一种新的多智能体 Q 学习算法

郭 锐[1] 吴 敏[1] 彭 军[1] 彭 姣[2] 曹卫华[1]

摘 要 针对非确定马尔可夫环境下的多智能体系统, 提出了一种新的多智能体 Q 学习算法. 算法中通过对联合动作的统计来学习其它智能体的行为策略, 并利用智能体策略向量的全概率分布保证了对联合最优动作的选择. 同时对算法的收敛性和学习性能进行了分析. 该算法在多智能体系统 RoboCup 中的应用进一步表明了算法的有效性与泛化能力.

关键词 多智能体, 增强学习, Q 学习
中图分类号 TP18

A New Q Learning Algorithm for Multi-agent Systems

GUO Rui[1] WU Min[1] PENG Jun[1] PENG Jiao[2] CAO Wei-Hua[1]

Abstract Due to the presence of other agents, the environment of multi-agent systems (MAS) cannot be simply treated as Markov decision processes (MDPs). The current reinforcement learning algorithms which are based on MDPs must be reformed before it can be applicable to MAS. Based on an agent′s independent learning ability this paper proposes a novel Q-learning algorithm for MAS —an agent learning other agents′ action policies through observing the joint action. The policies of other agents are expressed as action probability distribution matrixes. A concise and yet useful updating method for the matrixes is proposed. The full joint probability of distribution matrixes guarantees the learning agent to choose his/her optimal action. The convergence and performance of the proposed algorithm are analyzed theoretically. When applied to RoboCup, our algorithm showed high learning efficiency and good generalization ability. Finally, we briefly point out some directions of multi-agent reinforcement learning.

Key words Multi-agent systems, reinforcement learning, Q-learning

1 引言

在机器学习范畴内, 根据反馈的不同将学习分为监督学习和非监督学习两大类, 增强学习 (Reinforcement learning) 属于非监督学习, 是一种以反馈为输入的自适应学习方法, 通过与环境交互, 不断改进最终获得最优行为策略, 由于其在线学习和自适应学习的特点, 增强学习是解决智能体策略寻优问题的有效工具, 在各个领域获得了广泛的应用[1].

多智能体系统 (Multi-agent systems) 是人工智能领域一个活跃的研究分支, 多智能体学习技术的研究将有力地推动多智能体理论的发展[2]. Weiss 等人将多智能体学习分为三类: 乘积形式, 分割形式和交互形式[3], 即认为多智能体系统是一个学习智能体, 或是一个学习独立的系统, 或是一个个体独立但能通过合作学习提高学习效用性的系统. 在多智能体合作学习的情况下, Narendra 等人提出了合作进化学习技术[4], 由于智能体间是合作关系, 系统最大化累计回报的目标与单个智能体最大化累计回报的目标是一致的, 合作进化学习的性能在该种情况下能达到最优; Littman 提出的 Mini-Max Q 学习算法以两个智能体的系统为例, 讨论了多智能体间竞争学习的情况[5], 由于该算法仅以两个智能体的特殊 (非此即彼) 环境为例, 缺乏较好的泛化能力; Littman 还讨论了在半竞争情况下多智能体学习的情况, 提出了 FOF Q 学习算法[6], 解决智能体间既竞争又合作的情况下如何有效学习的问题; 国内高阳等人采用元对策理论提出了多智能体非零和 MDPs(Markov decision processes) 对策的增强学习模型和算法[7], 讨论了在非零和模型下通过预测对手的策略模型来修正自己的策略; Hu 提出了一般情况下多智能体学习的 BR 算法[8], 该算法在仅有一个采用固定策略的对手的情况下是收敛的, 同样, 由于基于的是较特殊的环境, 该算法缺乏较强的泛化能力.

本文基于智能体独立学习的情况, 提出一种多智能体 Q 学习算法, 理论分析表明该算法在一定条件下是收敛的, 由于采用了简洁的动作策略学习技术, 算法中多智能体系统的联合动作学习空间由指数空间降为线性空间, 有效地提高了学习性能, 并将算法应用到多智能体系统 — RoboCup 中. 实验结

收稿日期 2005-11-10 收修改稿日期 2006-4-28
Received November 10, 2005; in revised form April 28, 2006
湖南省自然科学基金项目 (06JJ50144) 和国家杰出青年科学基金项目 (60425310) 资助
Supported by the Provincial Natural Science Foundation of Hunan (06JJ50144) and the National Science Fund for Distinguished Youth Scholars of P.R.China (60425310)
1. 中南大学信息科学与工程学院 长沙 410083 2. 贵州省高速公路开发总公司 贵阳 550003
1. School of Information Science and Engineering, Central South University, Changsha 410083 2. Expressway Development Company of Guizhou, Guiyang 550003
DOI: 10.1360/aas-007-0367

果验证了算法的有效性和泛化能力. 最后简要给出多智能体增强学习研究的方向及进一步的工作.

2 多智能体 Q 学习

2.1 多智能体 Q 学习思想

由于多个智能体的存在, 多智能体系统中问题的求解需考虑智能体间的影响, 多智能体系统不能用马尔可夫模型描述, 基于马尔可夫模型的增强学习不能直接引入多智能体系统.

首先必须改进增强学习所依据的环境模型, 在多智能体系统中, 当前环境状态可因学习智能体自身的动作或其它智能体的动作而改变, 系统失去了封闭性, 后继状态不再仅由当前状态 s 与学习智能体的动作 a 决定, 多智能体增强学习中的回报函数和状态后继函数不能再用 $r(s,a)$ 和 $S'=\delta(s,a)$ 来表示.

再者, 在多智能体系统中, 学习智能体应学习其它智能体的策略, 系统当前状态到下一状态的变迁由学习智能体与其它智能体的动作决定, 当其它智能体的策略未知时, 将造成后继函数的不确定. 多数情况下, 其它智能体的行为是依据了一定的策略, 智能体在某状态下采取的动作是服从一定概率分布的随机行为. 以足球比赛为例, 某一球员不能准确预知对手的意图, 但通过常识与观察, 该球员可以判断：当我方控球后, 临近的对手会截球而不会去进攻, 防守队员将以较大的概率选择截球, 以较小的概率选择进攻, 这说明其它智能体的策略选择服从一定的概率分布, 根据先验知识与观察可以部分确定该概率分布, 也就部分确定了其策略. 因此, 通过在学习过程中对其它智能体的行为进行观察与统计, 可学习其它智能体的策略, 同时获知该策略对环境的影响, 确定回报规则函数 $r(s,\vec{\boldsymbol{a}})$ 和状态后继函数 $S'=\delta(s,\vec{\boldsymbol{a}})$. 为此, 引入统计方法, 通过对状态和动作向量的统计来学习其它智能体的策略.

2.2 多智能体 Q 学习算法

定义学习目标为学习策略 $\boldsymbol{\pi}: S\rightarrow A$,$S$ 为有限状态集, A 为智能体动作集合. 时刻 t 在状态 s_t 下, 智能体选择动作 $a_i\in A$ 的概率分布表示为 $\pi_t=\{P_1,P_2,...,P_i\}$, 策略 $\boldsymbol{\pi}=(\pi_1,\pi_2,...,\pi_t,...)$, 从状态 s_t 开始, 按策略 $\boldsymbol{\pi}$ 获得的期望累计折扣回报 $v^{\boldsymbol{\pi}}$ 为

$$v^{\boldsymbol{\pi}}(s_t)=E(\sum_{i=0}^{\infty}\gamma^i r_{t+i}) \tag{1}$$

其中, $0\le\gamma<1$ 为折扣因子, 反映了对当前回报与未来回报的取舍, r_t 指每次获得的有界回报, 由于是在非确定马尔可夫环境下进行学习, 累计回报加上期望运算, 最佳策略 π^* 是使 (1) 式获得最大值的策略.

为了描述多个智能体的行为引入动作向量 $\vec{\boldsymbol{a}}$, 对 Q 学习算法[9] 改进后有

$$\begin{aligned}Q(s,\vec{\boldsymbol{a}})&=E[r(s,\vec{\boldsymbol{a}})+\gamma V^*(\delta(s,\vec{\boldsymbol{a}}))]\\&=E(r(s,\vec{\boldsymbol{a}}))+\gamma E[V^*(\delta(s,\vec{\boldsymbol{a}}))]\\&=E(r(s,\vec{\boldsymbol{a}}))+\gamma\sum_{s'}P(s'\,\big|\,s,\vec{\boldsymbol{a}})V^*(s')\end{aligned} \tag{2}$$

对 (2) 式进行替换, 得

$$Q(s,\vec{\boldsymbol{a}})=E(r(s,\vec{\boldsymbol{a}}))+\gamma\sum_{s'}P(s'|s,\vec{\boldsymbol{a}})\max_{\vec{\boldsymbol{a}}'}Q(s',\vec{\boldsymbol{a}}') \tag{3}$$

$P(s'\,\big|\,s,\vec{\boldsymbol{a}})$ 表示在状态 s 下, 智能体执行联合动作 $\vec{\boldsymbol{a}}=(a_1,a_2,...,a_i)$ 后其后继状态为 s' 的概率, $\vec{\boldsymbol{a}}'$ 为新状态 s' 下的联合动作.

用 $\hat{Q}_t$ 表示第 t 次迭代后 Q 值的近似值, 则 Q 值能通过下式进行迭代

$$\hat{Q}_{t+1}(s,\vec{\boldsymbol{a}})\leftarrow(1-\alpha_t)\hat{Q}_t(s,\vec{\boldsymbol{a}})+\alpha_t[r_t+\gamma\max_{\vec{\boldsymbol{a}}'}\hat{Q}_t(s',\vec{\boldsymbol{a}}')] \tag{4}$$

α_t 是动态学习率. 用 π_1^* 表示学习智能体的最佳策略, $\hat{\pi}_t^i$ 表示学习智能体对智能体 i 在 t 时刻策略的近似估计, 将 (4) 式变为

$$\begin{aligned}&\hat{Q}_{t+1}(s,\vec{\boldsymbol{a}})\leftarrow(1-\alpha_t)\hat{Q}_t(s,\vec{\boldsymbol{a}})+\\&\alpha_t[r_t+\gamma\max_{\vec{\boldsymbol{a}}'}\sum\pi_1^*\prod_{i=2}^{n}\hat{\pi}_t^i\hat{Q}_t(s',\vec{\boldsymbol{a}}')]\end{aligned} \tag{5}$$

上式即为提出的多智能体 Q 学习算法, 在多智能体环境下智能体可通过该式进行学习.

2.3 算法收敛性和有效性分析

算法中引入动态学习率, 提出动作策略学习更新算法, 并对学习误差有限性进行证明, 这三个引理的提出与证明保证了算法的严谨性. 算法中学习是对智能体动作空间 $|A|$ 的搜索, 对其他智能体则选择策略中极大似然估计概率的动作, 证明该学习搜索在一定条件下等同于联合动作 $|A|^n$ 空间的全搜索, 这就证明了学习是收敛与完备的. 同时根据机器学习理论对算法有效性进行分析.

2.3.1 算法收敛性证明

1) 引理的定义和证明

引理 1. 学习率 $\alpha_t=\frac{1}{1+\beta C_t(s,\vec{\boldsymbol{a}})}$, 其中 $C_t(s,\vec{\boldsymbol{a}})$ 统计了在 t 次学习过程中状态动作对 $(s,\vec{\boldsymbol{a}})$ 出现的次数, β 为常数.

学习率 α_t 的选取基于以下思想：对出现次数多的状态动作对 $(s,\vec{\boldsymbol{a}})$, 因为已进行了多次 Q 值迭代逼近, 较多考虑上次的 Q 值；对出现次数少的状态动作对 $(s,\vec{\boldsymbol{a}})$, 较多考虑后继学习的效用性, 引入参数 $\beta \geq 1$ 是为了增大统计量 $C_t(s,\vec{\boldsymbol{a}})$ 的影响, 加快学习收敛的速度 (下面的讨论将表明 β 的选取不影响学习的收敛性). 随着 $C_t(s,\vec{\boldsymbol{a}})$ 的增大, $\alpha_t \to 0$, 该学习率削弱了 Q 值每次迭代的修改量, 这使得学习过程逐渐趋于平稳. 事实上, 学习率 α_t 的选取不会影响学习算法的收敛性, 先假设 (5) 式中的 $\max\limits_{\vec{\boldsymbol{a}}'}\sum \pi_1^* \prod\limits_{i=2}^{n} \overset{\wedge}{\pi}{}_t^i \hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 部分与 (4) 式中的 $\max\limits_{\vec{\boldsymbol{a}}'} \hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 部分等价. 显然, $(\alpha_t...)$ 序列为不完全几何级数 (β 为常数), $\sum_{t=1}^{\infty}\alpha_t = \infty$, $0<\alpha_t<1$, 但 $(\alpha_t^2...)$ 序列收敛, 即 $\sum_{t=1}^{\infty}\alpha_t^2 < \infty$. Q 学习中当学习率满足该条件时, 随着 $n\to\infty$, $\hat{Q}_t(s,\vec{\boldsymbol{a}})$ 以概率 1 趋近于 $Q(s,\vec{\boldsymbol{a}})$[10], 这表明了学习率 α_t 的选取是有效的.

引理 2. 智能体对其它智能体最优动作策略的学习是对其极大似然策略的学习.

(5) 式中 $\max\limits_{\vec{\boldsymbol{a}}'}\sum \pi_1^* \prod_{i=2}^{n} \overset{\wedge}{\pi}{}_t^i \hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 部分描述了智能体在策略 π_1^* 和其它智能体估计策略 $\overset{\wedge}{\pi}_i$ 下的联合动作概率, 该概率决定了在状态 s' 下, 对动作 $\vec{\boldsymbol{a}}'$ 选择的概率分布, $\overset{\wedge}{\pi}{}_t^i$ 反映了智能体 i 在当前状态下 (t 时刻) 动作的选择, 此时智能体 i 选择 $\overset{\wedge}{\pi}{}_t^i$ 中最大概率动作.

设智能体 i 可选择 m 个动作 (假设系统中智能体同构), 则初始时智能体 i 的策略 $\overset{\wedge}{\pi}{}_t^i$ 的概率向量为 $(1/m, 1/m, ...)$, 即认为智能体 i 对所有动作的选择概率相等, 学习智能体通过统计对智能体 i 的策略进行学习, 更新规则如下：

规则 1. 设 $\overset{\wedge}{\pi}{}_t^i = (x_1/m, x_2/m, x_3/m, ...)$ 代表智能体 i 在 t 时刻的 (近似) 动作概率向量, 如果智能体 i 在 $t+1$ 时刻选择了动作 j, 则

$$\overset{\wedge}{\pi}{}_t^i = (\frac{x_1}{m+1}, \frac{x_2}{m+1}, ..., \frac{x_j+1}{m+1}, ...) \qquad (6)$$

证明. 在 t 时刻 $\sum_{k=1}^{m} x_k = m$, 则在 $t+1$ 时刻 $x_1+x_2+...+x_j+1+...=m+1$, 即 $t+1$ 时刻按式 (6) 更新的概率向量分量之和也等于 1. 又 $x_j < m \Rightarrow x_j+1 < m+1$, 即更新后概率向量的分量依然小于 1. □

规则 1 在学习初期赋予各动作相等的概率, 反映了学习智能体对其它智能体策略的未知, 学习智能体通过规则 1 学习其它智能体的策略, 更新规则保证执行频度高的动作的概率分量不断增大, 同时保证低频动作也具有被选择的概率, 这使得学习智能体对其它智能体策略的学习具有一定的冗错能力与灵活性. 最终, 学习智能体将建立其它智能体的策略模型, 用概率分布 $\overset{\wedge}{\pi}$ 表示, 是学习智能体对其它智能体策略的极大似然估计.

引理 3. 设 a 表示在 k 次学习中智能体 i 有 a 次未选择最佳动作, 则 $\exists n \in N$, 且 $a < n$, 即 a 有界.

证明. 智能体在某状态下采取的动作是服从一定概率分布 (策略) 的随机行为, 设该策略收敛于最佳策略 (可非固定), 即 $\overset{\wedge}{\pi}_i \to \pi_i^*$, π_i^* 代表最佳策略. 对策略 $\overset{\wedge}{\pi}_i$ 的第 t 个 (即 t 时刻) 分向量 $\overset{\wedge}{\pi}{}_t^i$, 有 $\lim_{k\to\infty} |\overset{\wedge}{\pi}{}_{tk}^i - \pi_{tk}^{i*}| = 0$, 式中 $\overset{\wedge}{\pi}{}_{tk}^i$ 与 π_{tk}^{i*} 分别表示智能体 i 第 k 次迭代估计策略与最佳策略, k 代表学习迭代次数, $\overset{\wedge}{\pi}{}_{tk}^i = (x_1/m+k, x_2/m+k, ..., k+1-a/m+k, ...)$, a 表示在 k 次学习中智能体 i 有 a 次未选择该最佳动作 (相应状态下), 设该动作为第 j 个动作, 即 $x_j = k+1-a$, 而 $\pi_{tk}^{i*} = (y_1/m+k, y_2/m+k, ..., k+1/m+k, ...)$, 有

$$\begin{aligned}
&\lim_{k\to\infty} |\overset{\wedge}{\pi}{}_{tk}^i - \pi_{tk}^{i*}| = \\
&\lim_{k\to\infty} [\frac{(x_1-y_1)^2}{(m+k)^2} + ... + \frac{(k+1-a-k-1)^2}{(m+k)^2} + ...]^{1/2} = \\
&\lim_{k\to\infty} [\frac{\sum\limits_{i=1}^{j-1}(x_i-y_i)^2 + \sum\limits_{i=j+1}^{m}(x_i-y_i)^2 + a^2}{(m+k)^2}]^{1/2} = \\
&\lim_{k\to\infty} [\frac{\sum\limits_{i=1}^{m}(x_i-y_i)^2 + a^2}{(m+k)^2}]^{1/2} \quad (\text{其中可设}, x_j = y_j = 0) \\
&= \lim_{k\to\infty} [\frac{\sum \left(\frac{x_i-y_i}{k}\right)^2 + \frac{a^2}{k^2}}{(1+\frac{m}{k})^2}]^{1/2} = 0 \qquad (7)
\end{aligned}$$

显然, $k\to\infty$ 时, 式 (7) 分母趋于 1, 式 (7) 中分子每部分都趋于 0, 所以 $\exists n \in N$, 使得 $a < n$, 即 a 有界. □

2) 算法收敛性证明

证明. 在 (4) 式中, $\max\limits_{\vec{\boldsymbol{a}}'} \hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 部分的目的是在状态 s' 下对所有的联合动作 $\vec{\boldsymbol{a}}'$ 进行搜索, 并求最大 Q 值, 这保证了学习的收敛性；(5) 式中的 $\max\limits_{\vec{\boldsymbol{a}}'}\sum \pi_1^* \prod_{i=2}^{n} \overset{\wedge}{\pi}{}_t^i \hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 部分对学习智能体的动作空间 $|A|$ 进行搜索, 对其它智能体则选择策略 $\overset{\wedge}{\pi}{}_t^i$ 中最大概率动作, 即选择其极大似然估计动作, 该搜索在一定条件下等同于联合动作 $|A|^n$ 空间的全搜

索, (5) 式中 $\max_{\vec{\boldsymbol{a}}'}\sum\pi_1^*\prod_{i=2}^n\overset{\wedge}{\pi}_t^i\hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 与 (4) 式中的 $\max_{\vec{\boldsymbol{a}}'}\hat{Q}_t(s',\vec{\boldsymbol{a}}')$ 等价.

(5) 式中学习是对智能体动作空间 $|A|=m$ 进行搜索, 设 $p_1=1/m$, 学习智能体采用盲目搜索对其所有动作进行尝试, 对智能体 i 则选择其策略概率中最大概率动作, 根据引理 2 其选择概率 $p_i=\frac{k+1-a_i}{m+k}$, a_i 表示在 k 次学习迭代中智能体 i 有 a_i 次未选择该动作, 则用概率表示对联合动作空间 (最优动作) 的搜索

$$p=\sum_{l=1}^{m}\frac{\prod_{i=2}^{n}(k+1-a_i)}{(m+k)^{n-1}}p_1 \tag{8}$$

上式中 p 表示联合动作选择全概率, l 取值 $1-m$ 代表对学习智能体 m 个动作的搜索, k 为学习迭代次数, 如果 p 随着 k 的增大趋于 1 就表明该搜索等同于 $|A|^n$ 空间的全搜索 (从效用上保证选取最优动作), 也就证明了学习算法的收敛性, 对 (8) 式进行变换有

$$p=\frac{\prod_{i=2}^{n}(k+1-a_i)}{(m+k)^{n-1}}\frac{m}{m}$$

$$=\frac{\prod_{i=2}^{n}(k+1-a_i)}{(m+k)^{n-1}}$$

设 $a=\max a_i$, 假设在 k 次学习迭代中智能体未选择最优动作的误差数皆最大, 有

$$p>\frac{(k+1-a)^{n-1}}{(m+k)^{n-1}}=[\frac{1-\frac{a-1}{k}}{1+\frac{m}{k}}]^{n-1} \tag{9}$$

对 (9) 式右端取极限有, $\lim_{k\to\infty}[\frac{1-\frac{a-1}{k}}{1+\frac{m}{k}}]^{n-1}$ 当 $k\to\infty$ 时分母趋于 1, 根据引理 3 可知 a 有界, 因此 $(a-1)/k\to 0$, 分子趋于 1, 即 $\lim_{k\to\infty}[\frac{1-\frac{a-1}{k}}{1+\frac{m}{k}}]^{n-1}=1$, 根据极限夹逼准则, 有

$$1>\lim_{k\to\infty}p>\lim_{k\to\infty}[\frac{1-\frac{a-1}{k}}{1+\frac{m}{k}}]^{n-1}=1\quad\Rightarrow p=1 \tag{10}$$

由 (10) 式可知, p 随 k 的增大趋于 1, 该学习搜索等同于 $|A|^n$ 空间的全搜索, 同时根据引理 1 知, 式 (5) 中的学习率的选取是有效的, 该学习算法是收敛的. □

2.3.2 算法有效性分析

1) 算法误差分析

算法中智能体的动作根据其策略来选择, 其最优动作是策略中的最大概率动作, 如果策略概率分布有误, 将造成联合最优动作选择的弃真而影响算法的有效性. 下面对算法中 $\overset{\wedge}{\pi}_t^i$ 概率误差进行分析, 设智能体 i 可选 m 个动作, 对应 $\overset{\wedge}{\pi}_t^i$ 概率向量有 m 个分量, 那么在一次 (t 时刻) 动作的选择中其假设空间为 m, 在 k 次学习迭代中学到该动作策略的概率为 (该动作至少被观察到 $l\geq k/2$ 次)

$$\sum_{j=l}^{k}C_k^j(\tfrac{1}{m})^j(\tfrac{m-1}{m})^{k-j},\ \text{其中}\ l\geq k/2$$

则学习空间为

$$|H|=[\sum_{j=l}^{k}C_k^j(\tfrac{1}{m})^j(\tfrac{m-1}{m})^{k-j}]^{-1},\ \text{其中}\ l\geq k/2$$

根据 PAC 准则 [11], 当学习次数 (样本数) k 满足: $k\geq(\ln|H|-\ln\delta)/\varepsilon$ 时, 学习智能体至少以 $1-\delta$ 的概率学到该假设, 且该假设的泛化误差概率不超过 ε. 设 $m=10$, $\delta=0.01$,$\varepsilon=0.05$, 此时 $k\geq 112$, 也就是说学习次数不到 115 时就将以 0.99 的概率学到该假设且误差概率不超过 0.05, 因此当学习次数较大 (如远大于 115) 时对策略的学习将达到很好的效果, 联合最优动作选择弃真的概率非常小.

2) 算法可行性讨论

(4) 式中算法要求在状态 s' 下对所有联合动作 $\vec{\boldsymbol{a}}'$ 进行搜索, 文献 [12] 中提出的改进 BR 学习算法即是通过联合动作的全概率分布来保证每一个动作都能被搜索, 对具有 n 个智能体的多智能体系统, 当每个智能体动作空间为 $|A|$ 时, 学习搜索空间为指数空间 $|A|^n$, 当 n 与 $|A|$ 增大时面临维数灾难, 学习效率将急剧下降. (5) 式中算法仅对学习智能体的动作空间 $|A|$ 进行搜索, 对其它智能体则选择策略最大概率动作, 总的学习空间为线性空间 $|A|$, 这有效地降低了算法的复杂度. 其实, 由于算法中联合动作选择策略非常简洁, 根据 Occam's razor 定理[6], 这将使学习算法性能较优.

3 学习算法在 RoboCup 中的应用

RoboCup 机器人仿真足球比赛是一个标准仿真足球比赛平台, 参赛每方开启 11 个智能体 (队员), 比赛环境属于多智能体系统, 其目的是提供一个实时的充满噪声的环境来进行多智能体的研究. 本文通过 RoboCup 平台来验证算法的有效性.

作为球员的智能体应具有踢球、截球等基本技能, 而球队作为一个整体还应具有高层战术策略. 高

层策略不仅关注个体本身还包括个体间的合作与对抗, 如何选择高层策略是一个复杂的问题. 高层策略可看成是智能体在多智能体环境下如何选择最优动作的策略, 因此多智能体 Q 学习是解决该问题的有力工具, 通过学习智能体可获得高层策略.

在比赛中, 球员获球后可根据当前状态执行相应动作, 采用以下特征来描述状态 s：

1) 队员与球坐标；

2) 半径 R 内的对手坐标 op[n][2], R <3, n 指对手数；

3) 半径 L 内的队友坐标 team[m][2], 3< L <25, m 指队友数；

4) 球员与球门两端构成的三角区域内的对手坐标 opp[k][2], k 指区域内的对手数.

考虑 3 名进攻队员 (9, 10, 11) 与 4 名防守队员 (1, 2, 3, 4) 的对抗, 其中一名进攻队员具有学习能力, 学习获球后的动作策略, 其余队员为固定策略. 进攻队员的目标是成功射门与提高控球时间. 进攻队员控球时可选三个动作：带球, 传球, 射门, 防守队员控球时可选：带球, 传球, 清球 (把球踢出界外), 当然动作应在可行的情况下进行选择 (譬如处在越位的情况下就不能传球). 非控球队员按既定策略动作. 训练场景如图 1 所示.

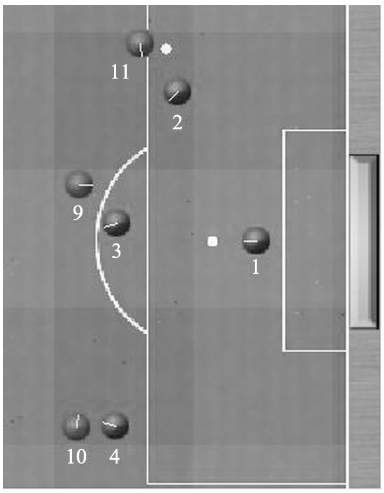

图 1　学习训练场景

Fig. 1　The training episode

动作评价标准如下：对带球而言, 如果带球成功突破对手后且在半径 1.5 米内没有对手, 给其回报 1；对传球而言, 如果队友成功获球后且在该队友 1.5 米半径内没有对手, 给其回报 1；如果射门成功, 给其回报 2. 学习场景为进攻方从最左方某一位置发球开始至射门成功或球超出以上区域. 训练场景通过教练程序设置并记录结果. 通过训练, 学习智能体可获得最优动作策略.

图 2 是不同算法性能间的比较, 纵坐标表示学习智能体在 100 次实验中的平均动作失误数, 这里将不恰当 (不成功) 的带球、传球与射门视为失误动作, 横坐标代表学习步数, 智能体在每 20 学习步后进行性能测试. 同比对照的是采用文献 [12] 中 BR 算法的学习智能体的性能.

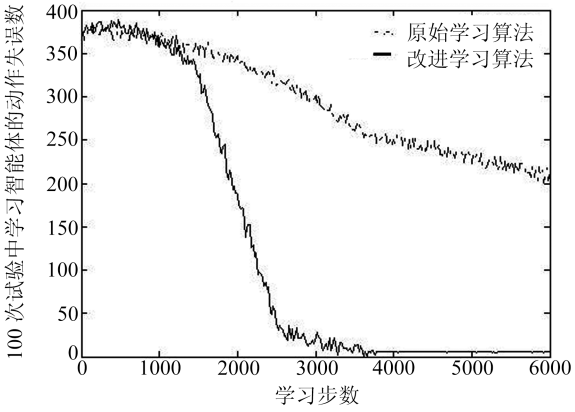

图 2　不同学习算法性能的比较

Fig. 2　The compare of two algorithms′ performance

根据性能曲线, 采用本文学习算法的智能体在 1300 个学习步后性能得到提高, 大约在 2600 个学习步后性能接近最优, 在 4000 个学习步左右即可收敛；采用 BR 算法的智能体的性能在相当多的学习步 (6000 个以上) 后性能依然次于前者, 且尚未收敛. 根据实验结果可得出以下结论：1) 采用本文算法的智能体在一定学习步后其动作策略优于未学习前的策略；2) 本文算法效率优于联合动作全搜索算法. 实验结果表明本文算法具有良好的学习效率和泛化能力, 该算法在多智能体环境中的应用是合适的.

4　结论

针对非确定马尔可夫环境的多智能体系统, 本文提出一种新的 Q 学习算法. 该算法通过对联合动作的统计来学习其它智能体的策略, 并利用策略概率向量的全概率分布保证了对联合最优动作的选择, 在理论上保证了算法的收敛性. 该算法将多智能体环境下的学习空间由指数空间降为线性空间, 有效地提高了学习效率. 同时将该算法应用到多智能体系统 RoboCup 中, 实验结果表明了学习算法的有效性. 由于在一般情况下 (智能体都具有学习能力) 多智能体系统中的学习是一个动态目标问题, 因此, 进一步的研究包括：提出新的多智能体系统学习收敛性的判断；改进学习算法加快学习过程的收敛, 使算法更适应在线学习的情况.

References

1 Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 1996, **4**(2): 237∼285

2 Sandip S C. Adaption, coevolution and learning in multiagent systems. In: Proceedings of AAAI Spring Symposium, AAAI Technical Report SS-96-01, AAAI, 1996. 57∼62

3 Weiss G, Dillenbourg P. What is multi in multiagent learning? *Collaborative Learning, Cognitive and Computational Approaches*. Amsterdam, Holland: Pergamon Press, 1998. 64∼80

4 Narendra P, Sandip S, Maria G. Shared memory based cooperative coevolution. In: Proceedings of IEEE International Conference on Evolutionary Computation, IEEE, 1998. 570∼574

5 Littman M L. Markov games as a framework for multiagent reinforcement learning. In: Proceedings of the 11th International Conference on Machine learning, Morgan Kaufmann, 1994. 157∼163

6 Littman M L. Friend-or-foe: Q-learning in general-sum games. In: Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann, 2001. 322∼328

7 Gao Yang, Zhou Zhi-Hua, He Jia-Zhou, Chen Shi-Fu. Research on Markov game-based multiagent reinforecement learning model and algorithms. *Journal of Computer Research and Development*, 2000, **37**(3): 257∼263
(高阳，周志华，何佳洲，陈世福. 基于 Markov 对策的多 Agent 强化学习模型及算法研究. 计算机研究与发展，2000，**37**(3): 257∼263)

8 Hu J, Wellman M P. Nash Q-Learning for General-Sum stochastic games. *Journal of Machine Learning*, 2003, **4**: 1039∼1069

9 Mitchell T M. *Machine Learning*. USA: McGraw-Hill Companies Inc. 1997, 367∼387

10 Watkins C J C H, Dayan P. Technical note Q-learning. *Journal of Machine Learning*, 1992, (8): 279∼292

11 Haussler D. Quantifying inductive bias: AI learning algorithms and valiant′s learning framework. *Artificial Intelligence*, 1988, **36**(2): 177∼221

12 Weinberg M, Rosenschein J S. Best-response multiagent learning in non-stationary environments. In: Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, IEEE, 2004. 506∼513

郭 锐 博士研究生, 研究方向为智能系统、智能交通系统. E-mail: csumail@sina.com
(**GUO Rui** Ph. D. candidate in School of Information Science and Engineering at Central South University. His research interest covers intelligent systems and intelligent transport systems.)

吴 敏 教授, 研究方向为先进控制理论及应用、过程控制和智能系统. 本文通信作者. Email: min@csu.edu.cn.
(**Wu Min** Professor in School of Information Science and Engineering at Central South University. His research interest covers advanced control theory and its application, process control, and intelligent systems. Corresponding author of this paper.)

彭 军 教授, 研究方向为智能系统与机器人技术.
(**PENG Jun** Professor in School of Information Science and Engineering at Central South University. Her research interest covers multi-agent systems and robotics technology.)

彭 姣 工程硕士, 研究方向为计算机智能应用.
(**PENG Jiao** Engineering master student at Wuhan University. Her research interest covers intelligent application of computer.)

曹卫华 副教授, 研究方向为多智能体系统、机器人技术与过程控制.
(**CAO Wei-Hua** Associate professor in School of Information Science and Engineering at Central South University. His research interest covers multi-agent systems, robotics technology, and process control.)