УДК 519.865

А.В.Колногоров, Р.С.Федорук

МОДЕЛЬ УПРАВЛЕНИЯ КОНЕЧНОЙ ЦЕПЬЮ МАРКОВА С ДИСКОНТИРУЕМЫМИ ДОХОДАМИ В СЛУЧАЕ НЕИЗВЕСТНЫХ ПАРАМЕТРОВ

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

The control strategy for systems described by finite Markov chains with discounted incomes in case of unknown parameters is offered. This strategy makes consistent estimates of parameters at the beginning of control and then uses the optimal strategy corresponding to these estimates.

Ключевые слова: цепи Маркова с дисконтируемыми доходами, стратегия управления

Введение

Цепи Маркова с доходами являются удобным средством математического описания систем, соответствующих процессам окружающего мира. Задача данной работы заключается в описании модели управления такой системой в случае постоянных, но неизвестных до начала управления параметров. Суть модели состоит в том, что в начале функционирования системы делаются оценки параметров, а затем применяется оптимальная для полученных оценок стратегия управления.

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 $S = \{1, 2, \ldots, N\}$. Каждому состоянию $i \in S$ соответствует конечное множество K_i решений, элементы которого обозначим $k = 1, 2, \ldots, K_i$. Пространством политик K назовем прямое произведение множеств решений: $K = K_1 \times K_2 \times \ldots \times K_N$. Рассматривается задача принятия последовательных решений, состоящая в выборе решений при наблюдении текущих состояний в моменты $n = 0, 1, 2,\ldots$

Пусть F — множество всевозможных векторов со значениями в пространстве политик K и $f_1, f_2, \ldots, f_n, \ldots$ — элементы из F. Тогда стратегия π определяется, как последовательность $\pi = (f_1, f_2, \ldots, f_n, \ldots)$, где f_n — вектор, i-й элемент которого, обозначаемый $f_n(i)$, является решением, принимаемым в состоянии $i \in S$ в момент n. Стратегия $(f,f,\ldots,f,\ldots)$ 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

1) система получает доход r_i^k,

2) 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 $k \in K_i$. Кроме того,

$$\sum_{j \in S} p_{ij}^k = 1,\ p_{ij}^k \geq 0 \text{ при } i,\ j \in S,\ k \in K_i.$$

Рассмотрим процесс с переоценкой доходов. Пусть β, $0 \leq \beta < 1$, — коэффициент переоценки (дисконтирования). Смысл его состоит в том, что единица дохода через время n (например, n дней) будет стоить β^n единиц. Введение коэффициента переоценки с математической точки зрения ведет к ограниченности суммарного среднего дохода. Обозначим через ξ_n случайный доход, получаемый системой в момент времени n, а через $r(\pi, i, \beta) = M_{\pi,i}\left(\sum_{n=1}^{\infty} \beta^n \xi_n\right)$ — математическое ожидание полного дохода системы с начальным состоянием i и применяемой стратегией π. Цель управления состоит в определении такой стратегии π, которая максимизирует величину $r(\pi, i, \beta)$.

Ховардом [1] доказано, что оптимальная стратегия π^* в цепях Маркова с дисконтируемыми доходами

всегда является стационарной, причем максимизация $r(\pi,i,\beta)$ обеспечивается одной и той же стратегией π^* при всех *i*, а также предложен и доказан алгоритм нахождения этой стратегии. При $\beta \to 1$ зависимость от начального состояния исчезает, именно $\lim\limits_{\beta\to 1}\dfrac{r(\pi^*,i,\beta)}{r(\pi^*,\beta)}=1$ при всех $i\in S$, причем найдется такое β_0, что при всех $\beta \ge \beta_0$ стратегия π^* одинакова [2,3].

Метод определения неизвестных параметров

Будем предполагать известными множество состояний системы и возможные решения в каждом из состояний. Если доходы r_i^k являются детерминированными, как в рассматриваемом случае, 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 необходимо произвести их оценку, аналогично тому, как далее делается оценка вероятностей p_{ij}^k.

В процессе функционирования системы выполняется подсчет переходов из состояния в состояние при применении различных решений. 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*, а через $V_i^k(n)=\sum\limits_{j\in S}V_{ij}^k(n)$ — наблюдаемое к моменту времени *n* количество принятий решения *k* в состояния *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 $\hat{p}_{ij}^k(n)=\dfrac{V_{ij}^k(n)}{V_i^k(n)}$, которые имеют смысл при $V_i^k(n)>0$.

Пусть $p_{ij}^k \ge \delta > 0$ при всех i,j,k. В этом случае система на каждом шаге попадает в каждое состояние с вероятностью не меньшей δ. Тогда $M\left(\sum\limits_{k=1}^{K_i}V_i^k(n)\right)\ge \delta n$, и, следовательно, для любого $\varepsilon>0$ можно указать такое n_0, что при всех $n\ge n_0$ 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 $\varepsilon,\varepsilon_1>0$ можно указать такое n_0, что при всех $n\ge n_0$ выполнится неравенство

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

Модель управления при неизвестных параметрах

Модель управления состоит в том, что сначала на отрезке времени длины *N* выполняются подсчеты величин $V_{ij}^k(n)$, затем в момент времени *N* вычисляются оценки параметров $\hat{p}_{ij}^k(N)$, 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 вследствие применения этой стратегии, если начальное состояние равно *i*, обозначим через $r(\pi_1^*(N),i,\beta)$. Ее применение оправдано следующей теоремой.

Теорема. Пусть $p_{ij}^k \ge \delta > 0$ при всех i,j,k. Для любого $\varepsilon>0$ можно указать такое *N*, что выполнится предельное неравенство

$$\lim_{\beta\to 1}\frac{r(\pi_1^*(N),i,\beta)}{r(\pi^*,\beta)}\ge 1-\varepsilon. \qquad (2)$$

Доказательство. Пусть $\beta\ge\beta_0$, т. е. оптимальная стратегия π^* одинакова при всех рассматриваемых β. Пусть ε_1 таково, что при выполнении условия $\max\limits_{i,j,k}|\hat{p}_{ij}^k-p_{ij}^k|<\varepsilon_1$ оптимальная стратегия π^*, соответствующая вероятностям $\hat{p}_{ij}^k$, совпадает со стратегией π^*, соответствующей вероятностям p_{ij}^k. Выберем *N* из условия (1), тогда справедлива оценка $r(\pi_1^*(N),i,\beta)\ge(1-\varepsilon)\beta^N r(\pi^*,\beta)$, откуда следует справедливость предельного неравенства (2). Теорема доказана.

Заключение

Вывод, который можно сделать из работы: параметры системы, описываемой при помощи цепей Маркова с дисконтируемыми доходами, можно определить опытным путем. При устремлении к бесконечности времени испытаний вероятности, определенные опытным путем, сходятся к реальным вероятностям системы. Это позволяет обеспечить получение предельного дохода, сколь угодно близкого к оптимальному по относительной величине.

1. Ховард Р. Динамическое программирование и марковские процессы. М.: Советское радио, 1964. 192 с.
2. Баруча-Рид А. Элементы теории марковских процессов и их приложения. М.: Наука, 1969. 512 с.
3. Майн Х., Осаки С. Марковские процессы принятия решений. М.: Наука, 1977. 176 с.