УДК 519.837.3

О ТЕОРЕТИКО-ИГРОВОМ ПОДХОДЕ К ЗАДАЧЕ О ДВУРУКОМ БАНДИТЕ

А.В. Колногоров

Новгородский государственный университет им. Ярослава Мудрого

Россия, 173003, Великий Новгород, Большая Санкт-Петербургская ул., 41

E-mail: Alexander.Kolnogorov@novsu.ru

Ключевые слова: гауссовский двурукий бандит, пакетная обработка, игра с природой, минимаксный и байесовский подходы, основная теорема теории игр.

Аннотация: Рассматривается гауссовский двурукий бандит, который возникает при оптимизации пакетной обработки данных, если для обработки имеются два альтернативных метода с различными априори неизвестными эффективностями. Требуется в процессе управления определить более эффективный метод и обеспечить его преимущественное применение. С позиций теоретико-игрового подхода задача интерпретируется как игра с природой. Цели управления могут ставиться как в минимаксной, так и в байесовской постановках, каждая из которых имеет свои достоинства и недостатки. Объединяет обе постановки основная теорема теории игр, согласно которой минимаксный риск равен байесовскому, вычисленному относительно наихудшего априорного распределения, а минимаксная стратегия совпадает с соответствующей байесовской. В статье характеризуется класс априорных распределений, к которому принадлежит наихудшее, и приводится рекуррентное уравнение, позволяющее вычислить байесовский риск относительно априорного распределения из этого класса. Минимаксные стратегия и риск найдены численными методами, как байесовские, соответствующие наихудшему априорному распределению

1. Введение

Рассматривается задача о двуруком бандите, т.е. об игральном автомате с двумя рукоятками, называемых также действиями (см., например, [1, 2]). Выбор каждого действия сопровождается случайным выигрышем (доходом), распределение которого зависит только от выбранного действия и не меняется в процессе игры, но не известно игроку. Принимая в процессе игры решение о выборе того или иного действия, игрок может использовать всю имеющуюся к текущему моменту времени информацию о выбранных действиях и полученных в ответ выигрышах, т.е. формировать стратегию игры. Естественной целью игрока является максимизация математического ожидания полного дохода, поэтому его стратегия состоит в том, чтобы определить в процессе игры действие, которому соответствует больший ожидаемый одношаговый доход и обеспечить его преимущественный выбор. Проблема известна также как задача о моделировании целесообразного поведения [3,4] и об адаптивном управлении в случайной среде [5,6]. Она имеет многочисленные применения в медицине, технике, информационных технологиях.

Ниже рассматривается задача о двуруком бандите с гауссовскими (нормальными) распределениями доходов. Такая постановка естественно возникает при оптимизации пакетной обработки, если для обработки доступны два альтернативных действия с различными априори неизвестными эффективностями [7]. Задача исследуется в минимаксной постановке и интерпретируется как игра с природой. Нахождение минимаксных стратегии и риска выполняется с использованием основной теоремы теории игр как байесовских, соответствующих наихудшему априорному распределению. Приводятся результаты численного определения минимаксных стратегии и риска.

2. Основные результаты

Гауссовский двурукий бандит – это управляемый случайный процесс ξ_n, $n = 1, 2, \ldots, N$, значения которого интерпретируются как доходы, зависят только от текущих выбираемых действий $\{y_n\}$ и характеризуются гауссовскими плотностями распределения $f_{D_\ell}(x|m_\ell) = (2\pi D_\ell)^{1/2} \exp(-(x-m_\ell)^2/(2D_\ell))$, если $y_n = \ell$, $\ell = 1, 2$. Здесь m_1, m_2 – априори неизвестные математические ожидания, а D_1, D_2 – известные дисперсии, горизонт управления N также предполагается известным. Рассматриваемый двурукий бандит может быть описан векторным параметром $\theta = (m_1, m_2)$.

Стратегия управления σ в момент времени $n+1$ определяет, вообще говоря, вероятностный выбор действия y_n в зависимости от всей известной предыстории, т.е. чисел n_1, n_2 – применений обоих действий ($n_1 + n_2 = n$) и X_1, X_2 – соответствующих текущих полных доходов.

Определим функцию потерь. Если параметр известен, то следует всегда применять действие, соответствующее максимальному значению из m_1, m_2, при этом математическое ожидание полного дохода равно $N \max(m_1, m_2)$. Если же применяется стратегия σ, то математическое ожидание полного дохода меньше максимально возможного на величину, называемую функцией потерь и равную

$$L_N(\sigma, \theta) = N \max(m_1, m_2) - \mathbf{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi_n \right). \tag{1}$$

Здесь $\mathbf{E}_{\sigma,\theta}$ обозначает математическое ожидание, вычисленное по мере, порождаемой стратегией σ и параметром θ. В качестве множества допустимых значений параметра выберем следующее

$$\Theta = \{\theta : m_1 = m + v, m_2 = m - v; |m| \leqslant a < \infty, |v| \leqslant C < \infty\}.$$

Данное множество параметров гарантирует ограниченность функции потерь величиной $2NC$. Предполагается, что a достаточно велико.

Данная задача управления может рассматриваться как игра с природой (см., например, [8]). В этом случае множеством стратегий лица, принимающего решения (ЛПР), является $\{\sigma\}$, множеством стратегий природы – Θ, формула (1) описывает платежную функцию игры в нормальной форме. В развернутой форме игра может быть представлена в виде дерева, вершины которого соответствуют выбираемым действиям и получаемым в ответ доходам. Поскольку в процессе игры ЛПР получает информацию в виде выборки случайных величин, то данная игра является статистической. Наконец, отметим важную особенность игр с природой: в них только один

участник, а именно ЛПР, заинтересован в результате игры; природа к этому результату безразлична.

Опишем возможные цели управления. Весьма распространенным является подход, когда исследуется предельное (при $N \to \infty$) поведение среднего значения платежной функции $N^{-1}L_N(\sigma^0, \theta)$ для некоторой стратегии σ^0. Типичным требованием является малость или равенство нулю этой предельной величины на всем множестве параметров, соотвествующая стратегия σ^0 является оптимальной с точки зрения этой цели. При этом обычно рассматриваются другие виды доходов, например, бернуллиевские, и могут накладываться ограничения на множество стратегий. Например, в [3, 4] в качестве стратегий рассматриваются конечные автоматы и автоматы с переменной структурой, в [6] – рекуррентные алгоритмы, в [9] – правило UCB, в [10] – стратегии, сравнивающие действия на начальном этапе, а затем применяющие только лучшее по результатам сравнения. Ограничения на стратегии вводятся и в [5].

Более традиционный с точки зрения теории игр подход предполагает байесовскую и минимаксную формулировку целей. Этот подход использован в [1, 2]. Наиболее популярным является байесовский подход. Пусть $\lambda(\theta)$ есть априорная плотность распределения на множестве параметров. Величина

$$R_N^B(\lambda) = \inf_{\{\sigma\}} \int_\Theta L_N(\sigma, \theta)\lambda(\theta)d\theta \tag{2}$$

называется байесовским риском, а соответствующая оптимальная стратегия σ^B – байесовской стратегией. Байесовский подход очень популярен, так как он позволяет при любом априорном распределении $\lambda(\theta)$ найти байесовские стратегию и риск численными методами, решая рекуррентное уравнение типа уравнения Беллмана. Это свойство является и недостатком байесовского подхода: решение рекуррентного уравнения зависит как от вида априорного распределения, так и от распределений доходов, т.е. обычно не является универсальным. Примером общего результата является асимптотическая (при $N \to \infty$) оценка порядка роста $L_N(\sigma, \theta)$ как $\ln(N)$, полученная в [11]. Еще один недостаток байесовского подхода – отсутствие ясных критериев для выбора априорного распределения $\lambda(\theta)$.

Другим классическим подходом является минимаксный. В этом случае величина

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta) \tag{3}$$

называется минимаксным риском, соответствующая оптимальная стратегия σ^M – минимаксной стратегией. Минимаксный подход не зависит от априорного распределения и является робастным, так как обеспечивает ограниченность функции потерь, именно, $L_N(\sigma^M, \theta) \leqslant R_N^M(\Theta)$ на всем множестве параметров Θ. Недостатком минимаксного подхода является отсутствие прямых методов нахождения минимаксных стратегии и риска. Совместить достоинства байесовского и минимаксного подходов позволяет основная теорема теории игр, согласно которой при широких предположениях, выполняющихся в нашем случае, справедливо равенство

$$R_N^M(\Theta) = R_N^B(\lambda^0) = \sup_{\{\lambda\}} R_N^B(\lambda), \tag{4}$$

т.е. минимаксный риск (3) равен байесовскому риску (2), вычисленному относительно наихудшего априорного распределения, на котором байесовский риск максимален; при этом минимаксная стратегия совпадает с соответствующей байесовской.

Важно, однако, понимать, что прямое применение равенства (4) для нахождения минимаксных стратегии и риска практически невозможно в силу высокой вычислительной сложности. Однако для гауссовского двурукого бандита при достаточно больших a оказывается возможным описать класс априорных распределений, к которому принадлежит наихудшее, и это дает возможность кардинально упростить задачу. Результаты можно сформулировать в виде теоремы.

Теорема 1. *Для гауссовского двурукого бандита плотность наихудшего априорного распределения при $a \to \infty$ может быть выбрана в виде $\lambda_a(\theta) = \kappa_a(m)\rho(v)$, где $\kappa_a(m) = (2a)^{-1}$ – плотность равномерного распределения на множестве $|m| \leqslant 2a$, $\rho(v)$ – некоторая плотность распределения на множестве $|v| \leqslant C$. Соответствующий байесовский риск имеет предел*

$$R_N^B(\rho(v)) = \lim_{a\to\infty} R_N^B(\lambda_a(\theta)).$$

Обозначим $n_1' = n_1/D_1$, $n_2' = n_2/D_2$, $n' = n_1' + n_2'$, $U = (X_1 n_2 - X_2 n_1)/n'$ и рассмотрим рекуррентное уравнение

$$R(U, n_1, n_2) = \min(R^{(1)}(U, n_1, n_2), R^{(2)}(U, n_1, n_2)), \tag{5}$$

где $R^{(1)}(U, n_1, n_2) = R^{(2)}(U, n_1, n_2) = 0$ при $n_1 + n_2 = N$ и далее

$$\begin{aligned} R^{(1)}(U, n_1, n_2) &= g^{(1)}(U, n_1, n_2) + R(U, n_1 + 1, n_2) * f_{D_1 n_2^2 n'^{-1}(n' + D_1^{-1})^{-1}}(U), \\ R^{(2)}(U, n_1, n_2) &= g^{(2)}(U, n_1, n_2) + R(U, n_1, n_2 + 1) * f_{D_2 n_1^2 n'^{-1}(n' + D_2^{-1})^{-1}}(U) \end{aligned} \tag{6}$$

при $2 \leqslant n_1 + n_2 < N$, $n_1 \geqslant 1$, $n_2 \geqslant 1$. Здесь '$*$' означает свертку функций,

$$\begin{aligned} g^{(1)}(U, n_1, n_2) &= \int_{-C}^{0} 2|v| g(v; U, n_1, n_2)\rho(v)dv, \\ g^{(2)}(U, n_1, n_2) &= \int_{0}^{C} 2v g(v; U, n_1, n_2)\rho(v)dv, \\ g(v; U, n_1, n_2) &= \exp\left(\frac{2Uv}{D_1 D_2} - \frac{2v^2 n_1' n_2'}{n'}\right). \end{aligned} \tag{7}$$

Байесовская стратегия при $n \leqslant 2$ применяет действия по очереди. При $n > 2$ она в момент времени $n+1$ применяется то действие, которому соответствует меньшая величина $R^{(\ell)}(U, n_1, n_2)$. Байесовский риск (2) вычисляется по формуле

$$R_N^B(\rho(v)) = \int_{-C}^{C} 2|v|\rho(v)dv + \int_{-\infty}^{\infty} f_{D_1^2 D_2^2 (D_1+D_2)^{-1}}(U) R(U, 1, 1)dU, \tag{8}$$

где первое слагаемое в правой части описывает ожидаемые потери на первых двух шагах, когда действия применяются по очереди.

Поскольку оптимальная стратегия в начале управления применяет действия по очереди, то из теоремы следует, что дисперсии D_1, D_2 могут быть оценены с требуемой точностью на начальном этапе, если размеры пакетов достаточно велики. Поскольку байесовские стратегия и риск мало меняются даже при значительном изменении дисперсий (например, в пределах 5–10%), то вместо самих дисперсий в формулах (5)–(8) можно использовать их оценки. Это означает, что требование известности дисперсий D_1, D_2 можно снять, если размеры пакетов достаточно велики.

Вычисление байесовского риска выполнялось с использованием формул (5)–(8) при $N = 50$ в предположении, что плотность $\rho(v)$ сосредоточена в двух точках $v = -d_1 N^{-1/2}$ и $v = d_2 N^{-1/2}$ с вероятностями ρ, $1-\rho$. В соответствии с (4) наихудшее априорное распределение соответствует максимуму приведенного байесовского риска $r_N(\rho(v)) = N^{-1/2} R_N^B(\rho(v))$. При $D_1 = 1$, $D_2 = 0{,}75$ численными методами были найдены $d_1 \approx 1{,}56$, $d_2 \approx 1{,}49$, $\rho \approx 0{,}52$, $r_N(\rho(v)) \approx 0{,}61$.

Далее, для найденной стратегии были вычислены приведенные потери при $v = dN^{-1/2}$ в диапазоне $-15 \leqslant d \leqslant 15$ с шагом 0,2. Вычисления можно выполнять с помощью симуляций Монте-Карло или с использованием формул, приведенных в [7]. Оказалось, что максимальные приведенные потери приблизительно равны 0,61 и достигаются как раз при $d \approx -1{,}56$ и $d \approx 1{,}49$. Таким образом, найденная байесовская стратегия является уравнивающей и, следовательно, минимаксной, если $-15 \leqslant d \leqslant 15$.

3. Заключение

Работа выполнена при поддержке Российского фонда фундаментальных исследований (18-29-16223-мк).

Список литературы

1. Berry D.A., Fristedt B. Bandit Problems: Sequential Allocation of Experiments. London, New York: Chapman and Hall, 1985. 275 p.
2. Пресман Э.Л., Сонин И.М. Последовательное управление по неполным данным. М.: Наука, 1982. 256 с.
3. Цетлин М.Л. Исследования по теории автоматов и моделированию биологических систем. М.: Наука, 1969. 316 с.
4. Варшавский В.И. Коллективное поведение автоматов. М.: Наука, 1973. 408 с.
5. Срагович В.Г. Адаптивное управление. М.: Наука, 1981. 384 с.
6. Назин А.В., Позняк А.С. Адаптивный выбор вариантов. М.: Наука, 1986. 288 с.
7. Колногоров А.В. Гауссовский двурукий бандит и оптимизация групповой обработки данных // Пробл. передачи информ. 2018. Т. 54. № 1. С. 93-111.
8. Боровков А.А. Математическая статистика. Дополнительные главы: Учебное пособие для вузов. М.: Наука. Главная редакция физико-математической литературы, 1984. 144 с.
9. Смирнов Д.С., Громова Е.В. Модель принятия решений при наличии экспертов как модифицированная задача о многоруком бандите // МТИП. 2017. Т. 9. № 4. С. 69–87.
10. Lai T.L., Levin B., Robbins H., Siegmund D. Sequential Medical Trials (Stopping Rules/Asymptotic Optimality) // Proc. Nati. Acad. Sci. USA. 1980. Vol. 77, No. 6. P. 3135-3138.
11. Lai T.L., Robbins H. Asymptotically Efficient Adaptive Allocation Rules // Advances in Applied Mathematics. 1985. Vol. 6. P. 4-22.