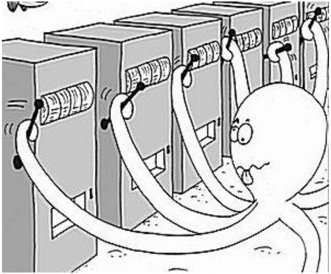
Model-based Interval Estimation for MDP

Alexander L. Strehl, Micheal L. Littman
Journal of Computer and System Science - Elsevier
IF: 1.497

Background

- Exploration or exploitation? **k-armed bandit problem**
- Probability Approximately Correct (PAC)
- Confidence Intervals
- Model-based Interval Estimation (MBIE)

k-armed bandit



- k arms, one choice at a time, a payoff, drawn from an unknown distribution with mean μ_i and variance
- **optimal long term**: always select the arm with highest known mean.

$$\mathrm{argmax}_i\{\mu_i\}$$

- Any problems?

k-armed bandit

- You don't get the chance to **explore** other machines
- Or, you don't get the chance to **exploit** what it was learned
- You just waste money to learn!
- Solution: Interval Estimation (IE)

Interval Estimation

- In each **trial**, construct **confidence interval** for the mean of the payoff distribution for each **arm**

$$[\hat{\mu}_i - \epsilon_i, \hat{\mu}_i + \epsilon_i]$$

- If the confidence interval is loose, the mean payoff may not be near optimal, we need to gain more information
- Model-based Interval Estimation (MBIE) is based on PAC optimality.

What is PAC?

- Probability Approximately Correct
- In a supervised learning classification problem:
- Set of observations instances X
- Set of hypothesis H
  - class of linear functions
  - class of polynomials
  - class of radial basis functions
- Set of concepts C
- Training set $\{(x_i, y_i)\}_i^m = \{(x_i, c(x_i))\}_i^m$

$$y = c(x), \quad c \in C$$

$$\hat{y} = h(x), \quad h \in H$$

$$|H|e^{-\epsilon m} < \delta$$

$$m \geq 1/\epsilon(ln|H| + ln(1/\delta))$$

[0]Watch this! It's good!
https://www.youtube.com/watch?v=qOMOYMOWCzU&t=1181s

MBIE

- Definitions
  - occupancy count $n(s,a)$
  - next-state count $n(s,a,s')$
  - model size limit m
- Reward Confidence Interval
  - immediate reward from taking action a from state s: r[1], r[2], ... r[n(s,a)]
  - empirical mean reward

$$\hat{R}(s,a) = \frac{1}{n(s,a)} \sum_{i=1}^{n(s,a)} r[i]$$

  - confidence interval

$$CI(R) = (\hat{R}(s,a) - \epsilon^R_{n(s,a)}, \hat{R}(s,a) + \epsilon^R_{n(s,a)})$$

$$\epsilon^R_{n(s,a)} = \sqrt{\frac{2/\delta_r}{2.n(s,a)}}$$

MBIE

- Transition Confidence Interval

$$\hat{T}(s'|s,a) = \frac{n(s,a,s')}{n(s,a)}$$

$$\epsilon^{T}_{n(s,a)} = \sqrt{\frac{2(ln(2^{|S|}) - 2) - ln(\delta_T)}{m}}$$

- from Hoeffiding bounds, with probability at least $1 - \delta$, R and T are in the confidence interval

MBIE

- Value iteration with confidence interval:

$$Q'(s,a) = \max_{\tilde{R}(s,a) \in CI(R)} \tilde{R}(s,a) + \max_{\tilde{T}(s,a) \in CI(T)} \gamma \sum_{s'} \tilde{T}(s'|s,a) \max_{a'} Q(s',a') \quad (1)$$

- MBIE-EB solves

$$Q'(s,a) = \tilde{R}(s,a) + \gamma \sum_{s'} \tilde{T}(s'|s,a) \max_{a'} \tilde{Q}(s',a') + \frac{\beta}{\sqrt{n(s,a)}} \quad (2)$$

[0]item page 5, difference between MBIE and MBIE-EB

Thank You!