

МОДЕЛЬ АДАПТИВНОГО УПРАВЛЕНИЯ КОНЕЧНОЙ ЦЕПЬЮ МАРКОВА СО СЛАБО ДИСКОНТИРУЕМЫМИ ДОХОДАМИ

А.В.Колногоров, Р.С.Федорук

Институт электронных и информационных систем НовГУ, Alexander.Kolnogorov@novsu.ru

Для марковских цепей со слабо дисконтируемыми доходами предложена адаптивная модель управления, оценивающая неизвестные параметры и осуществляющая затем оптимальное управление в соответствии с алгоритмом Ховарда. Дана математическая оценка оптимального времени проведения испытаний для частного случая системы, когда принимаемое решение не зависит от состояния системы.

Ключевые слова: управляемые цепи Маркова, адаптивное управление

Given Markov chain with slightly discounted incomes, an adaptive model which estimates the unknown parameters at the beginning of the control and then implements optimal control according to Howard's algorithm is proposed. The estimate of optimal test period for the case of the state-independent solutions is given.

Keywords: controlled Markov chains, adaptive control

Описание системы

Дадим определение конечной цепи Маркова с дисконтируемыми доходами. Рассмотрим систему, у которой пространство состояний S содержит конечное число элементов. Пусть S совпадает с множеством целых чисел $S = \{1, 2, \dots, N\}$. Каждому состоянию $i \in S$ соответствует конечное множество K_i решений, элементы которого обозначим $k = 1, 2, \dots, K_i$. Пространством политик K назовем прямое произведение множеств решений, т. е. $K = K_1 \times K_2 \times \dots \times K_N$. Рассматривается задача принятия последовательных решений, состоящая в выборе решений при наблюдении текущих состояний в моменты $n = 0, 1, 2, \dots$

Пусть F — множество всевозможных векторов со значениями в пространстве политик K и $f_1, f_2, \dots, f_n, \dots$ — элементы из F . Тогда стратегия π определяется, как последовательность $\pi = (f_1, f_2, \dots, f_n, \dots)$, где f_n — вектор, i -й элемент которого, обозначаемый $f_n(i)$, является решением, принимаемым в состоянии $i \in S$ в момент n . Стратегия $(f, f, \dots, f, \dots)$ обозначается f^∞ , где $f \in F$, и называется стационарной. Стационарная стратегия f^∞ состоит из политик, не зависящих от времени.

Если система находится в состоянии $i \in S$ и принимается решение $k \in K_i$, то система получает доход r_i^k и ее состояние в следующий момент времени определяется вероятностным законом p_{ij}^k ($j \in S$), где p_{ij}^k — вероятность того, что система из состояния i при выборе решения k попадает в состояние j . Предполагается, что доход r_i^k ограничен при всех $i \in S$ и $k \in K_i$. Допустимо использование случайных доходов, зависящих только от решений, выбираемых в данном состоянии, при этом следует потребовать ограниченности их математических ожиданий, которые также обозначим r_i^k . Кроме того, $\sum_{j \in S} p_{ij}^k = 1$, $p_{ij}^k \geq 0$ при $i, j \in S$, $k \in K_i$.

Пусть β , $0 < \beta < 1$, — коэффициент переоценки (дисконтирования) доходов. Смысл его состоит в том, что единица дохода через время n будет стоить β^n единиц. Введение коэффициента переоценки с математической точки зрения ведет к ограниченности суммарного дохода. Обозначим через ξ_n случайный доход, получаемый системой в момент времени n , через

$$V(\pi, i, \beta) = M_{\pi, i} \sum_{n=1}^{\infty} \beta^n \xi_n \quad \text{— математическое ожидание}$$

полного дохода системы, если ее начальное состояние равно i , а применяемая стратегия π . Цель управления состоит в определении такой стратегии π , которая максимизирует величину $V(\pi, i, \beta)$.

Ховардом (см., напр., [1]) доказано, что оптимальная стратегия π^* в цепях Маркова с дисконтируемыми доходами всегда является стационарной, причем максимизация $V(\pi, i, \beta)$ обеспечивается одной и той же стратегией π^* при всех i , а также предложен и доказан алгоритм нахождения этой стратегии. Ниже рассматриваются цепи Маркова со слабо дисконтируемыми доходами, т.е. при $\beta \rightarrow 1$. В этом случае зависимость от начального состояния исчезает, а именно $\lim_{\beta \rightarrow 1} \frac{V(\pi^*, i, \beta)}{V(\pi^*, \beta)} = 1$ при всех $i \in S$, причем при β , достаточно близких к 1, стратегия π^* одинакова.

Приведем пример. Предположим, что некоторая компания занимается перевозкой грузов из одного города в другой — города являются состояниями системы. У компании есть возможность перевезти груз несколькими способами: автомобилем, поездом, по воде или по воздуху. В разных городах доступный набор способов доставки разный — это решения. Вероятностные законы определяют вероятность перевозки из одного города в другой, а доходы — это доход, получаемый компанией от перевозки груза из одного города в другой в зависимости

от вида транспорта. В разные времена года эти параметры могут различаться, и для определения оптимальной стратегии можно использовать рассмотренную модель.

Адаптивная модель управления системой в случае неизвестных параметров

Рассматриваемый далее подход предложен в [2]. Будем предполагать известными множество состояний системы и возможные решения в каждом из состояний, а неизвестными — переходные вероятности p_{ij}^k . Если доходы r_i^k являются детерминированными, то их также можно считать известными, так как для их определения достаточно один раз выбрать решение k в состоянии i . Если же они являются случайными, то они тоже считаются неизвестными.

Предложенная модель сначала оценивает неизвестные параметры системы, а затем осуществляет оптимальное управление в соответствии с алгоритмом Ховарда на основе этих оценок. Таким образом, управление подстраивается под любую конечную цепь Маркова, т. е. является адаптивным управлением. Отметим, что адаптация предполагает наличие двух процессов: идентификации параметров и оптимального управления, причем эти два процесса конкурируют друг с другом. Если оценивание выполняется долго, то оценки будут близки к точным, однако из-за обесценивания доходов вследствие дисконтирования не будет возможности получить достаточно большой полный доход. Если же оценивание выполняется не очень точно, но быстро, то дисконтирование незначительное, однако велика вероятность ошибки при определении оптимальной стратегии. Поэтому важным является условие слабой дисконтируемости доходов: оно позволяет посвятить достаточно времени получению близких к точным оценок параметров системы.

В процессе функционирования системы выполняется подсчет переходов из состояния в состояние при применении различных решений, а также полученных доходов. Обозначим через $V_{ij}^k(n)$ наблюдаемое к моменту времени n количество переходов системы из состояния i в состояние j при принятии в нем решения k , через $V_i^k(n) = \sum_{j \in S} V_{ij}^k(n)$ — наблюдаемое к моменту времени n количество принятий решения k в состоянии i . Очевидно, $V_{ij}^k(1) = 0$ при всех i, j, k . Если система в момент времени n находится в состоянии i , выполним сравнение $V_i^k(n)$ при всех k . Выбрать следует решение, которому соответствует минимальное значение $V_i^k(n)$, а если таких значений несколько, то любое из них (например, с наименьшим номером k или равновероятно). Текущие оценки переходных вероятностей следует выполнять по формулам $\hat{p}_{ij}^k(n) = \frac{V_{ij}^k(n)}{V_i^k(n)}$ при

$V_i^k(n) > 0$. Если, кроме того, полный доход, полученный в состоянии i в ответ на выбор решения k , равен $X_i^k(n)$, то оценкой математического ожидания дохода будет $\hat{r}_i^k(n) = \frac{X_i^k(n)}{V_i^k(n)}$.

Пусть $p_{ij}^k \geq \delta > 0$ при всех i, j, k . В этом случае система на каждом шаге попадает в каждое состояние с вероятностью не меньшей δ . Тогда $M \sum_{k=1}^{K_i} V_i^k(n) \geq \delta n$ и, следовательно, для любого $\varepsilon > 0$ можно указать такое n , что выполнится неравенство

$$\Pr \left\{ V_i^k > \frac{\delta n}{2K_i} \right\} > 1 - \varepsilon.$$

Так как оценки состоятельны, то в этом случае для любых $\varepsilon_1, \varepsilon > 0$ можно указать такое n , что выполнится неравенство

$$\Pr \left\{ \max_{i,j,k} \left| \hat{p}_{ij}^k(n) - p_{ij}^k(n) \right|, \left| r_i^k(n) - r_i^k \right| < \varepsilon_1 \right\} > 1 - \varepsilon. \quad (1)$$

Модель управления состоит в том, что сначала на отрезке времени длины N выполняются подсчеты величин $V_{ij}^k(n)$, $X_i^k(n)$, затем в момент времени N вычисляются оценки параметров $\hat{p}_{ij}^k(N)$, $\hat{r}_i^k(N)$, по ним с помощью алгоритма Ховарда определяется оптимальная стационарная стратегия, которая и будет затем применяться. Обозначим эту стратегию $\pi_1^*(N)$. Математическое ожидание полного дохода при применении этой стратегии, если начальное состояние равно i , обозначим через $V(\pi_1^*(N), i, \beta)$.

Теорема 1. Для любого $\varepsilon > 0$ можно указать такое N , что выполнится предельное неравенство

$$\lim_{\beta \rightarrow 1} \frac{V(\pi_1^*(N), i, \beta)}{V(\pi^*, \beta)} \geq 1 - \varepsilon. \quad (2)$$

Доказательство. Множество стратегий цепи Маркова является конечным. При этом алгоритм Ховарда основан на сравнении выражений, которые непрерывно зависят от параметров цепи. Поэтому из (1) следует, что для достаточно большого N вероятность ошибочного определения лучшей стратегии $P_{err}(N)$ может быть сделана сколь угодно малой. Как отмечено выше, при β , близких к 1, оптимальная стратегия одинакова, а зависимость полного ожидаемого дохода от начального состояния исчезает, т. е. $V(\pi^*, i, \beta) = V(\pi^*, \beta)(1 + \varepsilon(\beta))$, где $\varepsilon(\beta) \rightarrow 0$ при $\beta \rightarrow 1$. Поэтому

$$V(\pi_1^*(N), i, \beta) \geq (1 - P_{err}(N)) \beta^N V(\pi^*, \beta)(1 + \varepsilon(\beta)).$$

Выполняя предельный переход $\beta \rightarrow 1$ при фиксированном N , получаем

$$\lim_{\beta \rightarrow 1} \frac{V(\pi_1^*(N), i, \beta)}{V(\pi^*, \beta)} \geq 1 - P_{err}(N),$$

где вероятность $P_{err}(N)$ может быть сделана сколь угодно малой для достаточно больших N , что приведет к выполнению условия (2). Теорема доказана.

Более сложной является задача оптимального выбора N при заданном значении β , близком к 1. Эта задача решается точно в следующем разделе для цепей Маркова с одним состоянием. Для цепей общего вида она пока решается с помощью моделирования.

Оценка оптимального времени проведения испытаний в случае, если принимаемое решение не зависит от состояния системы

В данном разделе развиваются методы оптимального управления для систем с недисконтируемыми доходами на конечном отрезке времени, предложенные в [3]. Рассматривается система S из одного состояния s_0 , в котором возможно принятие двух решений $k = 1$ и $k = 2$. При этом система получает в зависимости от принятого решения случайный доход r^k , дисконтируемый на каждом шаге на коэффициент β , близкий к 1. В ситуации случайных доходов требуется затратить определенное время на оценку параметров системы. Обозначим время проведения испытаний системы $2N$. Пусть X^1 и X^2 — доходы системы, полученные от поочередного применения решений $k = 1$ и $k = 2$ на протяжении времени $2N$, каждое по N раз.

Будем считать, что доход системы имеет нормальное распределение. Математическое ожидание доходов r^k на шаге n и дисперсия соответствуют равенствам $M(r_n^k) = m^k$, $D(r_n^k) = 1$. Пусть система всегда начинает функционировать начиная с решения $k = 1$ и при этом $m^1 > m^2$, тогда

$$M(X^1) = m^1 \frac{1-\beta^{2N}}{1-\beta^2}, \quad M(X^2) = \beta m^2 \frac{1-\beta^{2N}}{1-\beta^2}. \quad (3)$$

Сравнение величин X^1 и X^2/β покажет, какое из решений при их применении обеспечивает получение максимального дохода системы во время проведения испытаний. На этом основании будет делаться вывод о том, какое из решений выбрать для управления системой на оставшемся промежутке времени ее функционирования.

Обозначим потери на первом и втором этапах соответственно $L_1(\beta, N, m_1, m_2)$ и $L_2(\beta, N, m_1, m_2)$, тогда целевой функцией при оценке оптимального времени проведения испытаний является $R(\beta) = \min_N \max_{m_1, m_2} (\max_{m_1, m_2} L_1(\beta, N, m_1, m_2), \max_{m_1, m_2} L_2(\beta, N, m_1, m_2))$.

Теорема 2. Оптимальные значения $N(\beta)$ и $R(\beta)$ при β , близких к 1, имеют вид

$$N(\beta) = \left(\frac{c\sqrt{2}}{\Delta M(1-\beta)} \right)^{\frac{2}{3}}, \quad R(\beta) = \left(\frac{c\sqrt{2\Delta M}}{1-\beta} \right)^{\frac{2}{3}}, \quad (4)$$

где $c \approx 0,17$.

Доказательство. Дадим оценку потерям дохода на первом и втором этапах. Пусть $(m^1 - m^2) = \Delta m$, $0 < \Delta m < \Delta M$. При этом потери возможны на половине первого этапа длины $2N$, т.е. за N шагов. Математическое ожидание потерь на первом этапе

$$L_1(\beta, N, m_1, m_2) = P_{err}^1 \Delta m (\beta^0 + \beta^1 + \dots + \beta^{2N-1}) = P_{err}^1 \Delta m \frac{1-\beta^{2N}}{1-\beta} = \frac{\Delta m}{2} \frac{1-\beta^{2N}}{1-\beta}, \quad (5)$$

где $P_{err}^1 = \frac{1}{2}$ — вероятность того, что применяемое на данном шаге решение не является оптимальным. Математическое ожидание потерь на втором этапе

$$L_2(\beta, N, m_1, m_2) = P_{err}^2 \Delta m (\beta^{2N} + \beta^{2N+1} + \dots) = P_{err}^2 \Delta m \frac{\beta^{2N}}{1-\beta}, \quad (6)$$

где P_{err}^2 — вероятность того, что по итогам испытаний на первом этапе была принята неоптимальная стратегия. Оценим вероятность ошибки:

$$P_{err}^2 = P\left(X^1 - \frac{X^2}{\beta} < 0\right) = P\left(X < -\Delta m \frac{1-\beta^{2N}}{1-\beta^2}\right), \quad (7)$$

где $X = \left(X^1 - m^1 \frac{1-\beta^{2N}}{1-\beta^2}\right) - \left(\frac{X^2}{\beta} - m^2 \frac{1-\beta^{2N}}{1-\beta^2}\right)$. С учетом (3) нетрудно проверить, что

$$M(X) = M\left(X^1 - m^1 \frac{1-\beta^{2N}}{1-\beta^2}\right) - M\left(\frac{X^2}{\beta} - m^2 \frac{1-\beta^{2N}}{1-\beta^2}\right) = 0,$$

$$D(X) = D(X^1) + D(X^2) = \beta^0 + \beta^2 + \dots + \beta^{4N-2} = \frac{1-\beta^{4N}}{1-\beta^2}.$$

Поэтому для (7) справедливо равенство

$$P_{err}^2 = \Phi\left(-\Delta m \sqrt{\frac{1-\beta^{2N}}{(1+\beta^{2N})(1-\beta^2)}}\right),$$

где $\Phi()$ — функция стандартного нормального распределения. С учетом (6), (7)

$$L_2(\beta, N, m_1, m_2) = \Delta m \frac{\beta^{2N}}{1-\beta} \Phi\left(-\Delta m \sqrt{\frac{1-\beta^{2N}}{(1+\beta^{2N})(1-\beta^2)}}\right). \quad (8)$$

Уточним оценки (5) и (8). Положим $\beta = 1 - \frac{1}{\tau}$,

где τ можно интерпретировать как эффективное время работы системы. Оценим (5) и (8) при больших τ . Для этого найдем предел выражений $\frac{1-\beta^{2N}}{1-\beta}$ и

β^{2N} при $\frac{N}{\tau} \rightarrow 0$. Обозначим $\frac{N}{\tau}$ как α , тогда

$$\frac{1-\beta^{2N}}{1-\beta} = \tau \left(1 - \left(1 - \frac{1}{\tau}\right)^{2N}\right) = \frac{N}{\alpha} \left(1 - \left(1 - \frac{\alpha}{N}\right)^{\frac{N}{\alpha}}\right)^{2\alpha} = \frac{N(1-e^{-2\alpha})}{\alpha}.$$

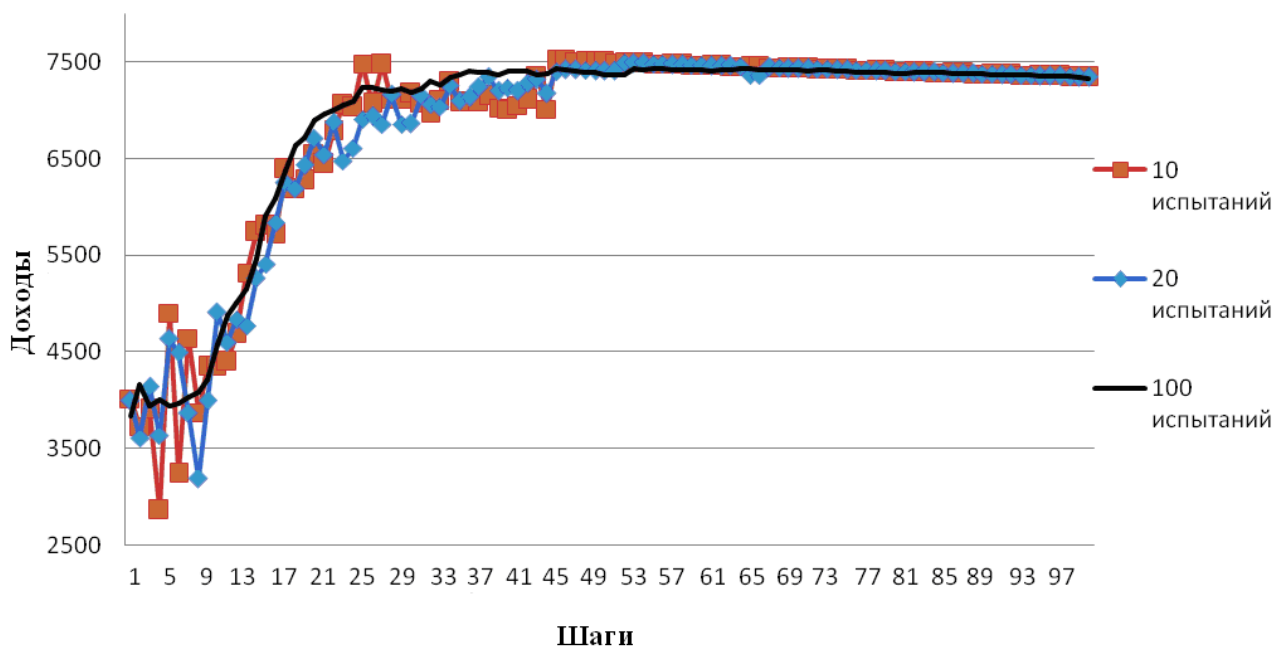
При нахождении первого предела воспользуемся правилом Лопиталя:

$$\lim_{\alpha \rightarrow 0} \frac{N(1-e^{-2\alpha})}{\alpha} = N \lim_{\alpha \rightarrow 0} \frac{1-e^{-2\alpha}}{\alpha} = N \lim_{\alpha \rightarrow 0} 2e^{-2\alpha} = 2N.$$

Второй предел

$$\lim_{\alpha \rightarrow 0} \beta^{2N} = \lim_{\alpha \rightarrow 0} \left(1 - \frac{\alpha}{N}\right)^{\frac{N}{\alpha} 2\alpha} = \lim_{\alpha \rightarrow 0} \frac{1}{e^{2\alpha}} = 1.$$

Обозначим $\Delta m \sqrt{\frac{N}{2}} = a$. Тогда



Изменение доходов в зависимости от времени проведения испытаний на примере матрицы 3×3 при различном количестве испытаний

$$L_1(\beta, N) = \max_{m_1, m_2} L_1(\beta, N, m_1, m_2) = \Delta MN,$$

$$L_2(\beta, N) = \max_{m_1, m_2} L_2(\beta, N, m_1, m_2) = c\tau\sqrt{\frac{2}{N}},$$

где $c = \max_{a>0} a\Phi(-a) \approx 0,17$. Легко видеть, что

$L_1(\beta, N)$, $L_2(\beta, N)$ являются соответственно возрастающей и убывающей функциями N . Поэтому $\min_N \max(L_1(\beta, N), L_2(\beta, N))$ достигается при $L_1(\beta, N) = L_2(\beta, N)$. Соответствующие $N(\beta)$ и $R(\beta)$ даются формулой (4). Теорема доказана.

Результаты моделирования для цепей Маркова общего вида

В общем случае оптимальное время обучения может определяться посредством математического моделирования, выполняемого с помощью специально подготовленной программы [4]. По результатам моделирования оказывается, что до определенного этапа доход действительно увеличивается с увеличением точности оценки параметров системы и, соответственно, с увеличением точности нахождения оптимальной стратегии. Однако после определенного этапа доход начинает снижаться в связи с тем, что точность нахождения параметров достаточно высока для нахождения оптимальной стационарной стратегии, но происходит большая переоценка доходов.

На приведенном графике максимальные значения достигаются для 5 испытаний — на 54 шаге (доход 7483), для 20 испытаний — на 52 шаге (доход 7489), для 100 испытаний — на 64 шаге (доход 7428).

Заключение

В работе предложен метод определения неизвестных параметров функционирующей системы. Компьютерное моделирование на основании описанного метода показало, что оценить параметры функционирующей системы, а также определить оптимальное время проведения испытаний системы для оценки параметров возможно. В работе также дана математическая оценка данного времени для частного случая системы, когда принимаемое решение не зависит от состояния системы.

1. Майн Х., Осаки С. Марковские процессы принятия решений. М.: Наука, 1977. 176 с.
2. Колногоров А.В., Федорук Р.С. // Вестник НовГУ. Сер.: Техн. науки. 2009. №50. С.29-30.
3. Колногоров А.В. // Проблемы передачи информации. 2000. Т.36. №4. С.117-127.
4. Федорук Р.С. Программа для ЭВМ «Расчет оптимального времени испытаний в цепи Маркова». Свидетельство о государственной регистрации программы для ЭВМ №2010614952 от 29.07.2010.