

УДК 519.248

МИНИМАКСНОЕ ОЦЕНИВАНИЕ И ПРОГНОЗИРОВАНИЕ ДЛЯ ГРУППИРОВАННЫХ БИНАРНЫХ ДАННЫХ С ИСКАЖЕНИЯМИ

М.А. ПАШКЕВИЧ

*Белорусский государственный университет
пр. Ф. Скорины, 4, Минск, 220050, Беларусь*

Поступила в редакцию 11 января 2005

Предлагаются методы робастного оценивания и прогнозирования для бета-иерархических моделей группированных бинарных данных с искаженными результатами наблюдений. Доказано, что оценки параметров моделей, оптимальные в смысле минимакса смещения, вычисляются на середине интервалов искажений, а прогноз, оптимальный в смысле минимакса риска, достигается при максимально возможных уровнях искажений. Теоретические результаты иллюстрируются компьютерным моделированием.

Ключевые слова: группированные бинарные данные, оценивание, прогнозирование, бета-иерархические модели, искажения, робастность.

Введение

Бета-иерархические модели широко используются при описании стохастических свойств группированных бинарных данных, поскольку в отличие от классической биномиальной модели позволяют учесть межгрупповую корреляцию и взаимную зависимость результатов наблюдений [1]. В основе таких моделей лежит бета-биномиальное распределение [2], для которого байесовская прогнозирующая функция представляется в явном виде. Это обуславливает практическую значимость рассматриваемого класса моделей для медицины и экономики [3]. При этом наиболее часто применяются бета-биномиальная и бета-логистическая варианты бета-иерархических моделей.

Бета-биномиальная модель (ББМ) была предложена Пирсоном [4] и развита далее в работе Скеллама [5]. Для оценивания параметров ББМ традиционно используются методы моментов и максимального правдоподобия [6], а прогнозирование на основе ББМ осуществляется с помощью байесовского предиктора, имеющего бета-распределение. Бета-логистическая модель (БЛМ) была предложена Хекманом и является обобщением ББМ [7]. При этом идентификация параметров БЛМ проводится методом максимального правдоподобия, а при прогнозировании используется подход, аналогичный ББМ. Обе модели в случае отсутствия искажений в наблюдаемых данных позволяют значительно повысить точность прогнозирования по сравнению с биномиальной моделью [8].

Однако искажения в данных существенно влияют на качество статистических выводов для бета-иерархических моделей. В предыдущих работах [9, 10] автором были получены выражения, позволяющие оценить робастность различных методов оценивания ББМ и БЛМ, а также увеличение риска прогнозирования для случая, когда уровни искажений известны с точностью до значений. При этом проведенные численные эксперименты показали, что даже при небольших уровнях искажений классические методы могут приводить к большим ошибкам в прогнозе. По-

этому актуальна задача разработки новых, устойчивых к искажениям методов статистического оценивания и прогнозирования для бета-иерархических моделей.

В данной работе предлагаются робастные методы оценивания и прогнозирования для БМ и БЛМ в случае, когда уровни искажений известны с точностью до интервала. При этом доказано, что оценки параметров моделей, оптимальные в смысле минимакса смещения, вычисляются при средних значениях уровней искажений, а прогноз, оптимальный в смысле минимакса риска, достигается при максимально возможных уровнях искажений. Полученные теоретические результаты иллюстрируются компьютерным моделированием.

Постановка задачи

Пусть результаты наблюдений описываются набором k бинарных векторов-строк $B = (B_1, B_2, \dots, B_k)$, $B_i \in \{0,1\}^{n_i}$, где $B_i = (B_{i1}, B_{i2}, \dots, B_{in_i})$ — результаты серии испытаний над i -м объектом, причем $B_{ij} = 1$, если в испытании j для объекта i некоторое случайное событие имело место, и $B_{ij} = 0$ в противном случае. Объекту номер i в испытании номер j поставлен в соответствие некоторый m -вектор факторов $Z_i \in R^m$, который описывает свойства объекта. При этом предполагается, что размеры серий испытаний n_1, n_2, \dots, n_k малы, а объекты обладают свойством "слабой неоднородности" [1].

Для описания стохастических свойств рассматриваемых данных используется бета-иерархическая модель, основанная на следующих предположениях.

П1. Для i -го объекта вероятность успеха p_i постоянна в течение всей серии испытаний.

П2. Вероятность успеха p_i является случайной величиной, имеющей бета-распределение с параметрами α_i^0, β_i^0 , причем p_1, p_2, \dots, p_k независимы в совокупности.

П3. Параметры бета-распределений α_i, β_i связаны с факторами Z_i выражениями $\alpha_i = f_\alpha(Z_i)$, $\beta_i = f_\beta(Z_i)$, где $f_\alpha(\cdot), f_\beta(\cdot)$ — некоторые функции.

В данной работе рассматриваются две наиболее широко используемые модели из этого класса: бета-биномиальная модель (БМ) и бета-логистическая модель (БЛМ), которые определяются следующими дополнительными предположениями:

БМ: $f_\alpha(Z_i) = \alpha^0$, $f_\beta(Z_i) = \beta^0$, $n_i = n$;

параметры модели: $n \in N$, $\alpha^0, \beta^0 \in R$.

БЛМ: $f_\alpha(Z_i) = \exp(Z_i^T a^0)$, $f_\beta(Z_i) = \exp(Z_i^T b^0)$;

параметры модели: $n_1, n_2, \dots, n_k \in N$, $a^0, b^0 \in R^m$.

Заметим, что размеры серий испытаний n для БМ и n_1, n_2, \dots, n_k для БЛМ предполагаются известными.

Пусть данные B подвержены аддитивным стохастическим искажениям, и наблюдается искаженная бинарная матрица \tilde{B} :

$$\tilde{B}_{ij} = B_{ij} \oplus \eta_{ij}, \quad (1)$$

где \oplus — операция сложения по модулю два, а $\{\eta_{ij}\}$ — независимые случайные бинарные величины, причем $i = 1, \dots, k$, $j = 1, \dots, n_i$. При этом для каждого i, j имеет место следующая зависимость случайной величины η_{ij} от наблюдений B_{ij} :

$$P\{\eta_{ij} = 1 | B_{ij} = 0\} = \varepsilon_0, \quad P\{\eta_{ij} = 1 | B_{ij} = 1\} = \varepsilon_1, \quad \varepsilon_0 \in [\varepsilon_0^{\min}, \varepsilon_0^{\max}], \quad \varepsilon_1 \in [\varepsilon_1^{\min}, \varepsilon_1^{\max}], \quad (2)$$

где $\varepsilon_0^{\min}, \varepsilon_0^{\max}, \varepsilon_1^{\min}, \varepsilon_1^{\max}$ — известные границы интервалов для уровней искажений $\varepsilon_0, \varepsilon_1$, точные значения которых неизвестны. Задача заключается в робастном оценивании параметров БМ и БЛМ и прогнозировании вероятностей $\{p_i\}$ на их основе в случае искажений (1), (2) с уровнями, известными с точностью до указанных интервалов.

Робастное оценивание параметров моделей

Рассмотрим задачу построения b -робастных оценок параметров бета-иерархической модели, обеспечивающих минимальное смещение при "наихудших" значениях уровней искажений из заданных интервалов. При этом предлагаемый подход применим сначала к оценкам, построенных по методу моментов (ММ-оценок). В работе [9] автором было доказано, что при известных уровнях искажений $\varepsilon_0, \varepsilon_1$ и истинных значениях параметров модели α^0, β^0 справедливы следующие асимптотические разложения для смещения ММ-оценки параметров ББМ:

$$\Delta\alpha(\varepsilon_0, \varepsilon_1) = (\alpha^0 + 2\beta^0 + 1)\varepsilon_0 + \left[\alpha^0(\alpha^0 + 1)/\beta^0\right]\varepsilon_1 + o(\varepsilon_0, \varepsilon_1), \quad (3)$$

$$\Delta\beta(\varepsilon_0, \varepsilon_1) = \beta^0 + \beta^0(\beta^0 + 1)/\alpha^0 \varepsilon_0 + (2\alpha^0 + \beta^0 + 1)\varepsilon_1 + o(\varepsilon_0, \varepsilon_1). \quad (4)$$

Обозначим вектор искомых параметров через $\theta = (\alpha, \beta)$, а вектор возмущений через $\varepsilon^0 = (\varepsilon_0^0, \varepsilon_1^0)^T$. Тогда, используя асимптотические разложения (3), (4), смещение ММ-оценки $\hat{\theta}$ можно представить как $\Delta\hat{\theta}_E = Q(\theta^0)\varepsilon + 1_m o(\|\varepsilon\|)$, где Q — (2×2) -матрица из соответствующих коэффициентов, 1_m — m -вектор, составленный из единиц, $\|\cdot\|$ — евклидова норма. Рассмотрим далее класс оценок вида $\hat{\theta}_c(\varepsilon) = \hat{\theta} - Q(\hat{\theta})\varepsilon$, где ε — некоторое значение из заданного интервала, которые будем называть оценками с компенсированным смещением. Тогда b -робастную оценку с компенсированным смещением можно определить как $\hat{\theta}_c^b(\varepsilon^{\min}, \varepsilon^{\max}) = \hat{\theta}_c(\varepsilon^*)$, где ε^* — решение следующей задачи оптимизации:

$$\max_{\varepsilon^0 \in \mathcal{E}} \left\| E\{\hat{\theta}_c(\varepsilon) - \theta^0 \mid \varepsilon^0\} \right\| \rightarrow \min_{\varepsilon \in \mathcal{E}}, \quad (5)$$

где θ^0 — неизвестное истинное значение для θ и $\varepsilon = [\varepsilon_0^{\min}, \varepsilon_0^{\max}] \times [\varepsilon_1^{\min}, \varepsilon_1^{\max}]$. Учитывая линейность по $\varepsilon_0, \varepsilon_1$ выражений (3), (4), доказано, что решение задачи (5) достигается на середине интервалов $[\varepsilon_0^{\min}, \varepsilon_0^{\max}]$, $[\varepsilon_1^{\min}, \varepsilon_1^{\max}]$, т.е. b -робастная оценка с компенсированным смещением вычисляется при средних уровнях искажений как

$$\hat{\theta}_c^b(\varepsilon^{\min}, \varepsilon^{\max}) = \hat{\theta}_c((\varepsilon^{\min} + \varepsilon^{\max})/2).$$

Асимптотические разложения, аналогичные выражениям (3), (4), были получены также для оценок максимального правдоподобия параметров ББМ и БЛМ [9, 10]. Это позволяет использовать результат теоремы 1 при построения на их основе b -робастных оценок с компенсированным смещением.

Робастное прогнозирование

Как было показано ранее в работе [11], оптимальный в смысле минимума среднеквадратичной ошибки байесовский прогноз для бета-иерархической модели с искажениями (1), (2) в случае известных уровней $\varepsilon_0, \varepsilon_1$ и параметров модели $\{n_i, \alpha_i^0, \beta_i^0\}$ определяется как

$$\hat{p}_\varepsilon^i(s; \varepsilon_0, \varepsilon_1) = \sum_{l=0}^n \omega_{sl}^i(\varepsilon_0, \varepsilon_1) (\alpha_i^0 + l) / (\alpha_i^0 + \beta_i^0 + n_i), \quad (6)$$

где весовые коэффициенты имеют вид

$$\omega_{sl}^i(\varepsilon_0, \varepsilon_1) = C_{n_i}^l W_{sl}^i(\varepsilon_0, \varepsilon_1) \alpha_i^{0[l+]} \beta_i^{0[(n_i-l)+]} / \sum_{j=0}^{n_i} (C_{n_i}^j W_{sj}^i(\varepsilon_0, \varepsilon_1) \alpha_i^{0[j+]} \beta_i^{0[(n_i-j)+]}),$$

$$W_{sj}^i = \sum_{l=\max(j,s)}^{\min(n_i, j+s)} C_j^{l-s} C_{n_i-j}^{l-j} \varepsilon_0^{l-j} (1-\varepsilon_0)^{n_i-l} \varepsilon_1^{l-s} (1-\varepsilon_1)^{j+s-l},$$

причем $y^{[z+1]} = \prod_{l=0}^{z-1} (y+l)$, $y \in R$, $z \in N$, а $s = \sum_{j=1}^{n_i} B_{ij}$ — наблюдаемое число успехов для объекта i . Рассмотрим теперь задачу построения минимаксного прогноза, обеспечивающего минимальное значение среднеквадратичной ошибки при "наихудших" значениях уровней искажений из заданных интервалов $\varepsilon_0 \in [\varepsilon_0^{min}, \varepsilon_0^{max}]$, $\varepsilon_1 \in [\varepsilon_1^{min}, \varepsilon_1^{max}]$.

Решение этой задачи будем искать в классе байесовских прогнозов $\hat{p}_\varepsilon^i(s; \varepsilon)$, определяемых выражением (6), параметром которых является вектор возмущающих воздействий $\varepsilon = (\varepsilon_0, \varepsilon_1)^T$. Тогда искомое значение параметра ε^* определяется из задачи оптимизации

$$\max_{\varepsilon^0 \in \mathcal{E}} r^2(\hat{p}_\varepsilon^i(s; \varepsilon) | \varepsilon^0) \rightarrow \min_{\varepsilon \in \mathcal{E}}, \quad (7)$$

где $r^2(\hat{p}_\varepsilon^i(s; \varepsilon) | \varepsilon^0)$ — среднеквадратичная ошибка прогноза $\hat{p}_\varepsilon^i(s; \varepsilon)$ при условии, что истинное значение уровней искажений равно ε^0 , а $\mathcal{E} = [\varepsilon_0^{min}, \varepsilon_0^{max}] \times [\varepsilon_1^{min}, \varepsilon_1^{max}]$. Решение задачи (7) получено ниже с использованием следующих вспомогательных результатов.

Утверждение 1. Среднеквадратичная ошибка некоторого произвольного прогноза $\hat{p}^i(s)$ выражается через байесовский прогноз $\hat{p}_\varepsilon^i(s; \varepsilon^0)$ и ряд распределения $\pi_s^{\varepsilon, i}(\varepsilon^0)$ как

$$r^2(\hat{p}^i) = \alpha_0^{i[2+1]} / (\alpha_0^i + \beta_0^i)^{[2+1]} + \sum_{l=0}^{n_i} (\hat{p}^i(s) - 2\hat{p}^i(l) \hat{p}_\varepsilon^i(l; \varepsilon^0)) \pi_s^{\varepsilon, i}(\varepsilon^0),$$

где $\{n_i, \alpha_i^0, \beta_i^0\}$ — истинные значения параметров модели и

$$\pi_s^{\varepsilon, i}(a, b, \varepsilon_0, \varepsilon_1) = \sum_{j=0}^{n_i} w_{sj}^j(\varepsilon_0, \varepsilon_1) \pi_j^{0, i}, \quad \pi_j^{0, i}(a, b) = C_{n_i}^j \frac{B(\alpha_i^0 + j, \beta_i^0 + n_i - j)}{B(\alpha_i^0, \beta_i^0)}.$$

Утверждение 2. Для прогноза $\hat{p}_\varepsilon^i(s; \varepsilon^*)$ справедливо следующее асимптотическое разложение среднеквадратичной ошибки по истинным значениям уровней искажений:

$$r^2(\hat{p}_\varepsilon^i(s; \varepsilon^*) | \varepsilon^0) = r_0^2 + \frac{n_i \beta_0^i}{(\alpha_0^i + \beta_0^i)(\alpha_0^i + \beta_0^i + n_i)^2} \varepsilon_0^0 + \frac{n_i \alpha_0^i}{(\alpha_0^i + \beta_0^i)(\alpha_0^i + \beta_0^i + n_i)^2} \varepsilon_1^0 + o(\varepsilon^0, \varepsilon^*).$$

При использовании этих утверждений доказано что при достаточно малых $\varepsilon_0^{max}, \varepsilon_1^{max}$ оптимальный в смысле минимакса среднеквадратичной ошибки прогноз $\hat{p}_\varepsilon^i(s; \varepsilon^*)$ вычисляется на верхней границе уровней искажений $\varepsilon^* = \varepsilon^{max}$. Полученный результат позволяет построить робастный прогноз для бета-иерархической модели, обеспечивающий минимальное значение среднеквадратичной ошибки при искажениях, известных с точностью до интервала.

Результаты компьютерных экспериментов

Для анализа эффективности предложенного метода b -робастного компенсированного оценивания параметров ББМ была проведена серия компьютерных экспериментов для $\alpha_0=0,5$; $\beta_0=9,5$; $n=10$; $k=1000$, причем генерировались 100 случайных выборок, на которые далее накладывались аддитивные стохастические искажения с уровнями от 0 до 0,05. На рис. 1 приводятся результаты b -робастного оценивания для трех способов компенсации, отличающихся значением ε^* . Как следует из рисунка, при $\varepsilon^* = (\varepsilon^{min} + \varepsilon^{max})/2$ область смещения имеет "компромиссное" расположение, симметричное относительно оси абсцисс и обеспечивающее наименьшее значение $\Delta\alpha$ в "наихудшем" случае. На рис. 2 приведены результаты компьютерного моделирования, подтверждающие этот теоретический результат. При этом процент экспериментальных точек, попадающих в асимптотическую область, колеблется от 77 до 90% для ММ-оценки и от 63 до 76% для МП-оценки

(соответствующие проценты отмечены числами на рис. 2). Это показывает, что предложенный метод b -робастного компенсированного оценивания является устойчивым к искажениям рассматриваемого типа.

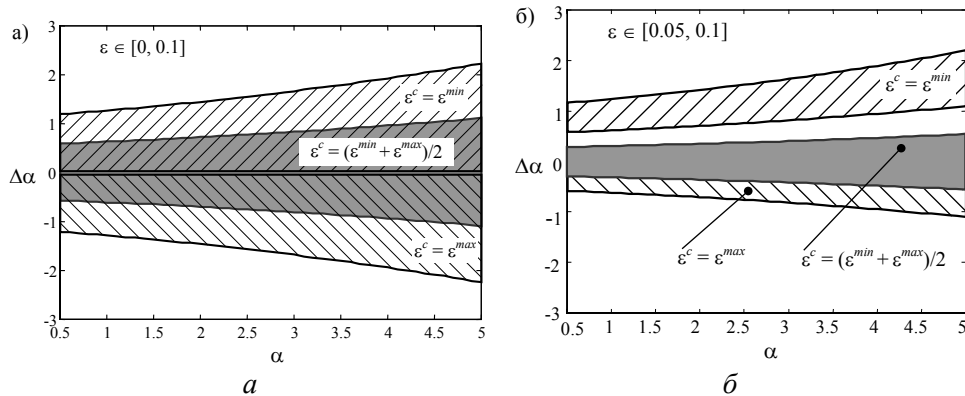





Рис. 1. Области смещения $\Delta\alpha$ для различных способов компенсации для b -робастной компенсированной ММ-оценки (а) и МП-оценки (б):  — настройка на нижнюю границу ($\varepsilon^c = \varepsilon^{\min}$);  — настройка на верхнюю границу ($\varepsilon^c = \varepsilon^{\max}$);  — b -робастная настройка ($\varepsilon^c = (\varepsilon^{\min} + \varepsilon^{\max}) / 2$)

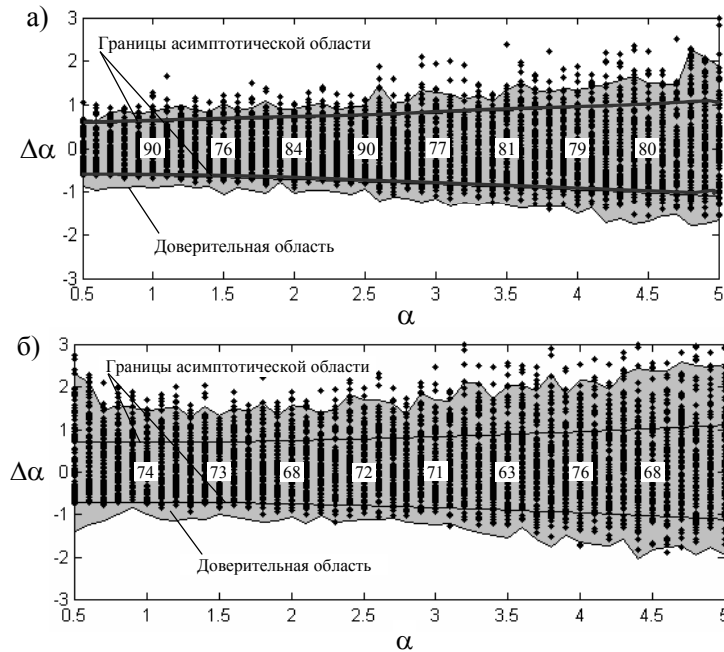




Рис. 2. Результаты компьютерного эксперимента для b -робастной компенсированной ММ-оценки (а) и МП-оценки (б):  — экспериментальная доверительная область;  — границы асимптотической области смещения

Заключение

Полученные результаты обобщают методы робастного оценивания и прогнозирования для бета-иерархических моделей группированных бинарных данных на случай вероятностей искажений, известных с точностью до интервала. При этом доказано, что оптимальные в смысле минимакса смещения оценки параметров вычисляются на середине интервалов, а прогноз, оптимальный в смысле минимакса риска, достигается при максимально возможных вероятностях искажений. Эффективность предложенных методов подтверждается результатами компьютерного моделирования.

MINIMAX ESTIMATION AND FORECASTING FOR CLUSTERED BINARY DATA WITH MISCLASSIFICATIONS

M.A. PASHKEVICH

Abstract

The paper proposes new minimax identification and forecasting techniques for the beta-mixed hierarchical models of the clustered binary data with misclassified observations. For the known misclassification probability intervals, it is proved that the b -robust bias-corrected estimator is calculated for the mean misclassification probabilities, while the optimal predictor that ensures the minimum risk of forecasting is reached on the upper bound of the intervals. The performance of the obtained theoretical results is verified via computer simulation.

Література

1. Diggle P.J., Heagerty P., Liang K.-Y., Zeger S.L. Analysis of Longitudinal Data. Oxford University Press, 2002, 379 p.
2. Jonson N.L., Kotz S., Kemp A.W. Univariate Discrete Distributions. Wiley-Interscience, New York, 1996.
3. Wilcox, R.R. // Journal of Educational Statistics. 1981. Vol. 6. P. 3–32.
4. Pearson E.S. // Biometrika, Vol. 17, 1925. P. 338–442.
5. Skellam J.G. // Journal of the Royal Statistical Society, Vol. B 10, 1948. P. 257–261.
6. Tripathi R.C., Gupta R.C., Gurland J. // Ann. Inst. Statist. Math. 1994. Vol. 46. P. 317–331.
7. Heckman J.J., Willis, R.J. // Journal of Political Economy, 1977, Vol. 85(11). P. 27–58.
8. Pfeifer P.E. // Journal of Interactive Marketing, 1998, Vol. 12 (2). P. 23–32.
9. Харин Ю.С., Пашкевич М.А. // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2003, № 1. С. 11–17.
10. Харин Ю.С., Пашкевич М.А. // Обзорение прикладной и промышленной математики. 2003. Т 10, Вып. 3. С. 246–247.
11. Pashkevich M., Dolgui A. // Applied Optimization, Kluwer Academic Publishers. 2004. Vol. 90. P. 55–70.