B A D A N I A O P E R A C Y J N E I D E C Y Z J E

Nr 3 2009

Robert KAPŁON*

ROZKŁAD *A PRIORI* W CZYNNIKU BAYESOWSKIM A WYBÓR MODELU KLAS UKRYTYCH

Na etapie wyboru liczby segmentów w analizie klas ukrytych kryteria informacyjne są często stosowane. Szczególne miejsce zajmuje tutaj kryterium bayesowskie BIC, które można wyprowadzić – dokonując pewnych uproszczeń – z koncepcji czynnika bayesowskiego. W czynniku tym pojawia się rozkład *a priori* parametrów, którego nie ma w BIC. Z tego względu w pracy podjęto próbę znalezienia takiego rozkładu *a priori*, aby skuteczność tak powstałego kryterium była większa niż skuteczność BIC.

Słowa kluczowe: *analiza klas ukrytych, czynnik bayesowski, rozkład a priori, kryterium informacji BIC, wybór modelu*

1. Wprowadzenie

Wyboru między kilkoma modelami konkurencyjnymi można dokonać, posługując się metodami wnioskowania statystycznego. Przykładem są dość często stosowane testy, oparte na ilorazie funkcji wiarygodności. Jeśli spełnione są warunki regularności, to rozkład chi-kwadrat jest rozkładem granicznym statystyki testowej. Gdy te warunki nie są spełnione – a tak jest w wypadku analizy klas ukrytych – rozkład jest nieznany, co utrudnia weryfikację hipotez. Utrudnia, lecz nie uniemożliwia, gdyż można, wykorzystując podejście bootstrapowe, aproksymować rozkład nieznanej statystyki [4].

Biorąc pod uwagę czasochłonność tych obliczeń, rezygnuje się z takiego podejścia i stosuje się kryteria informacyjne. Jednak i tutaj pojawia się problem, gdyż istnieje wiele kryteriów, o odmiennej podstawie koncepcyjnej, których ocena rozważanych modeli może być odmienna. Z tego też względu prowadzi się badania symulacyjne,

* Instytut Organizacji i Zarządzania Politechniki Wrocławskiej, ul. Smoluchowskiego 25, 50-372 Wrocław, e-mail: robert.kaplon@pwr.wroc.pl

zmierzające do rozstrzygnięcia, które kryteria i dla jakich modeli są najbardziej wiarygodne. W pracy [2] przeprowadzono takie badania dla binarnego modelu klas ukrytych. Okazało się m.in., że kryterium Akaike'a AIC częściej wskazywało model właściwy (wzorcowy) niż kryterium bayesowskie BIC. Ponieważ BIC jest przybliżeniem, przypadkiem szczególnym czynnika bayesowskiego, pojawia się więc pytanie, czy możliwe jest poprawienie wyników dla kryterium BIC poprzez dodatkowe/odpowiednie uwzględnienie rozkładu *a priori* parametrów?

W kontekście tak postawionego pytania celem opracowania jest dobór takiego rozkładu *a priori* parametrów binarnego modelu klas ukrytych w czynniku bayesowskim, aby dla rozważanych w pracy [2] modeli zwiększyć liczbę poprawnych wskazań w stosunku do kryterium BIC.

2. Czynnik bayesowski a BIC

Dokonując wyboru między dwoma konkurencyjnymi modelami M_0 i M_1 – przy założeniu, że preferencje odnośnie do wskazania któregoś z nich są identyczne – można posłużyć się czynnikiem bayesowskim (*bayes factor*):

$$BF = \frac{f(\mathbf{Y} \mid M_0)}{f(\mathbf{Y} \mid M_1)}.$$

Jeśli $BF > 1$, to model M_0 zostaje wybrany. Kluczową kwestią w obliczeniu tego czynnika jest znalezienie rozkładu *a posteriori* dla każdego modelu. Rozkład ten definiuje się następująco:

$$f(\mathbf{Y} \mid M) = \int f(\mathbf{Y} \mid \boldsymbol{\Theta}, M)\, p(\boldsymbol{\Theta} \mid M)\, d\boldsymbol{\Theta}. \tag{1}$$

Parametry modelu M reprezentowane są przez wektor $\boldsymbol{\Theta} = (\Theta_1, ..., \Theta_r)$; $p(\boldsymbol{\Theta}|M)$ jest rozkładem *a priori* i odzwierciedla on wiedzę badacza o nieznanych parametrach modelu, zanim próba zostanie pobrana.

Ze względu na trudności w obliczeniu całki we wzorze (1), dokonuje się aproksymacji. W tym celu rozkład *a posteriori* przedstawia się w postaci (por. [7], [5]):

$$f(\mathbf{Y} \mid M) = \int \exp[h(\boldsymbol{\Theta})]\, d\boldsymbol{\Theta}, \tag{2}$$

gdzie $h(\boldsymbol{\Theta}) = \log f(\mathbf{Y}, \boldsymbol{\Theta}|M)$. Rozwijając $h(\boldsymbol{\Theta})$ w szereg Taylora z resztą Peana w punkcie $\tilde{\boldsymbol{\Theta}}$ otrzymujemy:

$$h(\boldsymbol{\Theta}) = h(\tilde{\boldsymbol{\Theta}}) + (\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})^T h'(\tilde{\boldsymbol{\Theta}}) - 0.5(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}})^T H(\tilde{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}) + o(\| \boldsymbol{\Theta} - \tilde{\boldsymbol{\Theta}}) \|),$$

gdzie macierz hessianu $H(\widetilde{\boldsymbol{\Theta}}) = -h''(\widetilde{\boldsymbol{\Theta}})$. Ponieważ $\widetilde{\boldsymbol{\Theta}}$ jest wartością, w której funkcja przyjmuje maksimum (moda *a posteriori*), wzór (2) można więc zapisać w postaci:

$$f(\mathbf{Y} \mid M) = \int \exp[h(\boldsymbol{\Theta})]d\boldsymbol{\Theta} \approx \exp[h(\widetilde{\boldsymbol{\Theta}})]\int \exp[0.5(\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}})^T H(\widetilde{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \widetilde{\boldsymbol{\Theta}})]d\boldsymbol{\Theta}. \quad (3)$$

Przez podobieństwo funkcji podcałkowej (3) do wielowymiarowego rozkładu normalnego można ją zapisać w postaci:

$$f(\mathbf{Y} \mid M) \approx f(Y \mid \widetilde{\boldsymbol{\Theta}}, M)\, p(\widetilde{\boldsymbol{\Theta}} \mid M)(2\pi)^{r/2} \mid H(\widetilde{\boldsymbol{\Theta}}) \mid^{-1/2}. \quad (4)$$

Jest to tzw. aproksymacja Laplace'a, której błąd jest rzędu $O(n^{-1})$(por. [7], [6]).

Znalezienie wartości $\widetilde{\boldsymbol{\Theta}}$ bywa kłopotliwe, dlatego pewnym rozwiązaniem jest zastąpienie ich estymatorami największej wiarygodności (MLE) $\hat{\boldsymbol{\Theta}}$. Okazuje się, że błąd tej aproksymacji jest tego samego rzędu co (4), jednak należy pamiętać, że jest ona mniej dokładna zwłaszcza wtedy, gdy wzrasta wpływ rozkładu *a priori* w stosunku do funkcji wiarygodności [3]. Biorąc pod uwagę MLE, rząd błędu, dokonując jednocześnie monotonicznego przekształcenia (4) otrzymujemy:

$$-2\log f(\mathbf{Y} \mid M) = -2\log f(\mathbf{Y} \mid \hat{\boldsymbol{\Theta}}, M) - 2\log p(\hat{\boldsymbol{\Theta}} \mid M) - r\log 2\pi + \log \mid H(\hat{\boldsymbol{\Theta}}) \mid -O(n^{-1}).$$

Jeśli obserwacje są niezależne i pochodzą z tego samego rozkładu, próba jest duża, $H(\widetilde{\boldsymbol{\Theta}}) \approx n\mathbf{I}_1$, gdzie $\mathbf{I}_1$ jest macierzą informacji Fishera wyznaczoną dla jednej obserwacji, to powyższe równanie można zapisać w postaci:

$$-2\log f(\mathbf{Y} \mid M) = -2\log f(\mathbf{Y} \mid \hat{\boldsymbol{\Theta}}, M) - 2\log p(\hat{\boldsymbol{\Theta}} \mid M) - r\log 2\pi + r\log n + \log \mid \mathbf{I}_1 \mid -O(n^{-1/2}),$$

lub wykorzystując definicję kryterium bayesowskiego BIC:

$$-2\log f(\mathbf{Y} \mid M) = \mathrm{BIC} - 2\log p(\hat{\boldsymbol{\Theta}} \mid M) - r\log 2\pi + \log \mid \mathbf{I}_1 \mid -O(n^{-1/2}). \quad (5)$$

Z równania (5) można wnioskować, że zastąpienie czynnika bayesowskiego, a dokładnie rozkładu *a posteriori* przez BIC, zwiększa błąd aproksymacji do rzędu pierwszego, czyli $O(1)$. Oznacza to, że przy $n \to \infty$ oszacowanie zbiega do prawdziwej wartości $-2\log f(\mathbf{Y}|M)$ powiększonej o pewną stałą. Okazuje się jednak (por. [6]), że jeśli za rozkład *a priori* przyjąć wielowymiarowy rozkład normalny o wartościach średnich $\hat{\boldsymbol{\Theta}}$ oraz macierzy kowariancji $\mathbf{I}_1$, to równanie (5) redukuje się wtedy do BIC, a błąd aproksymacji do $O(n^{-1/2})$.

W kontekście ostatniej uwagi nasuwa się pytanie, czy możemy zaproponować jakiś inny rozkład *a priori* parametrów, którego uwzględnienie może poprawić skutecz-

ność kryterium BIC. Chodzi więc o to, aby znaleźć taki rozkład *a priori*, dla którego kryterium zdefiniowane następująco:

$$\mathrm{BICP} = \mathrm{BIC} - 2\log p(\hat{\boldsymbol{\Theta}} \mid M) \tag{6}$$

będzie odznaczało się większą skutecznością we wskazywaniu najlepszych modeli niż kryterium BIC.

3. Rozkład *a priori* w czynniku bayesowskim

W analizie klas ukrytych, w której występują zmienne binarne, należy oszacować prawdopodobieństwa tego, że zmienna o indeksie j (j = 1, ..., J) przyjmie wartość 1 pod warunkiem, że należy do klasy s (s = 1, ..., S) oraz prawdopodobieństwa przynależności do tejże klasy. Niech te prawdopodobieństwa wynoszą odpowiednio θ_{js} oraz π_s. Za rozkład *a priori* parametrów, przy założeniu ich niezależności, można przyjąć rozkład Dirichleta (por. [1]):

$$p(\boldsymbol{\theta} \mid \mathbf{a}) = \prod_{s=1}^{S}\prod_{j=1}^{J} \frac{1}{\mathrm{B}(a_s, a_s)}\, \theta_{js}^{a_s - 1}(1-\theta_{js})^{a_s - 1},$$

$$p(\boldsymbol{\pi} \mid \mathbf{a}) = \frac{\Gamma(a_1 + \ldots + a_S)}{\Gamma(a_1)\cdot\ldots\cdot\Gamma(a_S)} \prod_{s=1}^{S} \pi_s^{a_s - 1}.$$

Na nieznane parametry $\mathbf{a} = (a_1, \ldots, a_S)$ nałożono ograniczenie, tzn. przyjęto, że będą one równe w każdej klasie w obrębie rozważanego modelu. Dodatkowo, a to już wynika ze specyfiki rozkładu, każdy parametr jest większy od zera. Uwzględniając to, logarytm rozkładu łącznego dla modelu o S klasach można zapisać następująco:

$$\log p_S(\boldsymbol{\Theta} \mid a_S) = \log p_S(\boldsymbol{\theta}, \boldsymbol{\pi} \mid a_S) = \Delta_S(a_S) + (a_S - 1)\Phi_S(\boldsymbol{\Theta}, a_S), \tag{7}$$

gdzie:

$$\Delta_S(a_s) = \log \frac{\Gamma^{-S}(a_S)\Gamma(Sa_S)}{\mathrm{B}(a_S, a_S)}, \quad \Phi_S(\boldsymbol{\Theta}, a_S) = \sum_{s=1}^{S}\sum_{j=1}^{J} \log \theta_{js}(1-\theta_{js}) + \sum_{s=1}^{S} \pi_s .$$

4. Opis eksperymentu

W pracy [2] przeprowadzono eksperyment skuteczności kryteriów informacyjnych, w tym kryterium BIC. Plan eksperymentu zakładał, że znany jest model – na-

zwano go modelem wzorcowym. W konsekwencji obliczone kryteria dla tego modelu powinny być mniejsze niż dla modeli konkurencyjnych. Im częściej taka sytuacja występowała, tym bardziej wiarygodne było kryterium informacyjne.

Model wzorcowy otrzymywano w ten sposób, że generowano parametry modelu klas ukrytych (prawdopodobieństwa warunkowe), uwzględniając takie składowe eksperymentu jak: wielkość próby, liczbę zmiennych, podobieństwo klas, liczba klas ukrytych oraz ich wielkość. Różne kombinacje poziomów powyższych składowych pozwoliły na otrzymanie 20, 24 i 36 modeli wzorcowych odpowiednio o 1, 2 i 3 klasach ukrytych[1]. To z kolei dało podstawę do wygenerowania obserwacji i oszacowania parametrów modelu wzorcowego o liczbie klas w (w = 1, 2, 3) oraz modelu o jedną klasę więcej – $w + 1$ i jedną klasę mniej – $w - 1$. Oczywiście, jeśli $w = 1$, to oszacowano tylko model z dwiema klasami. Procedurę generowania prawdopodobieństw powtórzono 50 razy dla każdego modelu, szacując jednocześnie parametry 11 tys. modeli. Zgromadzony materiał statystyczny, w oparciu o (6) i (7), pozwolił obliczyć interesujące kryterium informacyjne w funkcji nieznanego parametru a_s dla klasy s:

$$\mathrm{BICP}_s = \mathrm{BIC}_s - 2\log p_s(\hat{\boldsymbol{\Theta}} \mid a_s)\,, \tag{8}$$

gdzie $\hat{\boldsymbol{\Theta}}$ jest estymatorem największej wiarygodności parametrów modelu.

Jeśli w jest modelem wzorcowym, to obliczona dla niego wartość kryterium BICP powinna być mniejsza od wartości tego kryterium obliczonego dla modeli konkurencyjnych, co odpowiada następującemu warunkowi:

$$\underset{\substack{t\in\{-1,1\}\wedge w>1\\ \text{lub } t=w=1}}{\forall} \mathrm{BICP}_{w+t} > \mathrm{BICP}_w, \quad w = 1, 2, 3. \tag{9}$$

Z kolei uwzględniając (8) i (9), poszukuje się takich wartości a_1, a_2, a_3, a_4, aby poniższa nierówność

$$\underset{\substack{t\in\{-1,1\}\wedge w>1\\ \text{lub } t=w=1}}{\forall} \log\frac{p_{w+t}(\hat{\boldsymbol{\Theta}} \mid a_{w+t})}{p_w(\hat{\boldsymbol{\Theta}} \mid a_w)} < \frac{1}{2}\mathrm{BICP}_{w+t,w}, \quad \text{gdzie} \quad \mathrm{BIC}_{w+1,w} = \mathrm{BIC}_{w+t} - \mathrm{BIC}_w \tag{10}$$

zachodziła jak najczęściej. Poszukiwania prowadzone są przy następujących warunkach:

a) wartości parametrów należą do przedziału (0, 2] i wyliczane są z dokładnością do części setnych;

b) parametry a_1, a_2, a_3, a_4 tworzą ciąg monotoniczny;

[1] W przywołanej pracy modelu wzorcowego z jedną klasą ukrytą nie rozważano. Tutaj jest to konieczne, gdyż poszukuje się wartości parametrów rozkładu *a priori*. Jeśliby z tego zrezygnować, to mogłoby się okazać, że poszukiwane parametry przyjmą wartość gwarantującą wybór modelu z dwiema klasami niezależnie od tego, czy model z jedną klasą byłby bardziej odpowiedni.

c) uwzględnia się tylko taki zbiór wartości parametrów, dla których efektywność kryterium $BICP_w$ jest nie mniejsza niż dla BIC_w,

Biorąc pod uwagę skomplikowaną naturę warunku (10), wykorzystano metodę przeszukiwania sieciowego (*grid search*) przestrzeni parametrów. Założono więc, zgodnie z punktem a), że $a_i = \{0.01, 0.02, ..., 2\}$ dla każdego $i = 1, 2, 3, 4$. Przestrzeń parametrów zdefiniowano jako iloczyn kartezjański: $a_1 \times a_2$, $a_1 \times a_2 \times a_3$, i $a_2 \times a_2 \times a_4$ odpowiednio dla modelu wzorcowego z 1, 2 i 3 klasami ukrytymi. Dla każdej zdefiniowanej kombinacji parametrów zliczano, ile razy nierówność (10) zachodzi. Maksymalne wartości, wiążące się ze stuprocentową skutecznością kryterium BICP, są równe liczbie modeli wzorcowych przemnożonej przez liczbę powtórzeń, a więc: 1000, 1200 i 1800.

W tym miejscu należy wspomnieć, że przyjęcie jako maksymalnej wartości parametru liczby 2 nie jest w istocie ważnym ograniczeniem. Wstępne badania symulacyjne pokazały, że zwiększenie zakresu zasadniczo nie wpływa na wyniki. Ta sama uwaga dotyczy rzędu dokładności parametrów. Ważne natomiast jest to, że w tym przedziale znajduje się wartość 1, czyli wartość, dla której rozważane kryterium redukuje się w przybliżeniu do kryterium BIC.

Nałożona monotoniczności w punkcie b) jest istotna, gdyż bez niej optymalne wartości parametru a_4 będą wyznaczane przy warunku, że model z 4 klasami nie będzie wybierany. Tym samym próba rozszerzenia zagadnienia na większą liczbę klas i włączenia modelu wzorcowego z 4 klasami pokazałaby, że otrzymane wartości są niewłaściwe. Inaczej jest w wypadku pozostałych parametrów. Przykładowo, z jednej strony a_3 dobierany jest tak, aby model z 3 klasami nie został wybrany, z drugiej natomiast wręcz odwrotnie.

Nowe kryterium nie powinno być mniej efektywne niż kryterium BIC. Wtedy jest sens jego wprowadzenia. Dlatego ze zbioru potencjalnych wartości parametrów wybrano te, dla których liczba poprawnych wskazań modeli wzorcowych dla BICP jest większa od BIC.

5. Eksperyment i jego wyniki

Otrzymano pokaźny zbiór wartości parametrów, spełniający warunki opisane w rozdziale trzecim. Aby dokonać ich wyboru, należy rozstrzygnąć, dla których z nich sumaryczna poprawność wskazań jest największa. Trzeba jednak pamiętać, że różne kombinacje składowych eksperymentu dostarczyły różnej liczby modeli wzorcowych, dlatego przed zsumowaniem otrzymane wyniki podzielono przez 20, 24 i 36 – odpowiednio dla modelu wzorcowego z 1, 2 i 3 klasami ukrytymi.

Zdecydowano również, że nie zostaną wybrane te wartości parametrów, dla których osiągnięto maksymalny wynik, gdyż – jak w każdym badaniu statystycznym – powtórzenie eksperymentu nie gwarantuje otrzymania tych samych wartości parame-

trów. Mając to na względzie, wybrano wszystkie te a_1, a_2, a_3, a_4, dla których sumaryczna poprawność wskazań nie różniła się więcej niż o 1% od wartości najlepszej. Następnie dla takiego zbioru zbudowano model regresji, w którym zmienną zależną były parametry, natomiast zmienna niezależna reprezentowała liczbę klas. W wyniku doboru odpowiedniego modelu i estymacji jego parametrów otrzymano

$$\hat{a}_s = \underset{(0,004)}{0,26} + \underset{(0,006)}{1,72}\,\frac{1}{s}, \quad s = 1, 2, 3, 4, \tag{11}$$

gdzie s oznacza liczbę klas rozważanego modelu, a wartości błędów oszacowań podano w nawiasie. Taki model jest bardzo dobrze dopasowany do danych empirycznych, na co wskazuje wysoka wartość współczynnika determinacji, przekraczająca 0,99.

Wyznaczone na podstawie równania (11) wartości parametrów w rozkładzie *a posteriori* przyczyniają się do większej skuteczności kryterium BICP niż BIC. Dokładne wartości poprawnych wskazań dla rozważanych kryteriów zawiera zamieszczona tabela.

Tabela. Sumaryczna liczba poprawnych wskazań

Model	Liczba poprawnych wskazań dla modelu wzorcowego		
	1 klasa	2 klasy	3 klasy
BICP	1000	947	501
BIC	1000	911	317

Źródło: Opracowanie własne.

Wzrost skuteczności kryterium BICP dla modelu wzorcowego z 2 klasami jest nieznaczny (niecałe 4%), gdyż BIC dość dobrze radziło sobie ze wskazywaniem właściwego modelu. Jednak efektywność BIC drastycznie się obniżyła, gdy model wzorcowy posiadał 3 klasy. Toteż dodanie do BIC rozkładu *a priori* spowodowało wzrost skuteczności o 58%.

Bibliografia

[1] CONGDON P., *Bayesian Models for Categorical Data*, Wiley 2005.

[2] KAPŁON R., *Liczba skupień w binarnym modelu klas ukrytych*, Raport Serii PRE, Politechnika Wrocławska, 2009.

[3] KASS R.E., RAFTERY A.E., *Bayes Factors*, Journal of the American Statistical Association, 1995, 90(430), s. 773–795.

[4] MCLACHLAN, G.J., *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), 1987, 36, s. 318–324.

[5] RAFTERY A.E., *Bayes factors and BIC – Comment on "A critique of the Bayesian information criterion for model selection"*, Sociological Methods and Research, 1999, 27, s. 411–427.

[6] RAFTERY A.E., *Bayesian model selection in social research (with discussion)*, Sociological Methodology, 1995, 25, s. 111–196.
[7] TIERNEY L., KADANE J.B., *Accurate approximations for posterior moments and marginal distributions*, Journal of the American Statistical Association, 1986, 81(393), s. 82–86.

Prior distributions for Bayes factors and latent class model selection

Estimating the values of parameters in latent class analysis, one needs to know the number of clusters in advance. It is crucial to determine a criterion which enables confirmation of the superiority of one number of classes over the others. A statistical approach, which is based on a likelihood ratio test (LRT), contends with the difficulties of assessing the null distribution of LRT statistics. As a remedy, information criteria like the Bayesian information criterion (BIC) can be used. This criterion is an approximation of a Bayes factor that depends on the prior distribution. Apparently, if one combines BIC and a suitable prior, the effectiveness of such a criterion increases in comparison to the standard BIC.

In this article we propose such a prior distribution. In order to do this, a simulation study is carried out and the data collected enable the construction of a nonlinear regression model. The number of classes and the values of the required parameter are chosen as the predictor and the dependent variable, respectively. Such an approach enables the estimation of the values of the parameters *a priori* given the number of clusters. The performance of the new criterion is better than the Bayesian information criterion by up to 58%.

Keywords: *latent class analysis, Bayes factor, prior distribution, BIC information criterion, model selection*