Обучение с подкреплением (Reinforcement Learning)

К. В. Воронцов, А.В. Зухба
vokov@forecsys.ru
a__l@mail.ru

ноябрь 2014

Содержание

Задача о многоруком бандите

A — множество возможных *действий*

$p_a(r)$ — неизвестное распределение *премии* $r \in \mathbb{R}$ за $\forall a \in A$

$\pi_t(a)$ — *стратегия* агента в момент t, распределение на A

Игра агента со средой:

1: инициализация стратегии $\pi_1(a)$
2: **для всех** $t = 1, \dots T, \dots$
3: агент выбирает действие $a_t \sim \pi_t(a)$;
4: среда генерирует премию $r_t \sim p_{a_t}(r)$;
5: агент корректирует стратегию $\pi_{t+1}(a)$;

$$Q_t(a) = \frac{\sum_{i=1}^{t} r_i [a_i = a]}{\sum_{i=1}^{t} [a_i = a]} \quad \text{— средняя премия в } t \text{ играх}$$

$$Q^*(a) = \lim_{t\to\infty} Q_t(a) \to \max_{a\in A} \quad \text{— \textit{ценность действия} } a$$

Примеры прикладных задач

- Управление технологическими процессами
- Управление роботами
- Показ рекламы в Интернете
- Управление ценами и ассортиментом в сетях продаж
- Игра на бирже
- Маршрутизация в телекоммуникационных сетях
- Маршрутизация в беспроводных сенсорных сетях
- Логические игры (шашки, нарды, и т.д.)

Задача о многоруком бандите впервые рассмотрена в статье
H. Robbins. Some aspects of the sequential design of experiments. Bulletin of the American Mathematics Society, 58:527–535, 1952.

Жадная стратегия

Множество действий с максимальной текущей оценкой ценности:

$$A_t = \operatorname{Arg}\max_{a\in A} Q_t(a)$$

Жадная стратегия — выбирать любое действие из A_t:

$$\pi_{t+1}(a) = \frac{1}{|A_t|}[a \in A_t]$$

Недостаток жадной стратегии — по некоторым действиям a можем так и не набрать статистику для оценки $Q_t(a)$.

ε*-жадная стратегия* (компромисс «изучение—применение»):

$$\pi_{t+1}(a) = \frac{1-\varepsilon}{|A_t|}[a \in A_t] + \frac{\varepsilon}{|A|}$$

Эвристика: параметр ε имеет смысл уменьшать со временем.

Метод UCB (upper confidence bound)

«Полужадная» стратегия: выбирать действие с максимальной верхней оценкой ценности:

$$A_t = \operatorname*{Arg\,max}_{a \in A} \left(Q_t(a) + \sqrt{\frac{2 \ln t}{k_t(a)}} \right),$$

где $k_t(a) = \sum_{i=1}^{t} [a_i = a]$.

Интерпретация:
чем меньше $k_t(a)$, тем менее исследована стратегия, тем выше должна быть вероятность выбрать a.

P. Auer, N. Cesa-Bianchi, P. Fischer. Finite-time analysis of the multiarmed bandit problem, Machine Learning, 2002.

Стратегия softmax (распределение Гиббса)

Мягкий вариант компромисса «изучение—применение»:
чем больше $Q_t(a)$, тем больше вероятность выбора a:

$$\pi_{t+1}(a) = \frac{\exp\bigl(Q_t(a)/\tau\bigr)}{\sum\limits_{b\in A}\exp\bigl(Q_t(b)/\tau\bigr)}$$

где τ — параметр *температуры*,
при $\tau \to 0$ стратегия стремится к жадной,
при $\tau \to \infty$ — к равномерной, т.е. чисто исследовательской

Эвристика: параметр τ имеет смысл уменьшать со временем.

Какая из стратегий лучше?

— зависит от конкретной задачи,

— решается в эксперименте

Модельные эксперименты в обучении с подкреплением

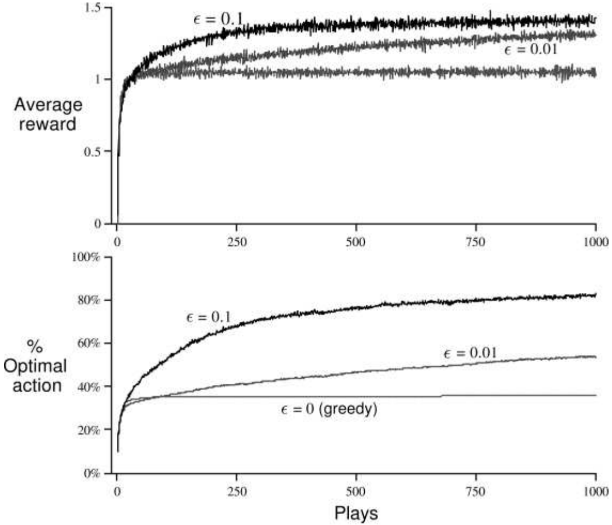
«10-рукая испытательная среда»:
Генерируется 2000 задач, в каждой задаче

$$|A| = 10,$$
$$p_a(r) = \mathcal{N}(r; Q^*(a), 1),$$
$$Q^*(a) \sim \mathcal{N}(0, 1).$$

Строятся графики зависимости
— среднего вознаграждения (average reward),
— доли оптимальных действий (% optimal action),
от числа шагов t, усреднённые по 2000 задачам.

Richard Sutton, Andrew Barto. Reinforcement Learning: An Introduction. The MIT Press. 1998, 2004.
http://webdocs.cs.ualberta.ca/~sutton/book/ebook/the-book.html
Русский перевод:
Р. Саттон, Э. Барто. Обучение с подкреплением. Изд-во «Бином». 2011.

Сравнение жадных и ε-жадных стратегий



Рекуррентная формула для эффективного вычисления средних

Общая формула вычисления Q_t для корректировки стратегии:

$$Q_{t+1}(a) = (1-\alpha_t)Q_t(a) + \alpha_t r_{t+1} = Q_t(a) + \alpha_t\big(r_{t+1} - Q_t(a)\big)$$

При $\alpha_t = \frac{1}{k_t(a)+1}$ это среднее арифметическое, $k_t(a) = \sum\limits_{i=1}^{t}[a_i = a]$

При $\alpha_t = \text{const}$ это экспоненциальное скользящее среднее

Условие сходимости к среднему:

$$\sum_{t=1}^{\infty} \alpha_t = \infty, \qquad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

Среднее арифметическое — для стационарных задач

Экспоненциальное скользящее среднее — для нестационарных (в этом случае сходимости нет, но она и не нужна)

Экспоненциальное скользящее среднее (напоминание)

Задача прогнозирования временного ряда $y_0, \ldots, y_t, \ldots$:
— простейшая регрессионная модель — константа $y_t = c$,
— наблюдения учитываются с весами, убывающими в прошлое,
— прогноз $\hat{y}_{t+1}$ методом наименьших квадратов:

$$\sum_{i=0}^{t} w_{t-i}(y_i - c)^2 \to \min_c, \quad w_i = \beta^i, \quad \beta \in (0, 1)$$

Аналитическое решение — формула Надарая-Ватсона:

$$c \equiv \hat{y}_{t+1} = \frac{\sum_{i=0}^{t} \beta^i y_{t-i}}{\sum_{i=0}^{t} \beta^i}$$

Запишем аналогично $\hat{y}_t$, оценим $\sum_{i=0}^{t} \beta^i \approx \sum_{i=0}^{\infty} \beta^i = \frac{1}{1-\beta}$,

получим $\hat{y}_{t+1} = \hat{y}_t \beta + (1 - \beta) y_t$, заменим $\alpha = 1 - \beta$:

$$\hat{y}_{t+1} = (1 - \alpha)\hat{y}_t + \alpha y_t = \hat{y}_t + \alpha(y_t - \hat{y}_t)$$

Метод сравнения с подкреплением (reinforcement comparison)

Идея: использовать не сами значения премий, а их разности со средней (эталонной) премией:

$\bar{r}_{t+1} = \bar{r}_t + \alpha(r_t - \bar{r}_t)$ — *средняя премия*

$p_{t+1}(a_t) = p_t(a_t) + \beta(r_t - \bar{r}_t)$ — *предпочтения* действий

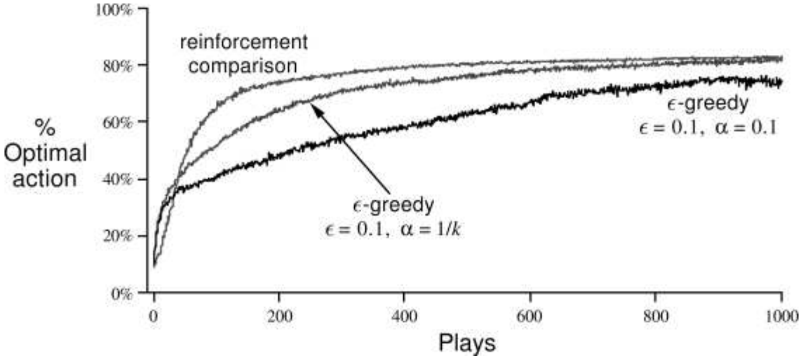
$$\pi_{t+1}(a) = \frac{\exp\big(p_{t+1}(a)\big)}{\sum\limits_{b\in A} \exp\big(p_{t+1}(b)\big)}$$ — softmax-стратегия агента

Эвристика: оптимистично завышенное начальное $\bar{r}_0$ стимулирует изучающие действия в начале

Экспериментальный факт: сравнение с подкреплением сходится быстрее ε-жадных стратегий.

Сравнение с подкреплением лучше ε-жадных стратегий

Эксперимент с 10-рукой испытательной средой:



Метод преследования (pursuit) жадной стратегии

Вместо собственно *жадной стратегии*

$$\pi_{t+1}(a) = \frac{[a \in A_t]}{|A_t|}$$

предлагается *преследование* (сглаживание) жадной стратегии:

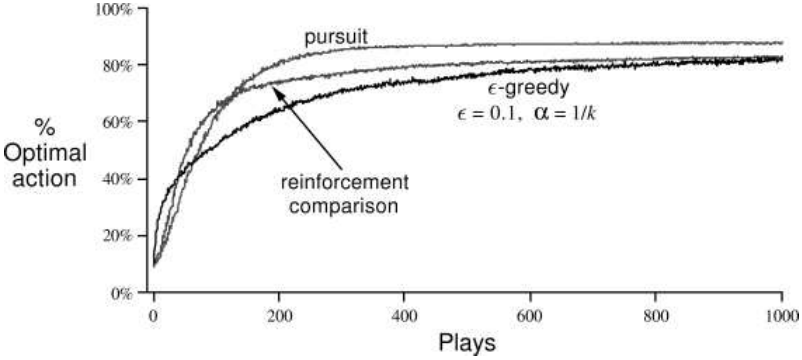
$$\pi_{t+1}(a) = \pi_t(a) + \beta \left(\frac{[a \in A_t]}{|A_t|} - \pi_t(a) \right)$$

Эвристика: начальное $\pi_0(a)$ можно взять равномерным.

Экспериментальный факт: метод преследования, сравнение с подкреплением и ε-жадные стратегии имеют каждый свою область применения.

Стратегия преследования ещё лучше

Эксперимент с 10-рукой испытательной средой:



Общая постановка задачи обучения с подкреплением

S — множество состояний среды

Игра агента со средой:

1: инициализация стратегии $\pi_1(a|s)$ и состояния среды s_1
2: **для всех** $t = 1, \dots T, \dots$
3: агент выбирает действие $a_t \sim \pi_t(a|s_t)$;
4: среда генерирует премию $r_{t+1} \sim p(r|a_t, s_t)$
и новое состояние $s_{t+1} \sim p(s|a_t, s_t)$;
5: агент корректирует стратегию $\pi_{t+1}(a|s)$;

Это *марковский процесс принятия решений* (МППР), если

$$\mathsf{P}\bigl(s_{t+1} = s', r_{t+1} = r \bigm| s_t, a_t, r_t,\ s_{t-1}, a_{t-1}, r_{t-1}, \dots, s_1, a_1\bigr) = \\ = \mathsf{P}\bigl(s_{t+1} = s', r_{t+1} = r \bigm| s_t, a_t\bigr)$$

МППР называется финитным, если $|A| < \infty$, $|S| < \infty$.

Выгода. Ценность состояния. Ценность действия

$R_t = r_{t+1} + r_{t+2} + \cdots + r_{t+k} + \cdots$ — *суммарная выгода*

Обобщение — *дисконтированная выгода*:

$R_t = r_{t+1} + \gamma r_{t+2} + \cdots + \gamma^{k-1} r_{t+k} + \cdots$

$\gamma \in [0, 1]$ — коэффициент дисконтирования:

чем выше γ, тем более агент дальновидный

Функция ценности состояния s при стратегии π:

$$V^\pi(s) = \mathsf{E}_\pi(R_t | s_t = s) = \mathsf{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s\right)$$

Функция ценности действия a *в состоянии* s при стратегии π:

$$Q^\pi(s, a) = \mathsf{E}_\pi(R_t | s_t = s,\ a_t = a) = \mathsf{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \,\middle|\, s_t = s,\ a_t = a\right)$$

E_π — мат.ожидание при условии, что агент следует стратегии π

Метод временных разностей TD(0)

Рекуррентная формула для ценности состояния $V^\pi(s)$:

$$\begin{aligned} V^\pi(s) &= \mathsf{E}_\pi\bigl(\textstyle\sum_{k=0}^\infty \gamma^k r_{t+k+1} \bigm| s_t = s\bigr) = \\ &= \mathsf{E}_\pi\bigl(r_{t+1} + \gamma \textstyle\sum_{k=0}^\infty \gamma^k r_{t+k+2} \bigm| s_t = s\bigr) = \\ &= \mathsf{E}_\pi\bigl(r_{t+1} + \gamma V^\pi(s_{t+1}) \bigm| s_t = s\bigr) \end{aligned}$$

Метод временных разностей TD (temporal difference)
После того, как выбрано a_t и стали известны r_{t+1}, s_{t+1}, оцениваем $V^\pi(s)$ экспоненциальным скользящим средним:

$$V(s_t) := V(s_t) + \alpha_t\bigl(r_{t+1} + \gamma V(s_{t+1}) - V(s_t)\bigr)$$

Утв. Если α_t уменьшается ($\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$), и все s посещаются бесконечное число раз, то $V(s) \overset{\text{пн}}{\to} V^\pi(s)$, $t \to \infty$

Метод SARSA (state–action–reward–state–action)

Рекуррентная формула для ценности действия $Q^\pi(s,a)$:

$$Q^\pi(s,a) = \mathsf{E}_\pi\left(\sum_{k=0}^\infty \gamma^k r_{t+k+1} \;\middle|\; s_t = s,\ a_t = a\right) =$$
$$= \mathsf{E}_\pi\left(r_{t+1} + \gamma Q^\pi(s_{t+1}, a_{t+1}) \;\middle|\; s_t = s,\ a_t = a\right)$$

Игра агента со средой:

1: инициализация стратегии $\pi_1(a|s)$ и состояния среды s_1
2: **для всех** $t = 1, \dots T, \dots$
3: агент выбирает действие $a_t \sim \pi_t(a|s_t)$:
$a_t = \arg\max_a Q(s_t, a)$ — жадная стратегия
(но возможны и другие: ε-жадная, по Гиббсу, ...)
4: среда генерирует $r_{t+1} \sim p(r|a_t, s_t)$ и $s_{t+1} \sim p(s|a_t, s_t)$;
5: агент разыгрывает ещё один шаг: $a' \sim \pi_t(a|s_{t+1})$;
6: $Q(s_t, a_t) := Q(s_t, a_t) + \alpha_t\bigl(r_{t+1} + \gamma Q(s_{t+1}, a') - Q(s_t, a_t)\bigr)$;

Метод Q-обучения

Аппроксимируем оптимальную функцию ценности действия:

$$Q^*(s,a) = \mathsf{E}\Bigl(r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1},a') \Bigm| s_t = s,\ a_t = a\Bigr)$$

Оценка $Q^*(s,a)$ экспоненциальным скользящим средним:

$$Q(s_t,a_t) := Q(s_t,a_t) + \alpha_t\Bigl(r_{t+1} + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\Bigr)$$

Утв. Если α_t уменьшается $(\sum_t \alpha_t = \infty,\ \sum_t \alpha_t^2 < \infty)$, и все s посещаются бесконечное число раз, то $Q \overset{\text{пн}}{\to} Q^*$, $t \to \infty$

Отличия от SARSA: выбрасывается шаг 5 и меняется шаг 6.

Многошаговое TD-прогнозирование

Хотелось бы иметь более надёжную оценку $V(s)$ или $Q(s, a)$, приближающуюся к дисконтированной выгоде R_t

$$\begin{aligned}
R_t^{(1)} &= r_{t+1} + \gamma V(s_{t+1}) \\
R_t^{(2)} &= r_{t+1} + \gamma r_{t+2} + \gamma^2 V(s_{t+2}) \\
&\dots \\
R_t^{(n)} &= r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{n-1} r_{t+n} + \gamma^n V(s_{t+n}) \\
R_t &= r_{t+1} + \gamma r_{t+2} + \dots + \gamma^{n-1} r_{t+n} + \dots
\end{aligned}$$

Премии $r_{t+2}, r_{t+3}, \dots$ в момент t неизвестны, но, оказывается, можно усреднять прошлые, а не будущие наблюдения, и асимптотически это приводит к тому же результату!

Метод временных разностей TD(λ)

Идея «следов приемлемости» $e(s)$:
будем корректировать $V(s)$ не только текущего s_t, но и недавно пройденных состояний, с коэффициентом затухания $\lambda \in [0, 1]$

Обновление $V(s)$ теперь не только для $s = s_t$:

1: $e(s_t) := e(s_t) + 1$;
2: **для всех** $s \in S$, $e(s) \neq 0$
3: $\quad V(s) := V(s) + e(s) \cdot \alpha_t \big(r_{t+1} + \gamma V(s_{t+1}) - V(s)\big)$;
4: $\quad e(s) := \gamma\lambda e(s)$;

Возможны варианты обновления следов приемлемости:
$e(s) := [s = s_t]$ — получаем метод TD(0)
$e(s) := \min\{\gamma\lambda e(s), 1\}$ — «заметающий след»
$e(s) := \big(e(s) < \varepsilon\big)\ ?\ 0 : e(s)$ — обнуление слишком старых следов

При $\lambda = 0$ имеем TD(0), при $\lambda = 1$ приближаемся к оценке R_t

Методы SARSA(λ) и $Q(\lambda)$

Идея следов приемлемости легко переносится на метод SARSA:

Обновление $Q(s, a)$ теперь не только для $s = s_t$:

1: $e(s_t, a_t) := e(s_t, a_t) + 1$;
2: **для всех** $s \in S$, $a \in A$: $e(s, a) \neq 0$
3: $\quad Q(s, a) := Q(s, a) + e(s, a) \cdot \alpha_t \big(r_{t+1} + \gamma Q(s_{t+1}, a') - Q(s, a)\big)$;
4: $\quad e(s, a) := \gamma \lambda e(s, a)$;

... и на Q-обучение, если положить
$a' := \arg\max_a Q(s_{t+1}, a)$;

Важная деталь: исследовательские действия должны прерывать следы приемлемости, иначе будут строиться неверные оценки оптимальной стратегии.

Адаптивный ε-жадный метод временных разностей

Идея: чем сильнее колебания (дисперсия) $Q_t(s,a)$, тем больше должна быть вероятность ε_t исследовательских действий.

$$f(s,a) = \left|\frac{\exp(Q_t(s,a)/\sigma) - \exp(Q_{t+1}(s,a)/\sigma)}{\exp(Q_t(s,a)/\sigma) + \exp(Q_{t+1}(s,a)/\sigma)}\right|$$

$$\varepsilon_{t+1}(s) = \varepsilon_t(s) + \delta\big(f(s_t,a_t) - \varepsilon_t(s)\big)$$

Рекомендации:

$\delta = 1/|A(s)|$, $A(s)$ — число возможных действий в состоянии s

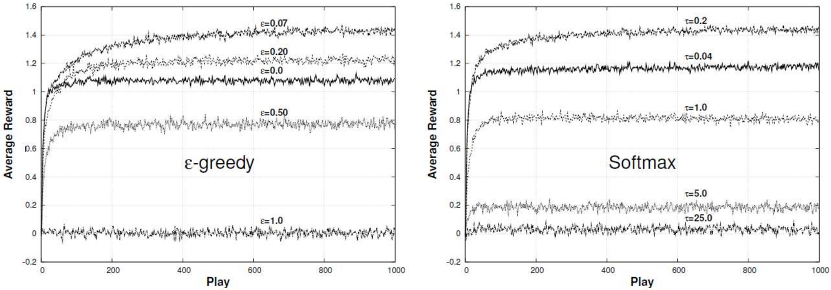
σ — *обратная чувствительность* (inverse sensitivity),

при $\sigma \to 0$ — чисто исследовательская стратегия

Инициализация: $\varepsilon_1(s) \equiv 1$ — чисто исследовательская стратегия

Michel Tokic. Adaptive ε-greedy exploration in reinforcement learning based on value differences // 33rd German conference on Advances in artificial intelligence. 2010. Pp. 203–210.

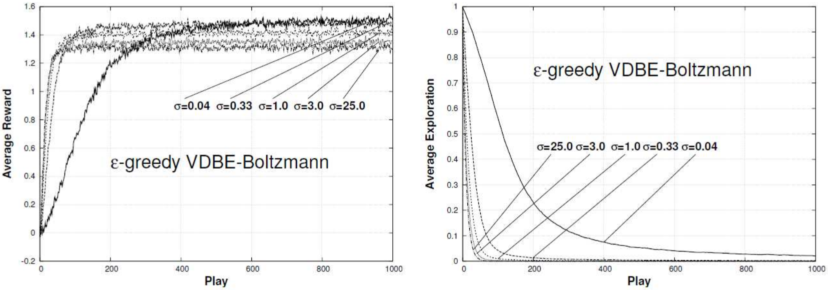
ε-жадные стратегии и softmax



ε-жадные стратегии чувствительны к выбору параметра ε

стратегия softmax чувствительна к выбору температуры τ

Адаптивный ε-жадный метод VDBE



Метод VDBE (value differences based exploration)

— обгоняет ε-жадные стратегии и softmax;

— постепенно уменьшает долю исследований ε;

— может легко сочетаться с другими методами

Резюме в конце лекции

- В обучении с подкреплением нет ответов учителя, есть только ответная реакция среды
- Задача о многоруком бандите — это простой случай среды с одним состоянием
- В общей задаче ценность состояний и действий зависит от состояний среды, но также может быть оценена экспоненциальным скользящим средним
- Компромисс «изучение–применение» обычно подбирается под задачу экспериментальным путём