ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 28 МАТЕМАТИКИ Выпуск 2
2021

А. В. Колногоров, **А. Н. Лазутченко** (Великий Новгород, НовГУ)**. Задача о двуруком бандите с мультиномиальным распределением доходов.**

УДК 621.391.1 : 519.713 : 517.977.5

Резюме: Рассматривается задача о двуруком бандите, характеризуемая конечным множеством допустимых значений одношаговых доходов, описываемых мультиномиальным распределением. Получены рекуррентные уравнения для вычисления байесовских стратегии и риска. Минимаксные стратегия и риск ищутся как байесовские, соответствующие наихудшему априорному распределению.

Ключевые слова: задача о двуруком бандите, мультиномиальное распределение, минимаксный и байесовский подходы, основная теорема теории игр.

Двурукий бандит (см., например, [1]) с конечным множеством доходов $\{a_0,\dots,a_k\}$ — это управляемый случайный процесс ξ_n, $n=1,2,\dots,N$, значения которого зависят только от текущих выбранных действий $y_n\in\{1,2\}$ и характеризуются распределением $\Pr(\xi_n=a_j|y_n=\ell)=p_{\ell j}$, где $p_{\ell j}>0$, $j=0,\dots,k$, $p_{l0}+\dots+p_{lk}=1$, $\ell=1,2$. Обозначим $\mathbf{p}_\ell=(p_{\ell1},\dots,p_{\ell k})$. Значения $a_0,\dots,a_k$ предполагаются фиксированными, поэтому рассматриваемый двурукий бандит полностью характеризуется параметром $\theta=(\mathbf{p}_1,\mathbf{p}_2)$. Пусть к моменту времени n оба действия применены n_1 и n_2 раз соответственно ($n_1+n_2=n$), причем в ответ на применение ℓ-го действия получены $\mathbf{m}_\ell=(m_{\ell0},\dots,m_{\ell k})$ значений $a_0,\dots,a_k$ ($m_{\ell0}+\dots+m_{\ell k}=n_\ell$). Тогда распределение статистики $(\mathbf{m}_\ell,n_\ell)$ является мультиномиальным

$$\Pr(\mathbf{m}_\ell,n_\ell|\mathbf{p}_\ell)=Mu(\mathbf{m}_\ell,n_\ell,\mathbf{p}_\ell)=\frac{n_\ell!}{m_{\ell0}!\dots m_{\ell k}!}p_{\ell0}^{m_0}\cdot\ldots\cdot p_{\ell k}^{m_k}.$$

Обозначим через Θ множество значений параметра, через $\lambda(\theta)=\lambda(\mathbf{p}_1,\mathbf{p}_2)$ — априорную плотность распределения на Θ, через $E(\mathbf{p}_\ell)=a_0p_{\ell0}+\dots+a_kp_{\ell k}$ — математическое ожидание одношагового дохода за применение ℓ-ого действия. Пусть стратегия управления σ описывает выбор действия y_{n+1} в зависимости от текущей предыстории процесса $\mathbf{m}_1,n_1,\mathbf{m}_2,n_2$. Функция потерь $L_N(\sigma,\theta)=N\max\left(E(\mathbf{p}_1),E(\mathbf{p}_2)\right)-\mathbf{E}_{\sigma,\theta}\left(\sum_{n=1}^N\xi_n\right)$ характеризует математическое ожидание потерь полного дохода, вызванное неполнотой информации. Здесь $\mathbf{E}_{\sigma,\theta}$ означает знак математического ожидания при выбранных σ, θ. Байесовский и минимаксный риски определяются следующим образом:

$$R_N^B(\lambda)=\inf_\sigma\int_\theta L_N(\sigma,\theta)d\theta,\quad R_N^M(\Theta)=\inf_\sigma\sup_\theta L_N(\sigma,\theta).$$

Положим $x^+=\max(x,0)$ и обозначим через $\mathbf{e}_j=(e_{j0},\dots,e_{jk}),j=0,\dots,k$ единичные векторы, заданные координатами $e_{jj}=1$ и $e_{ji}=0$ при $j\neq i$, через $\mathbf{0}=(0,\dots,0)$ — $(k+1)$-мерный нулевой вектор. Положим $\overline{\ell}=3-\ell$ и будем считать, что $\{\mathbf{m}_1,n_1,\mathbf{m}_2,n_2\}$ означает, что порядок записи $\mathbf{m}_1,n_1$ и $\mathbf{m}_2,n_2$ в скобках

несуществен. Байесовский риск можно найти рекуррентно с помощью уравнения

$$R\{\mathbf{m}_1, n_1, \mathbf{m}_2, n_2\} = \min\left(R^{(1)}\{\mathbf{m}_1, n_1, \mathbf{m}_2, n_2\}, R^{(2)}\{\mathbf{m}_1, n_1, \mathbf{m}_2, n_2\}\right),$$

где $R^{(1)}\{\mathbf{m}_1, n_1, \mathbf{m}_2, n_2\} = R^{(2)}\{\mathbf{m}_1, n_1, \mathbf{m}_2, n_2\} = 0$ при $n = N$ и далее

$$R^{(\ell)}\{\mathbf{m}_\ell, n_\ell, \mathbf{m}_{\overline{\ell}}, n_{\overline{\ell}}\} = g^{(\ell)}(\mathbf{m}_1, n_1, \mathbf{m}_2, n_2)$$
$$+(n_\ell + 1)^{-1} \sum_{j=0}^{k} R\{\mathbf{m}_\ell + \mathbf{e}_j, n_\ell + 1, \mathbf{m}_{\overline{\ell}}, n_{\overline{\ell}}\}(m_{\ell j} + 1), \quad \ell = 1, 2$$

при $n < N$, где

$$g^{(\ell)}(\mathbf{m}_1, n_1, \mathbf{m}_2, n_2)$$
$$= \iint_\Theta \left(E(\mathbf{p}_{\overline{\ell}}) - E(\mathbf{p}_\ell)\right)^+ F(\mathbf{m}_1, n_1, \mathbf{m}_2, n_2, \theta)\lambda(\mathbf{p}_1, \mathbf{p}_2)d\mathbf{p}_1 d\mathbf{p}_2, \quad \ell = 1, 2.$$

Здесь $d\mathbf{p}_\ell = dp_{\ell 1} \dots dp_{\ell k}$, $F(\mathbf{m}_1, n_1, \mathbf{m}_2, n_2, \theta) = Mu(\mathbf{m}_1, n_1, \mathbf{p}_1)Mu(\mathbf{m}_2, n_2, \mathbf{p}_2)$. Байесовский риск определяется по формуле

$$R_N^B(\lambda) = R\{\mathbf{0}, 0, \mathbf{0}, 0\}.$$

Байесовская стратегия предписывает выбирать то действие, которому соответствует меньшая текущая величина $R^{(\ell)}\{\mathbf{m}_1, n_1, \mathbf{m}_2, n_2\}$; в случае их равенства выбор может быть произвольным. Минимаксный риск может быть найден с использованием основной теоремы теории игр как байесовский, соответствующий наихудшему априорному распределению, т. е. $R_N^M(\Theta) = R_N^B(\lambda^0) = \sup_\lambda R_N^B(\lambda)$. В случае бернуллиевского двурукого бандита аналогичный подход использован в [2].

Исследование выполнено при финансовой поддержке РФФИ, научный проект № 20-01-00062.

СПИСОК ЛИТЕРАТУРЫ

1. *Berry D. A.*, *Fristedt B.* Bandit Problems: Sequential Allocation of Experiments. London, New York: Chapman and Hall, 1985, 275 p.

2. *Колногоров А. В.* О минимаксном подходе к оптимальному целесообразному поведению в стационарных средах на конечном времени. — Изв. АН СССР, Техническая кибернетика, 1988, в. 6, с. 143–146. // *Kolnogorov A. V.* A minimax approach to optimal expedient behavior in stationary environments over finite time. — Sov. J. Comput. Syst. Sci., 1989. v. 27, is. 4, p. 33–35.

UDC 621.391.1 : 519.713 : 517.977.5

Kolnogorov A. V., ***Lazutchenko A. N.*** (Velikiy Novgorod, Yaroslav-the-Wise Novgorod State University). **Two-armed bandit problem with a multinomial distribution of incomes.**

Abstract: We consider the two-armed bandit problem characterized by a finite set of admissible values of one-step incomes which are described by a multinomial distribution. We obtain recursive equations for computing Bayesian strategy and Bayesian risk. Minimax strategy and minimax risk are searched for as Bayesian corresponding to the worst-case prior distribution.

Keywords: two-armed bandit problem, multinomial distribution, minimax and Bayesian approaches, main theorem of the game theory.