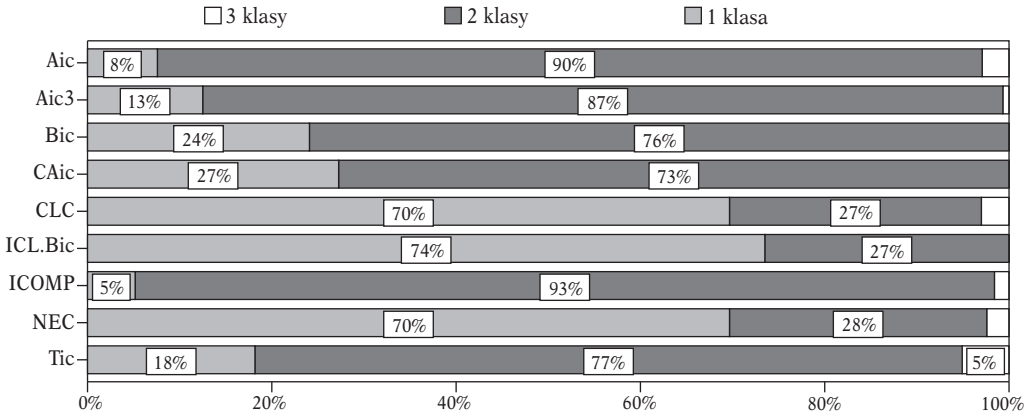
Dla każdego z 9000 modeli obliczono kryteria informacyjne, których wartości porównano (dla odpowiednich modeli) między klasami. Wybór dotyczący liczby klas dokonywany był w oparciu o najmniejszą wartość kryterium informacyjnego. Obliczenia przeprowadzono wykorzystując środowisko do obliczeń statystycznych R [26].

W tym miejscu należy podkreślić, że wybór modelu wzorcowego o 2 i 3 klasach podyktowany był względami praktycznych zastosowań *LCA*. Wynika z nich, że dość często nie potrzeba większej liczby klas do opisania związków w tabeli kontyngencji (por. modele w [23]). Oczywiście, nie zmienia to faktu, że możliwe jest rozszerzenie (zaproponowanie) eksperymentu o modele wzorcowe posiadające więcej niż 3 klasy. Podobna uwaga dotyczy przebiegów symulacyjnych, które można zwiększyć i otrzymać większą dokładność. Jednak dodatkowe badania autora pokazały (nie przedstawiono ich tutaj), że zwiększenie liczby przebiegów do 100, nie wpłynęło w istotny sposób na

otrzymane wyniki i późniejsze wnioski. Z tego względu redukcja do 50 przebiegów wydaje się być uzasadniona, gdyż znacznie skraca czas eksperymentu.

4. WYNIKI EKSPERYMENTU

W ujęciu sumarycznym, najgorzej sprawdziły się kryteria klasyfikacyjne. Gdy model wzorcowy posiadał 2 klasy, wtedy *CLC* i *ICL BIC* wskazały na 27% poprawnych modeli, a kryterium NEC o 1% więcej (rys. 1). Gdy należało wskazać na model z 3 klasami, owe kryteria wypadły jeszcze gorzej, gdyż odsetek poprawnych wskazań spadł odpowiednio do 7%, 4% i 19% (rys. 2).



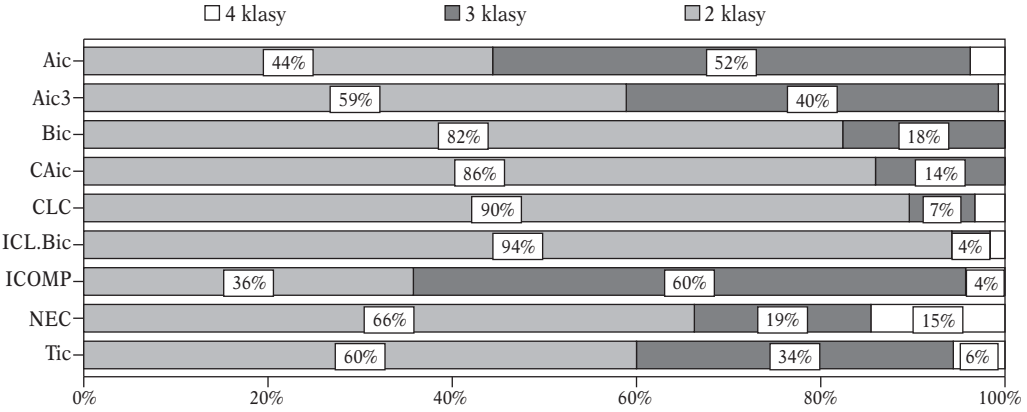
Rysunek 1. Ujęcie procentowe wyboru modelu dla każdego kryterium – model wzorcowy ma 2 klasy

Źródło: opracowanie własne.

Zdecydowanie najlepiej wypadły kryteria informacyjne ze szkoły Akaike'a. Wśród nich kryterium najczęściej wykorzystywane przy porównaniach modeli – *AIC*, poprawnie wskazało na 90% i 52% modeli (rys. 1 i 2). W konfrontacji z bayesowskim kryterium *BIC* (równie często stosowanym), które odnotowało 76% i 18% skuteczności, to raczej kryterium Akaike powinno być brane pod uwagę przy wyborze modeli.

Okazało się również, że bardziej ogólne kryterium *TIC* wypadło gorzej niż *AIC*. Oznacza to, że macierze informacji – o których wspominano w punkcie 2 – nie są równe.

Najwięcej zaufania wzbudza kryterium *ICOMP*, gdyż niezależnie od liczby klas modelu wzorcowego ma największą skuteczność osiągając odpowiednio 93% i 60% (por. rys. 1 i 2). Sprawdza się tutaj koncepcja, że im większa złożoność modelu, mierzona trudnością estymacji odwrotnej macierzy Fishera, tym większą karę należy nałożyć. W konsekwencji wykluczone zostają modele, w których parametry wykazują silną zależność, a błędy ich szacunku są duże.



Rysunek 2. Ujęcie procentowe wyboru modelu dla każdego kryterium – model wzorcowy ma 3 klasy

Źródło: opracowanie własne.

Analiza tabeli 2 i 3 pokazuje, że pewne charakterystyki modeli przyczyniają się do mniejszej poprawności wskazań modelu wzorcowego. Dla modelu z 2 klasami czynnik, jakim jest wielkość klasy, okazał się nieistotnie wpływać na przeciętną poprawność wskazań w wypadku wszystkich kryteriów[1]. Z kolei podobna analiza liczby zmiennych pokazała, że gorsze rezultaty, i to istotne statystycznie, osiągnięto dla 5 zmiennych. Zauważono również, że przy tak samo licznej próbie (1000), lepsze rezultaty pojawiły się, gdy tych zmiennych było 8. Jednak tylko w wypadku wszystkich kryteriów klasyfikacyjnych oraz kryterium *TIC* różnica ta okazała się statystycznie istotna. To może zastanawiać, gdyż wydawałoby się, że jeśli liczba obserwacji przypadających na jedną zmienną rośnie, wtedy poprawność wskazań nie powinna maleć. Postanowiono więc zweryfikować hipotezę, o korelacji między liczbą obserwacji przypadających na jedną zmienną a liczbą poprawnych wskazań. Ponieważ stworzona w tym celu zmienna jest porządkowa (ma 4 poziomy), więc wykorzystano współczynnik korelacji ρ Spearmana. Istotne korelacje na poziomie istotności 0.05 odnotowano dla kryteriów: AIC – 0.49, AIC3 – 0.51, ICOMP – 0.59. Okazuje się, że wraz ze wzrostem obserwacji przypadających na jedną zmienną wzrasta również liczba poprawnych wskazań – przynajmniej dla niektórych kryteriów.

Podobieństwo między klasami wydaje się ważnym elementem decydującym o skuteczności. Obliczony współczynnik korelacji ρ Spearmana potwierdził to przypuszczenie: wzrost podobieństwa między klasami obniża trafność wyboru modelu z 2 klasami (wzorcowego). Choć odnotowane wartości nie dla wszystkich kryteriów okazały się istotne statystycznie: *AIC* – –0.128 (0.55), *AIC3* – –0.347 (0.097), *BIC* – –0.645 (0.001), *CAIC* – –0.655 (0.001), *CLC* – –0.562, (0,004), *ICLBIC* – –0.546 (0,006), *ICOMP* – –0.366 (0.078), *NEC* – –0.571 (0.004), *TIC* – 0.352 (0.091), gdzie w nawiasach podano *p-value*.

[1] W tym celu przeprowadzono test *t*-Studenta. Tutaj, jak i w kolejnych testach za poziom istotności przyjęto wartość 0.1.

Tabela 2

Wyniki eksperymentu dla modelu wzorcowego z 2 klasami

Eks.	Próba	Zm.	Podobne	Klasa		AIC			AIC3			BIC			CAIC			CLC			ICL_BIC			ICOMP			NEC			TIC		
				1	2	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1	500	5	0.10	0.8	0.2	35	15	0	45	5	0	50	0	0	50	0	0	49	0	1	50	0	0	24	24	2	49	0	1	41	8	1
2	1000	5	0.10	0.8	0.2	30	19	1	44	6	0	50	0	0	50	0	0	50	0	0	50	0	0	21	28	1	50	0	0	48	2	0
3	1000	8	0.10	0.8	0.2	0	50	0	0	50	0	17	33	0	31	19	0	50	0	0	50	0	0	0	50	0	50	0	0	1	49	0
4	2000	8	0.10	0.8	0.2	0	50	0	0	50	0	1	49	0	1	49	0	50	0	0	50	0	0	0	50	0	50	0	0	0	46	4
5	500	5	0.15	0.8	0.2	1	45	4	6	43	1	45	5	0	48	2	0	50	0	0	50	0	0	3	41	6	50	0	0	17	27	6
6	1000	5	0.15	0.8	0.2	0	47	3	0	49	1	18	32	0	29	21	0	50	0	0	50	0	0	3	46	1	50	0	0	12	34	4
7	1000	8	0.15	0.8	0.2	0	50	0	0	50	0	0	50	0	0	50	0	23	24	3	44	6	0	0	49	1	23	25	2	0	50	0
8	2000	8	0.15	0.8	0.2	0	50	0	0	50	0	0	50	0	0	50	0	20	28	2	38	12	0	0	50	0	20	29	1	0	47	3
9	500	5	0.20	0.8	0.2	0	42	8	0	49	1	0	50	0	0	50	0	50	0	0	50	0	0	0	48	2	50	0	0	1	39	10
10	1000	5	0.20	0.8	0.2	0	49	1	0	50	0	0	50	0	0	50	0	50	0	0	50	0	0	0	50	0	50	0	0	0	48	2
11	1000	8	0.20	0.8	0.2	0	50	0	0	50	0	0	50	0	0	50	0	0	47	3	0	50	0	0	50	0	0	47	3	0	50	0
12	2000	8	0.20	0.8	0.2	0	50	0	0	50	0	0	50	0	0	50	0	0	48	2	0	50	0	0	50	0	0	48	2	0	44	6
13	500	5	0.10	0.5	0.5	22	25	3	42	7	1	50	0	0	50	0	0	49	1	0	50	0	0	8	41	1	49	1	0	47	2	1
14	1000	5	0.10	0.5	0.5	3	46	1	13	37	0	49	1	0	49	1	0	50	0	0	50	0	0	3	47	0	50	0	0	45	5	0
15	1000	8	0.10	0.5	0.5	0	50	0	0	50	0	0	50	0	0	50	0	50	0	0	50	0	0	0	49	1	50	0	0	0	50	0
16	2000	8	0.10	0.5	0.5	0	50	0	0	50	0	0	50	0	0	50	0	50	0	0	50	0	0	0	50	0	50	0	0	0	47	3
17	500	5	0.15	0.5	0.5	0	41	9	0	48	2	9	41	0	18	32	0	50	0	0	50	0	0	0	49	1	50	0	0	4	38	8

cd. tabeli 2

Eks.	Próba	Zm.	Podobne	Klasa		AIC			AIC3			BIC			CAIC			CLC			ICL_BIC			ICOMP			NEC			TIC		
				1	2	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
18	1000	5	0.15	0.5	0.5	0	49	1	0	49	1	0	50	0	1	49	0	50	0	0	50	0	0	0	50	0	50	0	0	2	47	1
19	1000	8	0.15	0.5	0.5	0	50	0	0	50	0	0	50	0	0	50	0	0	46	4	1	49	0	0	50	0	0	46	4	0	50	0
20	2000	8	0.15	0.5	0.5	0	50	0	0	50	0	0	50	0	0	50	0	0	42	8	0	50	0	0	50	0	0	42	8	0	46	4
21	500	5	0.20	0.5	0.5	0	48	2	0	49	1	0	50	0	0	50	0	47	3	0	49	1	0	0	47	3	47	3	0	0	47	3
22	1000	5	0.20	0.5	0.5	0	49	1	0	50	0	0	50	0	0	50	0	48	2	0	50	0	0	0	50	0	48	2	0	0	48	2
23	1000	8	0.20	0.5	0.5	0	50	0	0	50	0	0	50	0	0	50	0	0	43	7	0	50	0	0	50	0	0	45	5	0	49	1
24	2000	8	0.20	0.5	0.5	0	49	1	0	50	0	0	50	0	0	50	0	0	44	6	0	50	0	0	50	0	0	47	3	0	48	2

Źródło: opracowanie własne.

Dla modelu wzorcowego z 3 klasami zmienna dotycząca wielkości klas okazała się nieistotnie wpływać na przeciętną trafność wyników[2]. Z kolei podobieństwo między klasami jest silnie skorelowane z liczbą poprawnych wskazań (*p-value*): *AIC* – –0.787 (0.000), *AIC3* – –0.766 (0.000), *BIC* – –0.693 (0.000), CAIC – –0.642 (0.000), CLC – 0.366 (0.028), ICLBIC – –0.481 (0.003), *ICOMP* – –0.849 (0.000), *NEC* – 0.133 (0.439), *TIC* – –0.726 (0.000). W największym stopniu ta zależność widoczna jest dla kryteriów ze szkoły Akaike, ze szczególnym wskazaniem na *ICOMP*. Kryteria klasyfikacyjne wykazują słabą lub nawet nieistotną zależność w odwrotnym kierunku: im większe podobieństwo między klasami, tym większa liczba poprawnych wskazań. Taki wniosek z oczywistych względów jest błędny i poddaje w wątpliwość użyteczność tej grupy kryteriów.

Ostatnią z rozważanych składowych eksperymentu zdefiniowano jako liczbę obserwacji przypadających na zmienną. Takie ujęcie, podobnie jak w modelu wzorcowym z 2 klasami, pozwoliło na obliczenie współczynnika korelacji między tą zmienną a liczbą poprawnych wskazań (*p-value*): *AIC* – 0.244 (0.151), *AIC3* – 0.211 (0.217), *BIC* – 0.279 (0.099), *CAIC* – 0.308 (0.067), *CLC* – –0.431 (0.009), *ICLBIC* – –0.331 (0.049), *ICOMP* – 0.226 (0.184), *NEC* – –0.424 (0.01), *TIC* – –0.045 (0.792). Tylko dla niektórych kryteriów korelacje okazały się istotne statystycznie. Wśród nich są kryteria (*BIC*, *CAIC*), dla których wzrost obserwacji w przeliczeniu na jedną zmienną skutkuje nieznacznym wzrostem poprawnych wskazań – jednak wartości współczynników korelacji są raczej niskie. Z kolei dla wszystkich kryteriów klasyfikacyjnych korelacje okazały się istotne, przy czym kierunek zależności budzi wątpliwości.

Współczynniki korelacji, choć użyteczne, opisują siłę związku liniowego. Z tego względu, warto opisać zależność (dowolną) między rozważanymi czynnikami eksperymentu a szansą na wybór modelu wzorcowego z 3 klasami[3]. W tym celu wykorzystano koncepcję uogólnionych modeli liniowych, zgodnie z którą przyjęto, że zmienna zależna ma rozkład dwumianowy, a funkcja wiążąca ma postać kanoniczną (logitowa). Dla tak zdefiniowanego modelu znany jest w literaturze przedmiotu problem nadmiernego rozproszenia (*overdispersion*), który objawia się tym, że w rzeczywistości wariancja jest znacznie większa niż wynika ona z modelu. W wypadku zgromadzonych danych taka sytuacja miała miejsce, więc jako remedium wykorzystano estymację opartą na *quasi* wiarygodności [1]. Do estymacji parametru skali użyto skorygowanej, o liczbę stopni swobody, wartości statystyki χ^2 Pearsona. Obliczenia przeprowadzono dla każdego kryterium (estymacja punktowa i przedziałowa – 95% przedział ufności) i zestawiono je w tabeli 4. Dla ułatwienia interpretacji wartości oszacowanych parametrów poddano transformacji za pomocą funkcji ekspotencjalnej. Ponieważ kryteria klasyfikacyjne odznaczają się bardzo małą skutecznością, dlatego pominięto je przy omówieniu wyników, które zestawiono w tabeli 4.

[2] Wykorzystano analizę wariancji. Ponieważ dla niektórych kryteriów nie można było utrzymać założenia o równości wariancji, dlatego dodatkowo użyto tzw. mocnego testu Welsha.

[3] Z pogłębionej analizy dla modelu wzorcowego z 2 klasami zrezygnowano, gdyż zmiany w liczbie poprawnych wskazań są relatywnie niewielkie, co z kolei objawiało się występowaniem osobliwej macierzy hessianu.

Tabela 3

Wyniki eksperymentu dla modelu wzorcowego z 3 klasami

Eks.	Próba	Zm.	Podobne	Klasa			AIC			AIC3			BIC			CAIC			CLC			ICL_BIC			ICOMP			NEC			TIC		
				1	2	3	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
1	500	6	0.1	0.35	0.35	0.3	33	15	2	42	8	0	50	0	0	50	0	0	41	8	1	45	5	0	25	14	11	23	14	13	31	15	4
2	1000	6	0.1	0.35	0.35	0.3	40	10	0	47	3	0	50	0	0	50	0	0	36	7	7	40	5	5	33	16	1	17	16	17	36	11	3
3	1000	8	0.1	0.35	0.35	0.3	49	1	0	49	1	0	50	0	0	50	0	0	49	1	0	50	0	0	29	21	0	32	5	13	48	1	1
4	2000	8	0.1	0.35	0.35	0.3	43	7	0	50	0	0	50	0	0	50	0	0	50	0	0	50	0	0	17	33	0	43	2	5	50	0	0
5	500	6	0.15	0.35	0.35	0.3	15	26	9	36	12	2	50	0	0	50	0	0	42	6	2	48	2	0	17	27	6	24	10	16	40	9	1
6	1000	6	0.15	0.35	0.35	0.3	10	37	3	29	20	1	48	2	0	49	1	0	49	1	0	50	0	0	10	38	2	46	3	1	41	8	1
7	1000	8	0.15	0.35	0.35	0.3	19	31	0	30	20	0	49	1	0	50	0	0	50	0	0	50	0	0	3	47	0	43	5	2	36	13	1
8	2000	8	0.15	0.35	0.35	0.3	7	43	0	11	39	0	39	11	0	45	5	0	50	0	0	50	0	0	2	48	0	49	0	1	28	22	0
9	500	6	0.2	0.35	0.35	0.3	2	41	7	5	43	2	31	19	0	36	14	0	48	2	0	50	0	0	3	42	5	37	8	5	13	31	6
10	1000	6	0.2	0.35	0.35	0.3	0	47	3	1	48	1	11	39	0	15	35	0	50	0	0	50	0	0	1	47	2	42	6	2	7	42	1
11	1000	8	0.2	0.35	0.35	0.3	1	49	0	2	48	0	15	35	0	23	27	0	50	0	0	50	0	0	0	50	0	41	7	2	3	47	0
12	2000	8	0.2	0.35	0.35	0.3	0	50	0	0	50	0	3	47	0	5	45	0	50	0	0	50	0	0	0	50	0	40	10	0	3	47	0
13	500	6	0.1	0.7	0.15	0.15	33	15	2	45	5	0	49	1	0	49	1	0	39	8	3	43	5	2	30	16	4	24	16	10	35	12	3
14	1000	6	0.1	0.7	0.15	0.15	46	4	0	49	1	0	50	0	0	50	0	0	28	15	7	31	13	6	44	5	1	16	20	14	31	15	4
15	1000	8	0.1	0.7	0.15	0.15	50	0	0	50	0	0	50	0	0	50	0	0	40	9	1	44	6	0	47	3	0	16	18	16	42	5	3
16	2000	8	0.1	0.7	0.15	0.15	50	0	0	50	0	0	50	0	0	50	0	0	48	2	0	50	0	0	40	10	0	34	10	6	44	0	6
17	500	6	0.15	0.7	0.15	0.15	30	20	0	46	4	0	50	0	0	50	0	0	41	6	3	44	5	1	23	21	6	25	13	12	37	10	3
18	1000	6	0.15	0.7	0.15	0.15	36	14	0	46	4	0	50	0	0	50	0	0	37	9	4	42	6	2	26	22	2	25	14	11	46	3	1

cd. tabeli 3

Eks.	Próba	Zm.	Podobne	Klasa			AIC			AIC3			BIC			CAIC			CLC			ICL_BIC			ICOMP			NEC			TIC		
				1	2	3	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4	2	3	4
19	1000	8	0.15	0.7	0.15	0.15	39	11	0	49	1	0	49	1	0	49	1	0	47	2	1	49	1	0	22	28	0	39	9	2	45	3	2
20	2000	8	0.15	0.7	0.15	0.15	19	31	0	37	13	0	50	0	0	50	0	0	50	0	0	50	0	0	12	38	0	48	2	0	42	6	2
21	500	6	0.2	0.7	0.15	0.15	6	39	5	16	33	1	47	3	0	50	0	0	32	12	6	44	4	2	6	40	4	21	16	13	24	16	10
22	1000	6	0.2	0.7	0.15	0.15	4	40	6	13	36	1	32	18	0	39	11	0	43	6	1	46	4	0	9	37	4	36	9	5	26	23	1
23	1000	8	0.2	0.7	0.15	0.15	3	47	0	6	44	0	39	11	0	45	5	0	42	5	3	49	1	0	2	47	1	35	11	4	7	41	2
24	2000	8	0.2	0.7	0.15	0.15	2	48	0	3	47	0	22	28	0	29	21	0	49	1	0	50	0	0	2	48	0	41	9	0	8	39	3
25	500	6	0.1	0.45	0.4	0.15	31	14	5	39	10	1	50	0	0	50	0	0	36	5	9	41	4	5	31	14	5	19	10	21	33	14	3
26	1000	6	0.1	0.45	0.4	0.15	33	16	1	45	5	0	49	1	0	50	0	0	36	8	6	41	6	3	35	12	3	14	15	21	32	14	4
27	1000	8	0.1	0.45	0.4	0.15	50	0	0	50	0	0	50	0	0	50	0	0	46	2	2	46	2	2	46	4	0	25	10	15	48	0	2
28	2000	8	0.1	0.45	0.4	0.15	49	1	0	49	1	0	50	0	0	50	0	0	50	0	0	50	0	0	37	13	0	40	6	4	43	1	6
29	500	6	0.15	0.45	0.4	0.15	27	20	3	43	7	0	50	0	0	50	0	0	39	9	2	46	4	0	27	18	5	22	19	9	36	11	3
30	1000	6	0.15	0.45	0.4	0.15	21	26	3	39	10	1	50	0	0	50	0	0	48	1	1	48	1	1	24	23	3	37	8	5	45	3	2
31	1000	8	0.15	0.45	0.4	0.15	29	21	0	40	10	0	50	0	0	50	0	0	50	0	0	50	0	0	19	31	0	45	2	3	45	2	3
32	2000	8	0.15	0.45	0.4	0.15	15	35	0	22	28	0	48	2	0	49	1	0	50	0	0	50	0	0	10	40	0	42	6	2	36	13	1
33	500	6	0.2	0.45	0.4	0.15	3	35	12	12	36	2	42	8	0	46	4	0	47	3	0	49	1	0	5	38	7	34	8	8	20	24	6
34	1000	6	0.2	0.45	0.4	0.15	2	41	7	5	44	1	25	25	0	27	23	0	50	0	0	50	0	0	5	41	4	45	2	3	12	31	7
35	1000	8	0.2	0.45	0.4	0.15	2	48	0	4	46	0	25	25	0	29	21	0	50	0	0	50	0	0	2	47	1	35	14	1	6	44	0
36	2000	8	0.2	0.45	0.4	0.15	0	50	0	0	50	0	10	40	0	12	38	0	50	0	0	50	0	0	0	50	0	38	12	0	1	43	6

Źródło: opracowanie własne.

Tabela 4

Wyniki estymacji punktowej i przedziałowej szans wskazania modelu wzorcowego z 3 klasami

	AIC			AIC3			BIC			CAIC			ICOMP			TIC		
	Exp(*b*)	*L*	*P*	Exp(*b*)	*L*	*P*	Exp(*b*)	*L*	*P*	Exp(*b*)	*L*	*P*	Exp(*b*)	*L*	*P*	Exp(*b*)	*L*	*P*
Stała	0.26	0.10	0.68	0.11	0.03	0.35	0.02	0.00	0.1	0.02	0.00	0.10	0.40	0.22	0.72	0.21	0.08	0.52
Próba/zmienne																		
A	3.40	1.27	9.52	6.79	2.11	24.24	28.08	5.65	199.39	30.97	5.31	336.9	4.12	2.21	7.84	1.52	0.61	3.86
B	1.14	0.43	3.04	1.28	0.37	4.50	4.47	0.93	27.23	5.46	0.91	56.32	1.16	0.64	2.10	1.13	0.44	2.88
C	0.79	0.30	2.10	1.26	0.37	4.42	3.40	0.69	20.85	3.69	0.57	38.84	1.93	1.07	3.54	1.23	0.48	3.14
D	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Wielkość klas																		
E	0.68	0.29	1.58	0.50	0.17	1.37	0.49	0.14	1.64	0.49	0.12	1.83	0.28	0.16	0.48	0.46	0.20	1.01
F	0.60	0.25	1.39	0.56	0.20	1.54	0.4	0.11	1.36	0.47	0.12	1.76	0.33	0.19	0.57	0.62	0.28	1.35
G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Podobieństwo																		
0.20	36.06	13.80	107.59	68.54	22.28	251.22	20.31	5.93	93.78	12.6	3.62	58.03	33.33	18.15	64.42	15.57	7.14	36.59
0.15	4.37	2.01	9.98	2.72	1.02	7.84	0.28	0.03	1.66	0.13	0.00	1.17	5.97	3.68	9.88	1.21	0.51	2.92
0.10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
Parametr skali	**9.05**			**10.3**			**11.17**			**12.00**			**3.33**			**7.65**		

Próba/zmienne: liczba obserwacji przypadających na zmienną (A-250, B-166.67, C-125, D-83.33); Wielkość klas: E-(0.70, 0.15, 0.15), F-(0.45, 0.40, 0.15), G-(0.35, 0.35, 0.30); Podobieństwo: podobieństwo między klasami.

Źródło: opracowanie własne.

Największy wzrost szans wskazania prawidłowego modelu odnotowano dla zmiennej podobieństwo. Przykładowo, dla kryterium *AIC3* są one ponad 68 razy większe dla podobieństwa na poziomie 0.20 niż 0.10. Przy porównaniu podobieństwa 0.15 i 0.10 ten iloraz szans jest zdecydowanie mniejszy i wynosi 2.72. Ze względu na dość wysoką wartość parametru skali – co sugeruje duży rozrzut wyników – przedział ufności jest szeroki. A więc mamy 95% zaufanie co do tego, że najmniejsze wartości, jakie może ten iloraz przyjąć, wynoszą odpowiednio 22.28 i 1.02. W wypadku podobieństwa na poziomie 0.15 wartość ta jest bliska 1 i tym samym zbliża się do poziomu, który pozwala traktować oszacowany parametr jako nieistotny statystycznie.

Dla kryterium *ICOMP* szanse wskazania modelu wzorcowego są 33 razy większe dla podobieństwa 0.20 i prawie 6 razy większe dla podobieństwa 0.15 w porównaniu z podobieństwem 0.10. Warto odnotować, że ze względu na prawie 3-krotnie niższą wartość parametru skali dolna granica przedziału ufności dla podobieństwa 0.20 wynosi 18.15 i jest zbliżona do wartości dla kryterium *AIC3*. W wypadku pozostałych kryteriów oszacowane ilorazy szans są statystycznie istotne dla podobieństwa 0.20, a tylko dla kryterium *AIC* istotność odnotowano dla podobieństwa 0.15.

Potwierdzają się wnioski z przeprowadzonej analizy wariancji, a dotyczącej wielkości klas i jej wpływu na poprawność wskazań, z jednym wyjątkiem. Otóż dla kryterium *ICOMP* szanse wskazania modelu wzorcowego maleją ponad trzykrotnie jeśli klasy będą różniły się pod względem wielkości. Innymi słowy, szanse wskazania modelu wzorcowego są ponad trzykrotnie większe dla zbliżonych co do wielkości klas (poziom G) w porównaniu z modelami o klasach różnych wielkości (poziom E i F).

Wcześniejsza analiza składowej eksperymentu wyrażającej liczbę obserwacji przypadających na jedną zmienną pokazała, że występuje nieduża, ale statystycznie istotna korelacja między tą zmienną a liczbą poprawnych wskazań dla kryteriów *BIC* i *CAIC*. Okazuje się jednak, że jeśli rozważa się szansę na wskazanie modelu wzorcowego, wtedy jest ona dodatkowo istotna dla kryteriów *AIC*, *AIC3*, *ICOMP* – szanse wyboru modelu wzorcowego są odpowiednio o ponad 3, 6 i 4 razy większe dla dużej liczby obserwacji (poziom A) niż dla małej (poziom D). W wypadku pozostałych poziomów zmiennej (B i C) szanse te w porównaniu z poziomem D nieistotnie różnią się od 1.

5. PODSUMOWANIE

Badania symulacyjne przeprowadzone przez Celeux i Soromenho [13], a dotyczące kryterium klasyfikacyjnego *NEC*, były bardzo obiecujące. We wnioskach stwierdzono, że to kryterium nie ma tendencji do przeszacowywania liczby klas jak *AIC*, jak i również nie jest tak bardzo konserwatywne jak *BIC*. Co więcej, poprawność wskazań modelu opartego na mieszaninie rozkładów normalnych była zbliżona do kryterium *ICOMP*. Niestety, wyniki eksperymentu przedstawione w niniejszym opracowaniu pokazują, że *NEC* jak i pozostałe kryteria klasyfikacyjne (*CLC*, *ICLBIC*) odznaczają się bardzo niską skutecznością we wskazaniu właściwego modelu opartego na analizie klas ukrytych (*LCA*). Można by przypuszczać, że to specyfika modelu *LCA* jest przyczyną tych rozbieżności. Jednak pogłębione symulacje przeprowadzone przez Biernacki i Govaert [6] dla modelu opartego na mieszaninie wielowymiarowych rozkładów normalnych skłaniają

do tego samego wniosku. Wydaje się więc, że przez wzgląd na niską skuteczność, nie powinno się rekomendować kryteriów klasyfikacyjnych jako kryteriów wyboru. Przemawia za tym jeszcze jeden argument, a mianowicie zachowanie się tych kryteriów w odniesieniu do składowych eksperymentu. Eksperyment dla modelu wzorcowego z 3 klasami pokazał, że zależność między liczbą poprawnych wskazań a podobieństwem między klasami i liczbą obserwacji przypadających na jedną zmienną ma charakter odwrotny niż w wypadku pozostałych kryteriów.

Kryteria *AIC* oraz *BIC* są najczęściej wykorzystywane i porównywane ze sobą. Z tego względu warto zaznaczyć, że większą skuteczność, niezależnie od liczby klas modelu wzorcowego, ma kryterium Akaike. Jest ono również bardziej wiarygodne od kryterium ogólniejszego – *TIC*. Najlepsze jednak wyniki odnotowano dla kryterium *ICOMP*. Z tego względu właśnie to kryterium powinno być rekomendowane przy wyborze liczby klas modelu *LCA*.

Warty podsumowania jest wpływ składowych eksperymentu na wyniki, szczególnie dla modelu wzorcowego z 3 klasami. Okazało się, że wzrost podobieństwa między klasami obniża skuteczność kryteriów. Podobnie dzieje się w sytuacji, gdy liczba obserwacji przypadających na zmienną spada. Z kolei wielkość klas nie miała istotnego wpływu, z wyjątkiem kryterium *ICOMP*. Dla niego szanse na wskazanie modelu wzorcowego wzrastają, jeśli wielkości klas będą zbliżone.

Politechnika Wrocławska, Instytut Organizacji i Zarządzania

LITERATURA

[1] Agresti A., [2002], *Categorical Data Analysis*, Wiley-Interscience Publication.
[2] Akaike H., [1973], *Information theory and an extension of the maximum likelihood principle*, [w:] Petrov B.N., Csaki F. (eds.), Second international symposium on information theory (pp.), Budapest: Academiai Kiado, s. 267-281.
[3] Andrews L., Currim I.S., [2003], *A Comparison of segment retention criteria for finite mixture logit models*, „Journal of Marketing Research", 40(2), s. 235-243.
[4] Biernacki C., Celeux G., Govaert G., [1999], *An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model*, Pattern Recognition Letters, 20 (3), s. 267-272.
[5] Biernacki C., Celeux G., Govaert G., [2000], *Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (7), 719-725.
[6] Biernacki C., Govaert G., [1999], *Choosing Models in Model-based Clustering and Discriminant Analysis*, „Journal of Statistical Computation and Simulation", 64, 49-71.
[7] Bozdogan H., [1987], *Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions*, Psychometrika, 52, s. 345-370.
[8] Bozdogan H., [1988], *ICOMP: A new model-selection criterion*, [w:] Bock H., (eds.), Classification and related methods of data analysis, s. 599-608, Amsterdam, Elsevier Science (North-Holland).
[9] Bozdogan H., [2000], *Akaike's Information Criterion and Recent Developments in Information Complexity*, „Journal of Mathematical Psychology", 44, s. 62-91.
[10] Bozdogan H., [1990], *On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models*, Communications in statistics theory and methods, 19, s. 221-278.
[11] Bozdogan H., [1993], *Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix*. [w:] Opitz O., Lausen B., Klar R. (eds.), Information and Classification. Springer, Heidelberg, s. 40-54.

[12] Burnham K.P, Anderson D., [2002], *Model Selection and Multi-Model Inference*, Springer.
[13] Celeux G., Soromenho G., [1996], *An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model*, Classification Journal, 13, s.195-212.
[14] Dempster A.P., Laird N.M., Rubin D.B., [1977], *Maximum likelihood from incomplete data via the EM algorithm*, „Journal of the Royal Statistical Society", Ser. B, No. 1(39), s. 1-22.
[15] Goodman L.A., [1974], *Exploratory Latent Structure Analysis Using Both Identifiable And Unidentifiable Models*, „Biometrika" 61, s. 215-231.
[16] Heinen T., [1996], Latent Class And Discrete Latent Trait Models: Similarities And Differences, Thousand Oaks, California: Sage.
[17] Kapłon R., [2002], *Analiza danych dyskretnych za pomocą metody LCA*, „Taksonomia 9", Prace Naukowe AE we Wrocławiu.
[18] Kass R.E., Raftery A.E., [1995], *Bayes factors*, „Journal of the American Statistical Association", 90, s. 773-795.
[19] Konishi S., Kitagawa G., [1996], *Generalized information criteria in model selection*, Biometrika, 83, s. 875-890.
[20] Konisi S., Kitagawa G., [2003], *Asymptotic theory for information criteria In model selection-functional approach*, „Journal of Statistical Planning and Inference", 114, s. 45-61.
[21] Konisi S., Kitagawa G., [2008], *Information Criteria and Statistical Modeling*, Springer.
[22] Kullback S., [1997], *Information Theory and Statistics*, Dover Publications.
[23] McCutcheon A.L., [1987], *Latent Class Analysis*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-064. Thousand Oaks, CA: Sage.
[24] McLachlan G.J., Peel. D., [2000], *Finite Mixture Models*, New York, Wiley.
[25] McLachlan G.J., [1987], *On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture*, Journal of the Royal Statistical Society Series C (Applied Statistics), 36, s. 318-324.
[26] R Development Core Team, [2009], R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
[27] Raftery A.E., [1999], *Bayes factors and BIC – Comment on "A critique of the Bayesian information criterion for model selection*, Sociological Methods and Research, 27, s. 411-427.
[28] Takeuchi K., [1976], *Distribution of information statistics and a criterion of model fitting*, Mathematical Sciences, 153, s. 12-18, (In Japanese).
[29] Weakliem D.L., [1999], *A Critique of the Bayesian Information Criterion for Model Selection*, Sociological Methods and Research, 27, s. 359-397.
[30] Wolfe J.H., [1971], *A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinomial distributions*, Technical Bulletin STB 72-2, US Naval Personnel and Training Research Laboratory, San Diego.

Praca wpłynęła do redakcji w listopadzie 2009 r.

KRYTERIA WYBORU LICZBY SKUPIEŃ W BINARNYM MODELU KLAS UKRYTYCH – ANALIZA SYMULACYJNA

Wykorzystanie analizy klas ukrytych (*LCA*) wymaga przyjęcia *a priori* liczby klas. W celu rozstrzygnięcia, ile ma ich być, można wykorzystać kryteria informacyjne. Procedura selekcji sprowadza się do: szacowania kilku modeli o różnej liczbie klas, obliczenia wartości kryterium informacyjnego oraz wyboru modelu, dla którego odnotowano najmniejszą wartość tego kryterium. Ponieważ istnieje wiele kryteriów informacyjnych, więc należy zadecydować, które powinno rozstrzygać. Niestety, nie można jednoznacznie wskazać na konkretne kryterium, gdyż w zależności od klasy modelu, zmienia się ich wiarygodność. Taki wniosek wynika z badań symulacyjnych. Biorąc pod uwagę fakt, że najczęściej badania takie dotyczyły mieszanek rozkładów normalnych, dlatego celem niniejszego opracowania jest rozszerzenie tych badań o analizę klas ukrytych.

Słowa kluczowe: analiza klas ukrytych, liczba skupień, kryteria informacyjne, analiza symulacyjna

CRITERIA FOR CHOOSING THE NUMBER OF CLUSTERS OF THE BINARY LATENT CLASS MODEL – SIMULATION ANALYSIS

When using latent class analysis the number of clusters need to be known in advance. In order to decide on this, one can use information criteria. In such a case selection procedure is as follows: estimating a few models with different number of classes, computing information criteria and choosing a model for which a criterion takes the smallest value. Because there are many information criteria one need to determine which of them ought to be decisive. Unfortunately, by virtue of the differences among these criteria, their reliability alter depending on model class. Simulations confirm it as well. Taking into account the fact that simulations mainly concern finite mixtures of normal density functions, therefore in this paper we broaden research to latent class analysis.

Key words: latent class analysis, the number of clusters, information criteria, simulations