基于最小最大逼近强化学习的误差分析*

吴沧浦　　刘念泉
(北京理工大学自动控制系　100081)　(北京燕山石化集团公司研究院)

摘　要　在基于动态规划的强化学习中, 利用状态集结方法可以减小状态空间的大小, 从而在一定程度上克服了维数灾的困难, 同时还可以加快学习速度。但状态集结是一种逼近方法, 由此产生的问题是, 状态集结后的Q - hat 强化学习收敛所得的最优Q 值函数与集结前相应的最优Q 值函数会有多大的误差。为此提出了基于最小最大逼近强化学习的误差估计。

关键词　强化学习, 马尔可夫决策问题, 动态规划, 函数逼近, 误差分析
分类号　TP 273

Error Analysis of Minimax- based Reinforcement Learning with State Aggregation

Wu Cangpu　　*Liu Nianquan*
(Beijing Institute of Technology)　(Rearch Institute of Yanhua Group Corporation)

Abstract　In dynamic programming- based reinforcement learning, state space size can be reduced and learning can be accelerated using state aggregation method. Although the real time minimax- based Q - hat- learning after state aggregation turns out to be a non- Markov decision process, it is shown that the corresponding Q - values sequence converge to the optimal Q value with probability one. Furthermore, since state aggregation is a kind of approximation method, an important issue caused is how high the error that will be produced between the minimax optimal Q values obtained in Q - hat- learning after state aggregation and that obtained originally. It is proved that under mild conditions, the error can be reduced as small as possible.

Key words　reinforcement learning, Markov decision problem, dynamic programming, function approximation, error analysis

1 引　　言

近年来, 人们将动态规划与随机逼近理论相结合, 提出许多基于动态规划的解决优化随机控制问题的强化学习算法[1- 4]。在将这些强化学习算法应用于实际问题时, 最重要的问题之一是如何才能比表格表示法更紧凑地描述并储存代价函数。这些基于动态规划的学习算法, 缺陷是其收敛理论的一个共同假定是以表格表示法来描述代价函数。业已证明, 使用其它函数逼近方法可能对这些算法的收敛性造成不利影响。

本文论述了简单函数逼近方法, 即状态集结法与基于最小最大强化学习算法相结合的问题, 讨论了在状态集结下的基于最小最大判据在线Q - hat 学习[4] 与Watkins 的在线Q 学习相对应的收敛理论, 并进一步提出在基于最小最大强化学习中, 应用简单函数逼近方法时所产生误差的估计。

2 基于最小最大马尔可夫决策任务和Q - hat 学习

首先描述一下离散时间序列的马尔可夫决策模型。设S 表示马尔可夫决策问题(MDP) 的状态集合,A 表示每一状态可行动作的集合,$C(C \subset R)$ 表示即时的代价集合。这里S,A 和C 都是有限集, 每一时间步的当前状态i 是可观测的, 且决策a 由A 中

* 国家自然科学基金项目(69674005)
1998 - 10 - 05 收稿, 1999 - 01 - 11 修回

选择。当执行 a 后系统以概率 $P_{ij}(a)$ 转移到下一状态 j，伴随此状态转移付出即时代价 $r \in C(|r| < \infty$，对于所有的 $i, j, a)$。

按照 Heger[5] 的定义，对应于策略 π 定义总的折扣代价

$$R^{\pi} = \sum_{k=0}^{\infty} \gamma^k r_k^{\pi}$$

其中，$\gamma(0 < \gamma < 1)$ 为折扣因子，r_k^{π} 为相对于策略 π 在时间步 k 的即时代价。对于基于最小最大马尔可夫决策任务，以评价函数 V^{π} 来度量在最小最大判据下执行给定策略的代价。V^{π} 定义如下

$$V^{\pi}(i) = \sup\{v \in R \mid P > 0(R^{\pi} > v \mid I_0^{\pi} = i)\}$$
$$\forall i \in S \quad (1)$$

其中 I_0^{π} 为起始状态。换言之，$V^{\pi}(i)$ 是系统依照策略 π 由状态 i 出发可能发生的最差的代价。对于所有的 $i \in S$，定义最小最大最优值函数 $V^*(i) = \inf_{\pi} V^{\pi}(i)$。设 $c(i, a, j)$ 是在动作 a 下由状态 i 转移到状态 j 可能发生的最差的即时代价，即

$$c(i, a, j) = \sup\{r \in C \mid P(i, a, j, r) > 0\} \quad (2)$$

其中 $P(i, a, j, r)$ 表示对于给定的起始状态 i，动作 a 及下一状态 j，其即时代价为 r 的概率。设 $N(i, a)$ 为由于动作 a，从状态 i 可能到达的即时状态的集合，即

$$N(i, a) = \{j \in S \mid P(i, a, j) > 0\} \quad (3)$$

其中 $P(i, a, j)$ 表示如果执行动作 a，由起始状态 i 转移到下一状态 j 的概率。

对于策略 π，Q 值定义为

$$Q^{\pi}(i, a) = \max_{j \in N(i, a)} [c(i, a, j) + \gamma V^{\pi}(j)] \quad (4)$$
$$\forall i \in S, \quad a \in A$$

其中 $V^{\pi}(j) = \min_b Q^{\pi}(j, b)$。$Q$ - hat 学习的目的是估计由

$$Q^*(i, a) = \max_{j \in N(i, a)} [c(i, a, j) + \gamma V^*(j)] \quad (5)$$

定义的最小最大最优 Q 值。其中

$$V^*(j) = \min_b Q^*(j, b) \quad (6)$$

Q - hat 学习通常在线使用，即用与环境相互影响的实际经验来逐步改进 Q 值。每一段过程 k 过后，由起始状态 i，执行动作 a 相对应的 Q 值由

$$Q_k(i, a) = \max\{Q_{k-1}(i, a), r_k + \gamma V_{k-1}(j)\} \quad (7)$$

进行修正。其中

$$V_{k-1}(j) = \min_b Q_{k-1}(j, b)$$

对于所有状态和动作，假定 Q 值的原值 $Q_0(i, a)$ 已知，并满足 $Q_0(i, a) \leq Q^*(i, a)$。

对于 Q - hat 学习算法，Heger 证明了如下收敛定理：

定理 1 若在某一状态下所有允许的动作通常在该状态下可以无限地执行，那么随着 $k \to \infty$，对于所有的每一状态和动作二元组，式(7) 中的 $Q_k(i, a)$ 以概率 1 收敛于 $Q^*(i, a)$。

3 状态集结

由于存在维数灾问题，利用基于动态规划的学习去计算并储存代价函数的每一元素通常是不可行的。使用代价函数逼近的简洁方法表示，可将这一局限性缩小到一定程度，例如可用人工神经网络、多项式和决策树等方法来逼近代价函数。状态集结是最简单的逼近方法之一，这一方法可显著减小状态空间的大小，并加速基于动态规划的学习进程。

MDP 的状态集合 $S = (s^1, s^2, \cdots, s^N)$，按 $S = \bigcup_{i=1}^{M} S_i$ 和 $S_i \cap S_j = \phi(i \neq j)$，被分割成 M $(0 < M < N)$ 个簇 $S_1, S_2, \cdots, S_M$。下面用 s 表示原 MDP 中作为个体的状态，X^i 表示所有的 $s \in S_i$ 状态集结成的作为个体的簇状态。这样就形成一个新的簇状态集合 $x = (X^1, X^2, \cdots, X^M)$。在原 MDP 中，在线 Q - hat 学习基本上可从实际系统的 4 元组序列 (s_t, a_t, s_{t+1}, r_t) 得到(时间步 $t \in \{1, 2, \cdots, \infty\}$。这表示在当前动作 a_t 和相应的即时代价 r_t 下，从当前状态 s_t 转移到下一状态 s_{t+1}。这里假定 $s_t \in S_i, s_{t+1} \in S_j$。根据式(7)，$Q$ - hat 学习可将 Q 值进行更新。

但在状态集结后，在基于簇状态集合的 Q - hat 学习中，仅有 4 元组序列 $(X_t^i, a_t, X_{t+1}^j, r_t)$ 可观测，其中 X^i 和 X_{t+1}^j 分别来自 s_t 和 s_{t+1} 的映射。应注意，由于隐藏状态的存在，由此得出的决策过程本质上是非马氏的[6]。尽管如此，若 $s_t \in S_i$，则

$$Q_t(X^i, a) = \max\{Q_{t-1}(X^i, a), r_t + \gamma V_{t-1}(X^j)\} \quad (8)$$

如依照式(8) 应用 Q - hat 学习规则进行学习，仍可证明如下定理：

定理 2[7] 在上述基于最小最大决策任务中，状态集结后依照式(8) 而进行的 Q - hat 学习所得到的 $Q_t(X^i, a)$，将以概率 1 收敛到下式中的 $Q^*(X^i, a)$。

$$Q^*(X^i, a) = \max_{X^j \in N(X^i, a)} \{c(X^i, a, X^j) + \gamma \min_b Q^*(X^j, b)\} \quad (9)$$
$$\forall X^i \in x, \quad a \in A$$

其中

$$c(X^i, a, X^j) = \max\{c(s, a, s') \mid s \in S_i, s' \in S_j, P(s, a, s') > 0\}$$

$$N(X^i, a) = \{X^j \in \mathrm{x} \mid s \in S_i, s' \in S_j, P(s, a, s') > 0\}$$

4　基于最小最大强化学习函数逼近方法的误差估计

定理3　在上述基于最小最大决策任务中, 令 $K = (1, 2, \cdots, M)$, 如果

$$\max_{i \in K} \max_{s, s \in S_i} \max_{a \in A} |Q^*(s, a) - Q^*(s', a)| \leq \epsilon$$

成立, 则状态集结后的 Q 值误差估计

$$\max_{X^i \in \mathrm{x}} \max_{s \in S_i} \max_{a \in A} |Q^*(X^i, a) - Q^*(s, a)| \leq (3 - \gamma)\epsilon/(1 - \gamma)$$

证明　对于具有有限状态集 S 与有限动作集 A 的任一MDP, 设状态 S 被分割为 M 个簇 $S_1, S_2, \cdots, S_M$, 相应的簇状态集合为 $\mathrm{x} = (X^1, X^2, \cdots, X^M)$, 状态集结后的 Q 值为 $Q^*(X^i, a)$。设存在 ϵ 并使

$$\max_{i \in K} \max_{s, s \in S_i} \max_{a \in A} |Q^*(s, a) - Q^*(s', a)| \leq \epsilon \quad (10)$$

下面构造一新的MDP。为方便起见, 以下用MDPA 表示。此MDPA 在状态 $s \in S$ 与动作 $a \in A$ 下的最优 Q 值用 $\overline{Q}^*(s, a)$ 表示, 并定义为

$$\overline{Q}^*(s, a) = \max_{s \in S_i} Q^*(s', a) \quad (11)$$

其中, S_i 为 s 处于其中的状态子集, 即 $s \in S_i$, $Q^*(s, a)$ 为原MDP 的 Q 值。由此可见,MDPA 的 Q 值 $\overline{Q}^*$ 具有如下性质: 对于任一给定的 $i \in K$, 所有的 $s, s' \in S_i$, 所有的 $a \in A$, 下式成立。

$$\overline{Q}^*(s, a) = \overline{Q}^*(s', a) \quad (12)$$

由式(12) 可进一步推得,MDPA 的 Q 值 $\overline{Q}^*$ 还具有如下性质: 对于任一给定的 $i \in K$, 所有的 $X^i = \mathrm{x}$, 所有的 $s \in S_i$, 所有的 $a \in A$, 下式也成立。

$$\overline{Q}^*(X^i, a) = \overline{Q}^*(s, a) \quad (13)$$

新构造的MDP 模型MDPA 的状态转移概率记为 $\overline{P}_{ij}(a)$, 其最差的即时代价记为 $\overline{c}(i, a, j)$。对于任意选取的 $i, j \in S$ 和任意选取的 $a \in A$, 有

$$\overline{P}_{ij}(a) = P(i, a, j) \quad (14)$$

其中 $P(i, a, j)$ 为原MDP 的状态转移概率。对于任意选取的 $i, j \in S$ 和任意选取的 $a \in A$, 有

$$\overline{c}(i, a, j) = c(i, a, j) + \{\overline{Q}^*(i, a) - Q^*(i, a)\} - \gamma\{\overline{V}^*(j) - V^*(j)\} \quad (15)$$

其中, $c(i, a, j)$ 为原MDP 的最差即时代价, $V^*(j)$ 为式(6) 定义的原MDP 的最优值函数。$\overline{V}^*(j)$ 为由

$$\overline{V}^*(i) = \min_a \overline{Q}^*(i, a), \quad \forall i \in K \quad (16)$$

定义的MDPA 的最优值函数。由定义(11) 和(15) 不难推断, 所定义的MDPA 的 Q 值函数 $\overline{Q}^*$ 满足动态规划基本方程

$$\overline{Q}^*(i, a) = \max_{j \in N(i, a)} \{\overline{c}(i, a, j) + \gamma \overline{V}^*(j)\} \quad (17)$$

对所有的 $i \in S$, $a \in A$ 成立, 它与原MDP 的最优 Q 值的定义(5) 的形式完全相同。可见(11), (14) - (16) 所构造的新MDPA 模型是合理的。

为了证明定理 3 的结论, 先对 $\overline{Q}^*$ 与 Q^* 之差做出估计。$\forall i \in S$, $a \in A$, 由式(11) 得

$$|\overline{Q}^*(i, a) - Q^*(i, a)| = |Q^*(s^*(i), a) - Q^*(i, a)| \quad (18)$$

其中 $s^*(i) \in S_i$, S_i 为 i 所处于其中的状态子集, 且满足

$$Q^*(s^*(i), a) = \max_{s \in S_i} Q^*(s', a) \quad (19)$$

于是 $\forall i \in S$, $i, s^*(i) \in S_i$, $a \in A$, 有

$$|\overline{Q}^*(i, a) - Q^*(i, a)| \leq \max_{k \in K} \max_{s, s \in S_k} \max_{a \in A} |Q^*(s, a) - Q^*(s', a)| \leq \epsilon \quad (20)$$

由式(9) 可得

$$Q^*(X^i, a) = \sum_{k=0}^{\infty} \gamma^k r_k^{\pi} \quad (21)$$

其中

$$r_k^{\pi} = c(s^{\pi}(i_k), a_k^{\pi}, s^{\pi}(i_{k+1})) = \max\{c(s, a, s') \mid s \in S_{i_k}, s' \in S_{i_{k+1}}, P(s, a, s') > 0\}$$

$s^{\pi}(i_k)$ 和 $s^{\pi}(i_{k+1})$ 分别为策略 π 下第 k 时段与第 $k+1$ 时段的状态。类似地, 由式(13) 和(17) 可得

$$\overline{Q}^*(X^i, a) = \sum_{k=0}^{\infty} \gamma^k \overline{r}_k^{\pi'} \quad (22)$$

其中

$$\overline{r}_k^{\pi'} = \overline{c}(s^{\pi'}(i'_k), a_k^{\pi'}, s^{\pi'}(i'_{k+1}))$$

由式(15), (20), (16) 和(6) 可得

$$|\overline{c}(i, a, j) - c(i, a, j)| \leq 2\epsilon \quad \forall i, j \in S, \quad a \in A \quad (23)$$

由式(23), (21) 和(22) 得

$$|Q^*(X^i, a) - \overline{Q}^*(X^i, a)| \leq 2\epsilon \sum_{k=0}^{\infty} \gamma^k = \frac{2\epsilon}{1 - \gamma} \quad (24)$$

联合式(20) 和(24) 可知, $\forall X^i \in \mathrm{x}$, $\forall s \in S_i$, $\forall a \in A$, 下式成立。

$$|Q^*(X^i, a) - Q^*(s, a)| =$$

$$\left|Q^*(X^i,a)-\overline{Q}^*(X^i,a)+\overline{Q}^*(s,a)-Q^*(s,a)\right|\leq\frac{(3-\gamma)\epsilon}{1-\gamma} \tag{25}$$

可见定理3的结论成立。

5 结 论

在基于动态规划的强化学习中,利用状态集结方法可以减小状态空间的大小,从而在一定程度上克服了维数灾的困难;同时还可以加快学习速度。但状态集结是一种逼近方法,由此产生的重要问题是,状态集结后的Q-hat强化学习收敛所得的最优Q值函数与集结前相应的最优Q值函数会有多大误差。本文证明了只要在状态集结中,被集结的同簇里任意两个状态对应的Q值的差值足够小,则由集结而导致的最优Q值误差可以达到任意小。

参 考 文 献

1 C J C H Watkins. Learning from delayed rewards. UK: King's College, 1989

2 R S Sutton. Learning to predict by the methods of temporal differences. Machine Learning, 1988, 3: 9-44

3 A G Barto, S J Bradtke, S P Singh. Learning to act using real-time dynamic programming. Artificial Intelligence, 1995, 72: 81-138

4 M Heger. Consideration of risk in reinforcement learning. In: Proc of the 11th Int Conf. Morgan Kaufmann, 1994. 105-111

5 M Heger. The loss from imperfect value function in expectation-based and minimax-based tasks. Machine Learning, 1996, 22: 197-225

6 S D Whitehead, L J Lin. Reinforcement learning of non-Markov decision processes. Artificial Intelligence, 1995, 73: 271-306

7 Guofei Jiang, Cangpu Wu. Function approximation in minimax-based reinforcement learning. In: Proc of IC-SSSE'98. Beijing: Scientific & Technical Documents Publishing House, 1998. 335-345

作 者 简 介

吴沧浦 男,1932年生。1952年毕业于清华大学,1962年于中国科学院研究生毕业,现为北京理工大学教授,博士生导师。主要研究领域为系统最优化,大系统控制与决策,神经网络技术与智能控制等。

刘念泉 男,1957年生。1988年在北京化工学院获硕士学位,现为北京燕山石化集团公司研究院工程师,博士研究生。主要研究方向为智能控制理论及其应用。