強化学習と確率ロボットのモデル比較

2015/12/17

@mabo0725

趣旨

- 工作ロボットやゲーム制御の強化学習と自動運転での確率ロボットが混同される場合がある
  （同時期に共に著しい成果が達成されているため）
- 両モデルはマルコフ過程での状況(観測)と最適行動の学習は同じであるが方法は決定的に異なる
- 両モデルの比較を行う（邦訳がある)
  – 強化学習:Reinforcement Learning (Sutton)
  – 確率ロボット:Probabilistic Robotics（Thurn)

強化学習

- 状況(s)　報酬(r)　行動(a)　戦略(π)
- 現在の行動(a)選択は前回選択した状況(s)のみに依存する(マルコフ決定過程)が前提
  - Markov Decision Process(MDP)
- 状況(s)に於いて将来の報酬が最大となる行動(a)を選択する
- 将来の報酬が最大となる価値関数V(s)もしくは行動関数Q(s,a)の算出が目的

価値関数V　行動関数Q

- 将来状況(s')　行動(a')の漸化式で表す

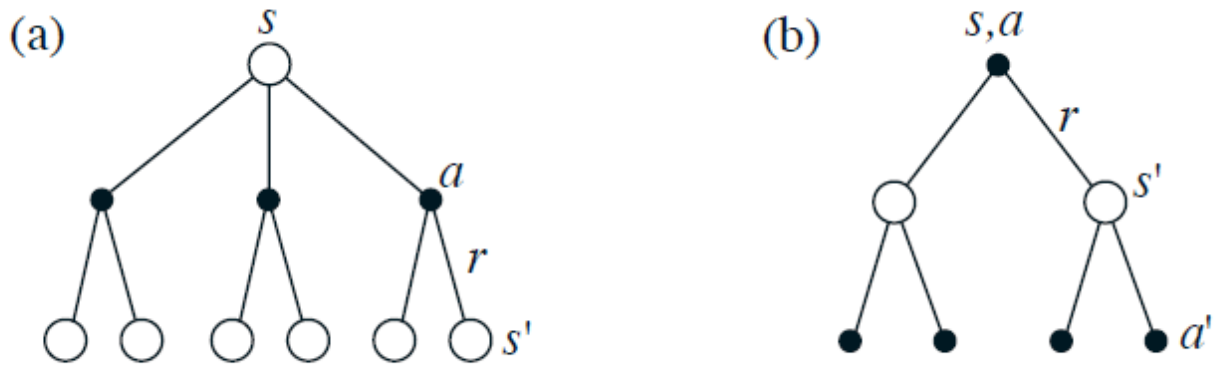


Figure 3.4: Backup diagrams for (a) v_π and (b) q_π.

価値関数V　行動関数Q

- MDPでは次回の価値は将来の漸化式で計算

$$
\begin{aligned}
v(s) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s'} p(s'|s,a) \Big[r(s,a,s') + \gamma v(s') \Big], \qquad (4.5)
\end{aligned}
$$

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a] \qquad (4.6) \\
&= \sum_{s'} p(s'|s,a) \Big[r(s,a,s') + \gamma v_\pi(s') \Big].
\end{aligned}
$$

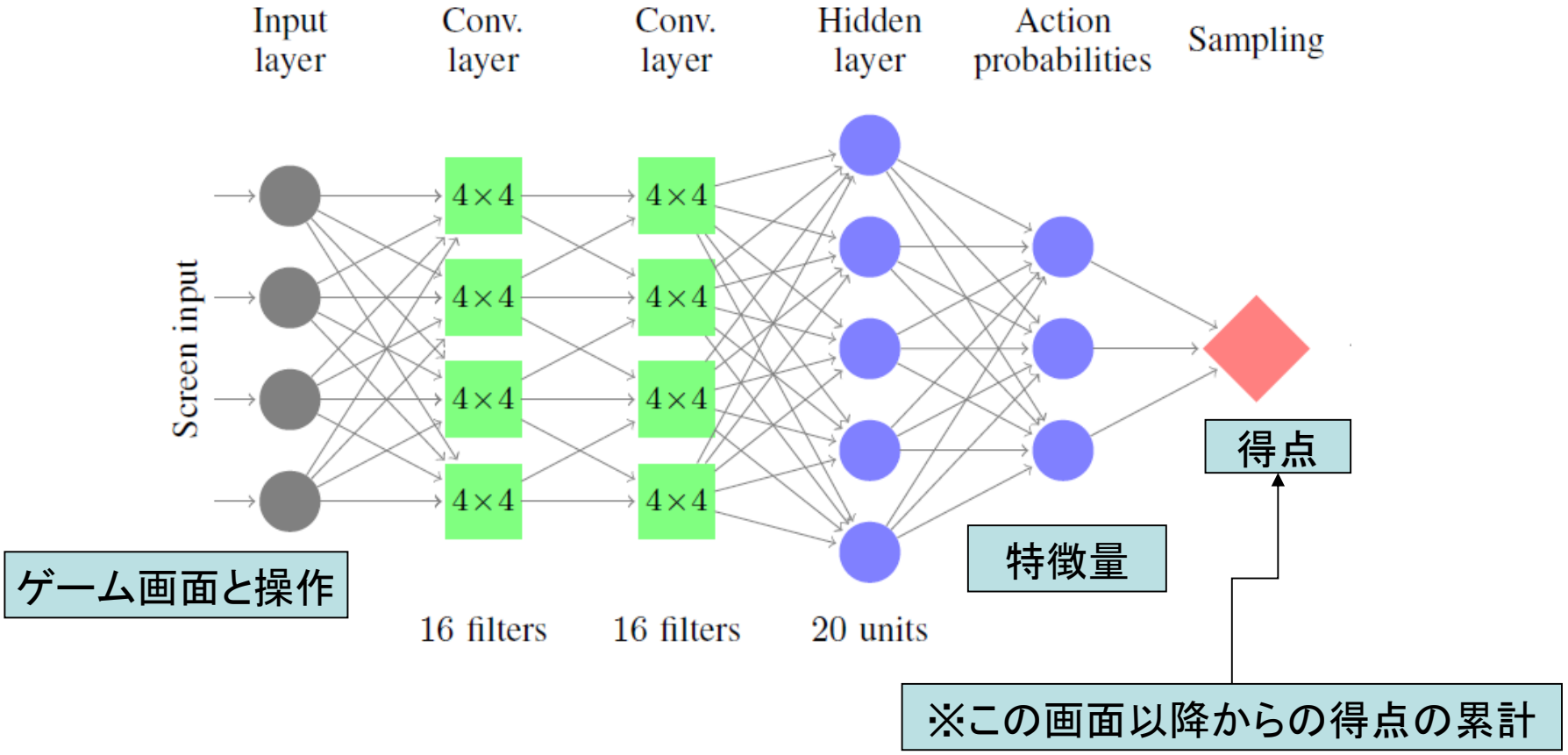
γ:価値の減少率　将来の価値は低く見積もる

価値関数V 行動関数Qの算出方法

- 将来の状況と行動の分岐を繰返し展開して末端からBuckUpで関数を算出する
- 将来価値にγ:減少率があるので無限に展開しなくて良い(n期先まで展開)。
  - 動的計画法 (遷移を定常になるまで繰返し)
  - モンテカルロ法 (全場合の抽出と出現確率）
  - TD(n)法 (V関数のパラメータをSDGで計算)
  - Sarsa(n)法(Q関数のパラメータをSDGで計算)
  - ニューロ法　(Q関数をニューロで汎用化)

DQN(Deep Q-Learning Net)

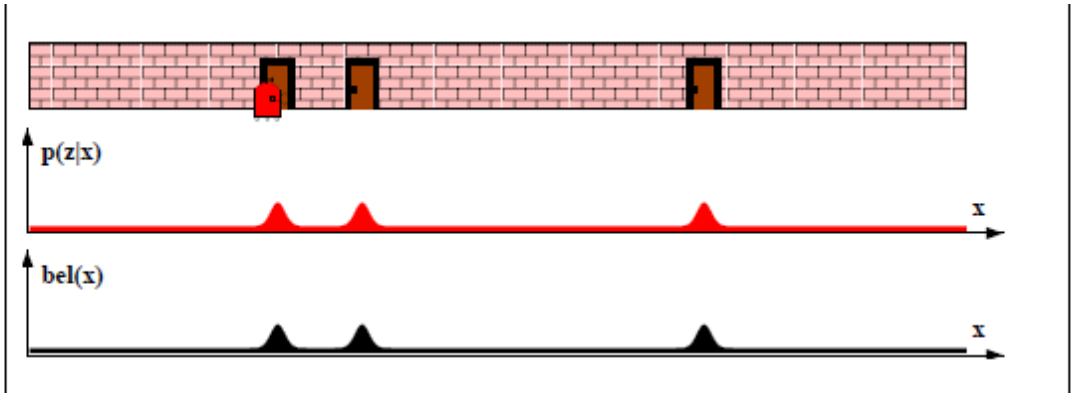
- 行動関数QをDeep CNNで訓練する



確率ロボット

- 信念確率belを事前確率として観測結果pによる事後確率で信念確率を更新するモデル
- 事例で説明

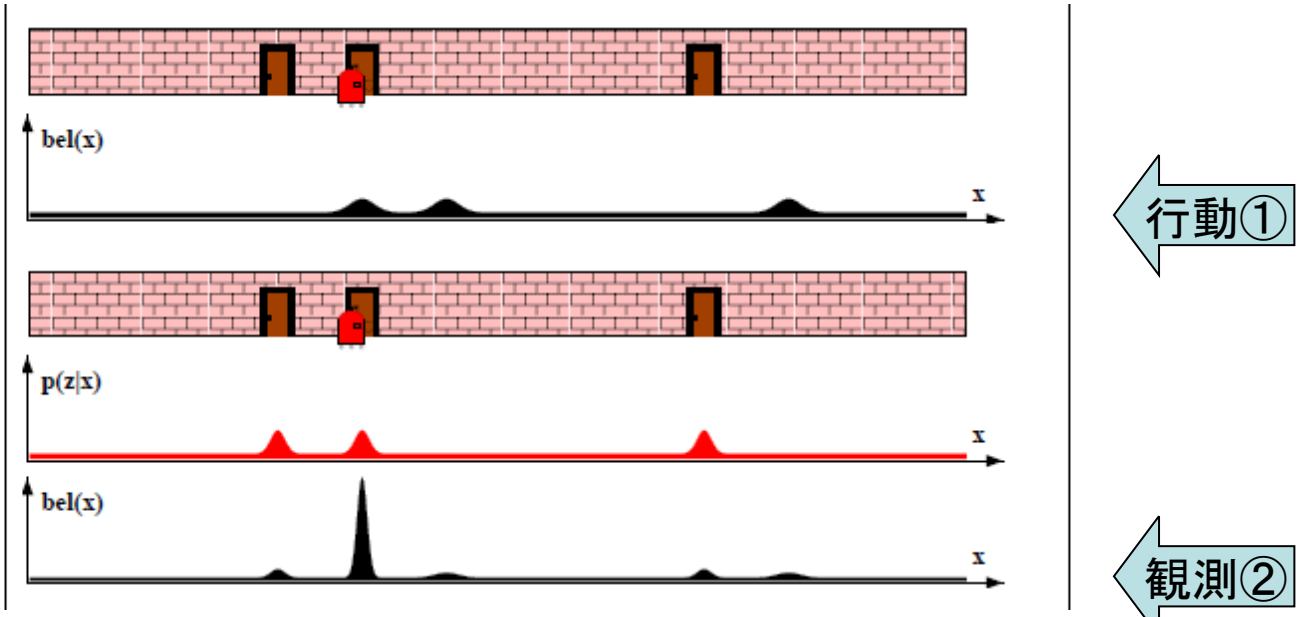
  ロボットにはドアの地図情報とドアを識別するセンサーがある



ドアの地図情報とドアの認識で自分の位置の信念確率

行動と観測による信念の更新

- 信念確率belを行動uと観測結果pで2回更新



自分の位置を観測結果で更新

ベイズ信念更新アルゴリズム

1: **Algorithm Bayes_filter**($bel(x_{t-1}), u_t, z_t$):
2: for all x_t do
3: $\overline{bel}(x_t) = \int p(x_t \mid u_t, x_{t-1})\ bel(x_{t-1})\ dx$
4: $bel(x_t) = \eta\ p(z_t \mid x_t)\ \overline{bel}(x_t)$
5: endfor
6: return $bel(x_t)$

Table 2.1 The general algorithm for Bayes fi ltering.

3行目	行動uによる状況xに対する信念の更新	矢印①
4行目	観測pによる状況xに対する信念の更新	矢印②

まとめ

- 強化学習も確率ロボットもマルコフ過程下の学習で最適な行動を選択する
- 強化学習は将来報酬の価値関数を学習して、価値関数が最大になる行動を選択する。
- 確率ロボットは行動と観測により事後確率が最大の信念確率に従って行動する
  - 自動運転ではOn-lineなので事後確率の高速化のため粒子フィルターを採用している