第 26 卷第 4 期
2009 年 4 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 26 No. 4
Apr. 2009

文章编号：1000−8152(2009)04−0371−06

多智能体强化学习及其在足球机器人角色分配中的应用

段　勇[1], 崔宝侠[1], 徐心和[2]

(1. 沈阳工业大学 信息科学与工程学院, 辽宁 沈阳 110178;

2. 东北大学 人工智能与机器人研究所, 辽宁 沈阳 110023)

摘要: 足球机器人系统是一个典型的多智能体系统, 每个机器人球员选择动作不仅与自身的状态有关, 还要受到其他球员的影响, 因此通过强化学习来实现足球机器人决策策略需要采用组合状态和组合动作. 本文研究了基于智能体动作预测的多智能体强化学习算法, 使用朴素贝叶斯分类器来预测其他智能体的动作. 并引入策略共享机制来交换多智能体所学习的策略, 以提高多智能体强化学习的速度. 最后, 研究了所提出的方法在足球机器人动态角色分配中的应用, 实现了多机器人的分工和协作.

关键词: 多智能体系统; 强化学习; 朴素贝叶斯分类器; 机器人足球; 角色分配

中图分类号: TP242　　**文献标识码**: A

Multi-agent reinforcement learning and its application to role assignment of robot soccer

DUAN Yong[1], CUI Bao-xia[1], XU Xin-he[2]

(1. College of Information Science and Engineering, Shenyang University of Technology, Shenyang Liaoning 110178, China;
2. Institute of AI and Robotics, Northeastern University, Shenyang Liaoning 110023, China)

Abstract: Robot soccer is a typical multi-agent system. The action selected by each robot player not only depends on the current field state, but is also impacted by other players. Hence, the decision-making strategy of robot soccer obtained by reinforcement learning needs the information of the joint-state and the joint-action of multiple agents. A multi-agent reinforcement learning method based on the action prediction of agents is proposed. The Naive Bayes classifier is applied to predict the actions of other agents. Moreover, the sharing-policy mechanism is introduced into multi-agent reinforcement learning system for exchanging the learning policies among agents. It can increase the learning speed effectively. Finally, the proposed approach is applied to learn the role assignment strategy in realizing the cooperation and coordination between robots.

Key words: multi-agent system; reinforcement learning; naive bayes classifier; robot soccer; role assignment

1　引言(Introduction)

机器人足球比赛是近年来提出的多智能体系统(multi-agent system, MAS)开发平台[1], 在该平台上, 各种人工智能和机器人学等领域的研究成果可以得到检验和比较, 进而促进各学科的发展.

决策系统是整个足球机器人系统的中心枢纽, 它的效率和优劣直接影响比赛的成败. 决策系统可以分为3个层次: 角色分配、动作选择和动作执行. 其中角色分配是根据比赛状态信息进行态势分析, 从而为每个机器人球员动态地分配角色, 不同的角色在比赛中承担不同的任务. 目前的角色分配策略多是根据专家知识和实验的方法来设计的, 当专家知识不完备或经验缺乏的情况下, 往往难于设计完善的角色分配策略. 同时由于比赛需要面对不同的球队、不同的策略, 因此主观设计的角色分配策略可能缺少灵活性和适应性, 往往需要不断地调整和补充. 因此本文讨论基于强化学习(reinforcement learning, RL)的动态角色分配策略设计, 使机器人球员在比赛实践中逐渐学习合理、有效的角色分配策略.

对于MAS问题的研究强调多智能体之间的交互作用和相互影响, 每个智能体的动作策略需要考虑其他智能体的行为, 因此多智能体强化学习的最优策略体现的是组合状态到组合动作的映射. 为此, 基于MAS的强化学习理论和算法成为研究的热点

收稿日期: 2007−12−09; 收修改稿日期: 2008−10−31.
基金项目: 国家自然科学基金资助项目(60475036).

问题, 目前已得到广泛的研究[2,3]. 其中最重要的研究成果如下: 基于双矩阵决策(bimatrix games)和随机策略(stochastic games)的理论框架[3], Littman[4]提出了零和决策下的多智能体强化学习方法. 随后, Hu和Wellman对其扩展, 将智能体的学习目标定义为对Nash平衡点的学习, 并证明了算法的收敛性[5].

本文研究一种基于智能体行为预测的多智能体强化学习方法, 采用朴素贝叶斯分类器预测智能体选择的动作, 从而得到所有学习智能体的组合动作. 最后讨论了多智能体强化学习方法在足球机器人动态角色分配中的应用, 使机器人通过学习掌握角色分配策略.

2 基于动作预测多智能体强化学习算法 (Multi-agent RL based on action prediction)

在MAS中, 对学习智能体而言, 当前的环境状态可以通过自身执行的动作来改变, 但由于同时其他智能体也在进行学习, 它们执行动作也在改变环境状态. 因此后续的环境状态对学习智能体是不可推知的, 它与所有智能体的动作有关. 此外, 多智能体强化学习从环境中获得的强化信号是基于多智能体团队的, 而不是基于独立的学习个体. 在多智能体强化学习中, 所有学习智能体的状态变量构成联合状态, 它们执行的动作构成联合动作, 每个智能体学习各自的策略. 所有智能体的联合动作作用于环境并使环境状态发生改变, 同时得到回报. 回报根据某种信度分配策略分配给每个独立的智能体, 以完成各自的强化学习.

2.1 多智能体强化学习(Multi-agent RL)

Q学习是一种重要的强化学习算法[6], 为了提高算法的学习速度, 出现了多种Q学习的改进算法, 主要有多步Q(λ)学习、SARSA学习、HQ学习等. 其中Q(λ)学习是在单步Q学习的基础上, 利用贪婪策略的瞬时差分再次更新Q值. SARSA是一种在线Q学习方法, 在学习的过程中同时更新所有状态、动作对的Q值. 而HQ学习是利用分级的实现来提高学习的速度.

本节将研究基于多智能的Q学习方法, 其中多智能体强化学习的Q函数依赖于所有智能体执行的动作, 因此智能体p的Q函数更新规则为[3]:

$$\begin{aligned}&Q_t^p(s_t^p,\vec{a}_t)=\\&(1-\alpha_t)Q_{t-1}^p(s_t^p,\vec{a}_t)+\\&\alpha_t[r_t^p+\gamma\pi^1(\vec{s}_{t+1})\cdots\pi^n(\vec{s}_{t+1})Q_{t-1}^p(\vec{s}_{t+1})],\end{aligned}\tag{1}$$

$$\begin{aligned}&\pi^1(\vec{s}_{t+1})\cdots\pi^n(\vec{s}_{t+1})Q_t^p(\vec{s}_{t+1})=\\&\sum_{a^1\in A}\sum_{a^2\in A}\cdots\sum_{a^n\in A}P_t^1(\vec{s}_{t+1},a^1)\quad\cdots\\&P_t^n(\vec{s}_{t+1},a^n)Q_{t-1}^p(s_{t-1}^p,a^1,\cdots,a^n).\end{aligned}\tag{2}$$

其中, s_t^p为智能体p的状态变量, $\vec{a}_t=\{a^1,\cdots,a^n\}$表示所有智能体执行的联合动作, $\vec{s}_{t+1}$为下一时刻的联合状态. 智能体p的策略用它的动作集合A^p的概率分布π^p来表示. $P_t^p(\vec{s}_{t+1},a^p)$表示智能体p选在联合状态$\vec{s}_{t+1}$下, 选择动作a^p的概率.

多智能体强化学习的关键问题是如何获取多智能体联合状态和联合动作. 在执行强化学习时, 每个学习个体都能够感知自身的状态, 所以如果能建立良好的通信机制, 就很容易实现联合状态的共享. 然而, 由于所以智能体同时选择动作, 这样智能体无法得知其他智能体将要执行什么动作, 所以它也就无法得知联合动作. 在多数情况下, 其他智能体的行为并不是随意的, 而可以认为是依据一定概率分布的动作策略. 因此研究的多智能体强化学习系统由动作预测单元和强化学习单元构成, 在多智能体强化学习系统中, 每个学习智能体拥有各自的强化学习单元和共同的动作预测单元. 其中动作预测单元使用基于概率的方法来预测其他智能体的动作, 并向强化学习单元提供其他智能体所选择的动作及其预测概率. 强化学习单元将学习样例返回给动作预测单元来更新预测模型, 进而完成多智能体强化学习, 下面提出基于朴素贝叶斯分类器的智能体动作预测方法.

2.2 基于朴素贝叶斯分类器的动作预测(Action prediction based on naive Bayes classifier)

朴素贝叶斯分类器[7,8](naive Bayes classifier, NBC)是使用概率规则来实现推理过程, 并将结果表示为随机变量的概率分布. 朴素贝叶斯分类器算法逻辑简单, 算法开销小, 对于不同特点的数据分类性能稳定[9]. 因此本文使用朴素贝叶斯分类器构成多智能体强化学习的动作预测单元来预测其他智能体的动作.

朴素贝叶斯分类器应用在学习任务中, 将每个训练样本矢量$\boldsymbol{x}=\{x_1,x_2,\cdots,x_n\}$按照最大概率分配给决策类别$\boldsymbol{c}=\{c_1,c_1,\cdots,c_m\}$. 则有:

$$P(\boldsymbol{x}|c_j)=P(x_1,x_2,\cdots,x_n|c_j)=\prod_{i=1}^{n}P(x_i|c_j).\tag{3}$$

根据贝叶斯定理, $\boldsymbol{x}$属于c_j的后验概率为

$$P(c_j|\boldsymbol{x})=\frac{P(\boldsymbol{x}|c_j)P(c_j)}{P(\boldsymbol{x})}.\tag{4}$$

由于$P(\boldsymbol{x})$对于所有的类别均相同, 因此有

$$P(c_j|\boldsymbol{x})\propto P(\boldsymbol{x}|c_j)P(c_j)=P(c_j)\prod_{i=1}^{n}P(x_i|c_j).\tag{5}$$

朴素贝叶斯分类器将未知类别的样本矢量$\boldsymbol{x}$归属于类别c_j, 当且仅当:

$$P(c_j|\boldsymbol{x}) > P(c_k|\boldsymbol{x}), 1 \leqslant k \leqslant m. \quad (6)$$

在学习之前, 由于类别的概率未知, 可以假设各类别出现的概率相同, 即$P(c_1) = P(c_2) = \cdots = P(c_m) = 1/m$. 在学习过程中, 类别概率可以很容易地从类别在训练样本中出现的频率来估计:

$$P(c_j) = v_j/v. \quad (7)$$

其中v_j为训练样本集合中属于类别c_j的样本个数, v为训练样本总数.

使用 $m-$估计方法来估计式 (5) 中的概率 $P(x_i|c_j)$, 这样做的好处是可以在事件出现频率很小时出现一个有偏的过低估计(underestimate)概率. 则$P(x_i|c_j)$的$m-$估计为

$$P(x_i|c_j) = (v_{ji} + m \cdot p)/(v_j + m). \quad (8)$$

其中, v_{ji}为在属于类别c_j的训练样本中样本分量x_i为某种可能的样本个数. p可以看作是x_i的先验估计, 一种典型的方法是使p为假定均匀的先验概率. 而m是一个称为等效样本大小的常量, 它用来确定对于观察到的数据如何衡量p的作用.

根据公式(5), 可以得到智能体在联合状态$\vec{s}$下, 选择动作a^j的条件概率$P(a^j|\vec{s})$, 进而确定智能体所要选择的动作. 在多智能体强化学习过程中, 将每个学习步骤多智能体的联合状态作为训练样本集合, 样本集合的容量随着学习的进程而逐渐增加. 预测智能体的动作集合作为决策类别. 动作预测单元和强化学习单元在学习过程中同时进行, 最终实现完善的动作预测策略和动作选择策略. 在每一步强化学习结束后, 可以统计已学习过的样本中的v, v_j和v_{ji}等值, 从而根据公式(7)和(8)计算训练样本的先验概率$P(c_j)$和$P(x_i|c_j)$, 最后由式(5)预测智能体的动作选择后验概率.

2.3 多智能体强化学习实现方法(Performance method of Multi-agent RL)

对于多智能体强化学习问题, 联合状态和联合动作使它的策略搜索空间要远大于独立智能体强化学习, 更容易引起维数灾难问题. 所以本文将Glorennec和Jouffe[10]的模糊Q学习方法扩展到多智能体强化学习. 将状态的各分量分别进行模糊化, 并使用模糊推理系统(FIS)来逼近强化学习系统状态空间到动作空间的映射, 进而提高多智能体强化学习的速度.

根据公式(1)强化学习状态变量到联合动作的映射用模糊规则来表示, 由于在强化学习阶段, 智能体还没有掌握最优策略, 因此对于给定的模糊规则前提部分, 将所有可能动作的组合作为模糊规则的结论部分, 通过强化学习来确定模糊规则与前提部分最匹配的结论部分. 模糊规则j表示如下:

$$\begin{aligned} R_j: \text{If } s^1 \text{ is } F_j \text{ Then } \vec{a} \text{ is } \vec{a}_{j1} \text{ with } q_{j1}, \\ \text{Or } \vec{a} \text{ is } \vec{a}_{jl} \text{ with } q_{jl}, \\ \cdots \\ \text{Or } \vec{a} \text{ is } \vec{a}_{jL} \text{ with } q_{jL}. \end{aligned} \quad (9)$$

其中, s^1为智能体1的状态, F_j表示模糊集. $\vec{a}_{jl}$和q_{jl}为模糊规则的结论部分, 分别表示状态s^1的可能联合动作及相应的评估值, L表示候选联合动作的个数. 联合动作$\vec{a}_{jl} = \{a^1_{jl}, a^2_{jl}, \cdots, a^n_{jl}\}$中其他智能体的动作$a^2_{jl}, \cdots, a^n_{jl}$通过2.2节的方法预测, 选择当中具有最大概率的动作作为选择的动作. 然后基于此来选择学习智能体的动作a^1_{jl}, 首先得到状态s^1和联合动作$\vec{a}$的评估值:

$$\begin{aligned} &q(s^1, a^1, a^2, \cdots, a^{n-1}) = \\ &\sum_{a^n \in A} P^n(\vec{s}, a^n) \cdot q(s^1, a^1, a^2, \cdots, a^n). \end{aligned} \quad (10)$$

依次推算可以得到:

$$\begin{aligned} q(s^1, a^1) = &\sum_{a^2 \in A} \sum_{a^3 \in A} \cdots \sum_{a^n \in A} P^2(\vec{s}, a^2) P^3(\vec{s}, a^3) \cdots \\ &P^n(\vec{s}, a^n) \cdot q(s^1, a^1, a^2, \cdots, a^n). \end{aligned} \quad (11)$$

式(9)中各动作的概率可以通过动作预测单元得到. 最后, 使用Boltzman策略来选择学习智能体的动作a^1:

$$\text{prob}(a^1_k) = \frac{\exp(\beta \cdot q(s^1, a^1_k)/T)}{\sum\limits_{a^1_h \in A^1} \exp(\beta \cdot q(s^1, a^1_h)/T)}. \quad (12)$$

通过预测的其他智能体动作和由式(12)计算的学习智能体动作可以得到选择的联合动作$\vec{a}_{jl^*}$, 此时每条模糊规则只要被选择的联合动作$\vec{a}_{jl^*}$被激活, 它作为本次学习的模糊规则结论部分.

采用零阶T-S模糊推理系统模型得到智能体1的输出动作和相应的Q值:

$$a(s) = \sum_{j=1}^{N} \bar{\alpha}_j(s) \times a^1_{jl^*}, \quad (13)$$

$$Q(s, \vec{a}) = \sum_{j=1}^{N} \bar{\alpha}_j(s) \times q_{jl^*}. \quad (14)$$

其中$\bar{\alpha}_j(s)$表示规则j的规一化后件适应度. 学习智能体执行选择动作$a(s)$使智能体状态发生转移并得到回报r_t. 联合动作评估值q_{jl^*}的更新使用Q学习算法, 首先由瞬时差分得到:

$$\begin{aligned} \Delta Q = \alpha_t[&r^p_t + \beta \pi^1(\vec{s}_{t+1}) \cdots \\ &\pi^n(\vec{s}_{t+1}) Q^p_{t-1}(\vec{s}_{t+1}) - Q^p_{t-1}(s^p_t, \vec{a}_t)]. \end{aligned} \quad (15)$$

根据动作预测方法可以得到其他智能体的动作选择

概率, 然后由公式(2)可以得到ΔQ. q值的更新公式如下:

$$q_{jl^*}(t) = q_{jl^*}(t-1) + \Delta Q \cdot \bar{\alpha}_j(s) \cdot e_{jl^*}(t). \quad (16)$$

其中$e_{jl^*}(t)$为资格迹. 在强化学习过程中逐渐更新联合动作的评估值q和动作预测单元, 在学习之后, 选择具有最大q值的联合动作中的学习智能体动作作为模糊规则的结论部分, 从而确定学习策略(状态到动作的映射).

3 基于多智能体强化学习的动态角色分配学习(Role assignment learning based on multi-agent RL)

微型机器人足球比赛(MiroSot)是FIRA系列比赛中影响力最大的项目, 将MiroSot 5对5比赛作为本文的研究平台. 定义角色空间共有5类角色, 分别是: 主攻球员、协攻球员、协防球员、主防球员和守门员. 球队中5名机器人球员构成多智能体系统, 其中守门员由固定的机器人来充当, 它在整个比赛中只执行守门的任务. 而其他4名球员采用动态角色分配策略, 通过多智能体强化学习方法来使机器人球员学习角色分配策略.

3.1 动态角色分配学习(Dynamic role assignment learning)

机器人球员角色分配的决策因素如图1所示, 图中, d_{B2HG}和d_{B2OG}为球与我方球门和对方球门的距离; d_{R2B}和α_{RB}为待分配角色的机器人与球的距离和角度; d_{HR2B}和α_{HRB}为我方另一个与球位姿最好的机器人与球的距离和角度; d_{OR2B}和α_{ORB}为对方与球位姿最好的机器人与球的距离和角度; $\theta_{\mathrm{B}v} \in (-\pi, \pi]$表示球运动方向与球和对方球门中心连线的夹角.

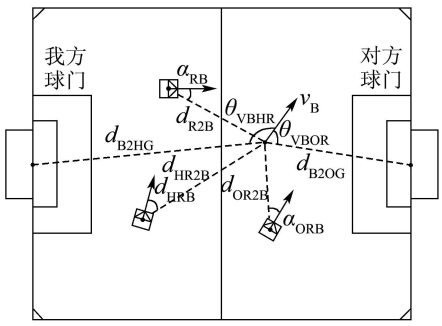

图 1 角色分配示意图

Fig. 1 Sketch of role assignment

机器球员角色分配主要考虑球的位置和运动方向, 机器人与对方球员的竞争因素以及机器人与队友的协作因素. 因此强化学习的状态变量定义如下:

$$s_1 = \mathrm{e}^{-\tau_1(d_{\mathrm{B2HG}}/d_{\mathrm{B2OG}})}, \quad (17)$$

$$s_2 = \mathrm{e}^{-\tau_2|\theta_{\mathrm{VBHG}}/\theta_{\mathrm{VBOG}}|}, \quad (18)$$

$$s_3 = \frac{1}{2}[\mathrm{e}^{-\tau_{31}(d_{\mathrm{R2B}}/d_{\mathrm{OR2B}})} + \mathrm{e}^{-\tau_{32}(\alpha_{\mathrm{RB}}/\alpha_{\mathrm{ORB}})}], \quad (19)$$

$$s_4 = \frac{1}{2}[\mathrm{e}^{-\tau_{41}(d_{\mathrm{R2B}}/d_{\mathrm{HR2B}})} + \mathrm{e}^{-\tau_{42}(\alpha_{\mathrm{RB}}/\alpha_{\mathrm{HRB}})}]. \quad (20)$$

式(17)~(20)中, $s_1, s_2, s_3, s_4 \in [0, 1]$. τ_1, τ_2, τ_{31}, τ_{32}, τ_{41}和τ_{42}均为正常数.

状态变量s_1和s_2分别表示球在球场的位置和球的运动方向. 状态变量s_3显示了我方球员和对方球员的竞争关系, 这里考虑的是与球位姿关系最好的对方球员. s_4体现了我方球员与队友的协作关系, 它可以用于判断谁更适合作为主球员, 这里只考虑另一个与球位姿关系最后的我方球员. 将状态变量s_1, s_2, s_3和s_4进行模糊化, 均表示为3个语言变量S, M和B, 其隶属度函数使用三角函数.

为了便于分配角色, 定义$R = \{1, 2, 3, 4\}$来表示主防队员、协防队员、协攻队员和主攻队员的编号, 所以多智能体强化学习系统的动作变量是机器人的角色编号: $R_l, l = 1, \cdots, 4$. 每名学习球员所选择的动作构成联合动作. 联合动作及相应的评估值作为学习系统模糊规则的结论部分. 根据对其他智能体动作的预测得到学习智能体所选择的动作(分配的角色), 强化学习系统的输出动作为相应的机器人角色编号的解模糊值$\bar{R}$. 然后根据输出值$\bar{R}$与角色编号的接近程度来确定机器人球员的角色. 接近程度用$\bar{R}$和编号的差值来表示, 即

$$e_l = |\bar{R} - R_l|, 1 \leqslant l \leqslant 4.$$

差值的最小值为: $e_{\min} = \min\{e_1, \cdots, e_4\}$, 则最终的机器人角色分配准则为

$$\mathrm{Role} = \begin{cases} \text{主防}, & e_{\min} = e_1, \\ \text{协防}, & e_{\min} = e_2, \\ \text{协攻}, & e_{\min} = e_3, \\ \text{主攻}, & e_{\min} = e_4. \end{cases} \quad (21)$$

同时考虑球与球门的距离以及控球时间等因素对角色分配强化学习的影响. 定义强化信号函数如下:

$$r_t^1 = \begin{cases} +1.0, & \text{我方进球}, \\ -1.0, & \text{对方进球}. \end{cases} \quad (22)$$

$$r_t^2 = \tau \cdot \Delta \bar{d}_{\mathrm{B2G}}, \quad (23)$$

$$r_t^3 = \begin{cases} +0.3, & \mathrm{CTR}_{T0} < \mathrm{CTR}_{T1}, \\ -0.3, & \text{其他}. \end{cases} \quad (24)$$

式(23)将球在球场的位置作为强化信号, $\Delta \bar{d}_{\mathrm{B2G}}$表示相邻学习步骤球与双方球门平均距离的差值, 即$\Delta \bar{d}_{\mathrm{B2G}} = \bar{d}_{\mathrm{B2HG}} - \bar{d}_{\mathrm{B2OG}}$, 其中$\tau$为一正的比例

系数. 公式(24)表示由控球时间决定的强化信号函数, 当我方球队控球时间增加时, 获得奖励, 其中CTR_{T0}和CTR_{T1}分别表示相邻学习周期累计我方的控球时间与对方控球时间的比值.

综合以上因素, 最终的强化信号函数为

$$r_t = \sum_{i=1}^{3} r_t^i. \tag{25}$$

3.2 策略共享机制(Policy sharing mechanism)

在学习过程中特定的阶段共享学习策略以实现多智能体的交互和协作. 由于微型机器人足球比赛采用集中控制方式, 因此可以很方便地实现相互之间的信息交换和共享. 每个独立强化学习系统都学习共同的决策策略, 因此在每个独立系统完成强化学习以后要将学习结果合并成一个共同的模糊推理系统. 共享策略的方法是修改所有独立强化学习系统的状态－动作评估函数, 对于智能体k的第j条规则的第l个候选动作对应的q值为q_{jl}^k, 策略共享的方法是使每个独立学习系统动作的q值等于所有独立学习系统相应q值的平均值, 即$q_{jl} = \dfrac{1}{4} \cdot \sum_{m=1}^{4} q^m(j,l)$.

3.3 回报信度分配策略(Reward assignment strategy)

回报函数r_t是所有独立强化系统输出动作共同作用环境的结果, 因此需要将回报r_t分配给每个学习个体. 为每个独立学习系统设置一个回报分配函数δ^k, $k = 1,2,3,4$. 定义$\delta^k = w_r + w_a + w_p$, 其中$w_r$表示角色权重, 认为主球员要比协助球员重要, 因此独立强化系统分配的角色为主攻或主防, 它得到较大的权重. 当分配的角色能够正确执行选择的动作, w_a取较大的权重. w_p表示该学习周期拥有控球权的角色应获得较大的权重. 最后根据每个角色的回报分配函数计算各强化学习系统所分配的回报, 即$r_t^k = \delta^k \cdot r_t / \sum_{k=1}^{4} \delta_k$. 分配回报时进行规一化处理, 以保证整体回报不变. 各强化学习系统根据重新分配的回报进行强化学习.

4 实验结果及分析(Experimental results and analysis)

使用FIRA标准仿真平台Robot Soccer Simulator进行学习实验, 利用多智能体强化学习方法来使机器人球员学习角色分配策略. 在多智能体强化学习过程中, 所有球员同时执行一次角色分配视为完成一次强化学习训练, 强化学习每进行200次为一个阶段, 在完成每个阶段后利用学习的结果作为动态角色分配控制器. 然后进行5场比赛, 每次比赛时间为5分钟, 比赛对手为强化学习的训练对手.

学习期间, Boltzman策略中参数T应随着学习次数的增加而衰减, 这种变化方式表明在学习的初期以较大的概率随机选择动作, 以便遍历所有可选动作; 在学习的后期, 以较大的概率选择具有较高评估值的动作, 以使学习尽快收敛. 此外, 其他主要参数选择如下: 学习率$0.3 \leqslant \alpha \leqslant 0.8$, 折扣算子$0.9 \leqslant \gamma \leqslant 0.99$, 以保证算法的学习速度和收敛性. 图2为在每个学习阶段预测单元对其他球员选择动作预测的正确率, 可见朴素贝叶斯分类器预测的准确性逐渐增加. 图3为在每个学习阶段的测试比赛中我方控球时间占整个比赛时间的百分比. 图4显示了在整个学习过程中, 我方平均竞胜球数. 从实验结果可见我方球队的竞胜球数也逐渐增多, 能够反映出我方球队的进攻能力和防守能力在学习过程中逐渐提高. 由于足球比赛具有一定的不确定性, 所以双方的得分会出现一定的波动, 但实验结果仍能证明利用强化学习方法来建立足球机器人动态角色分配策略是比较有效的.

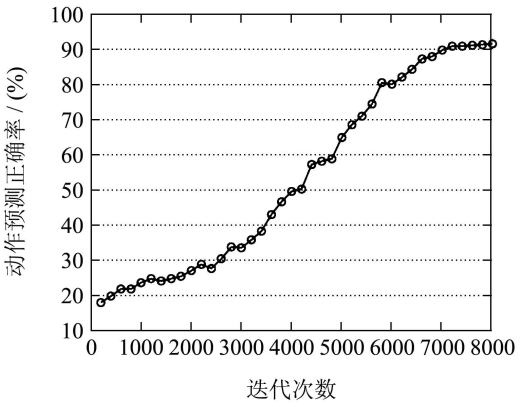

图 2　动作预测正确率

Fig. 2　Right rate of action prediction

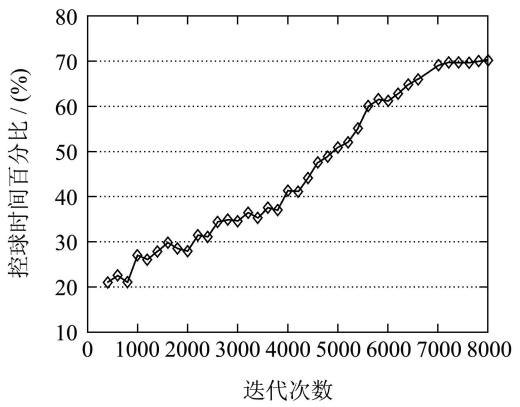

图 3　我方球队控球时间百分数

Fig. 3　Ball possession percent of our team

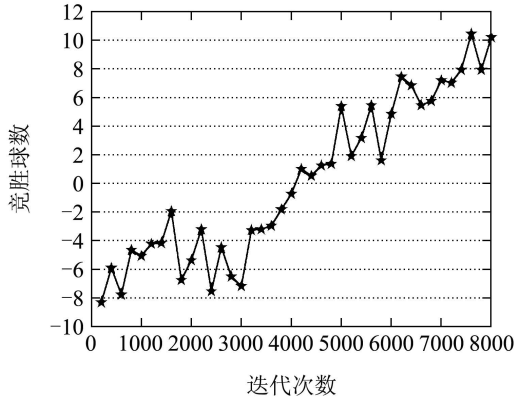

图 4 我方竞胜球数

Fig. 4 Score differences of two teams

5 结论(Conclusion)

本文研究了一种基于动作预测和策略共享机制的多智能体强化学习方法. 由于在多智能体团队中, 每个智能体所选择动作不仅决定于自身的状态信息, 还要不可避免地受到其他智能体执行动作的影响. 因此使用朴素贝叶斯分类器动作预测单元并对其他智能体的选择动作进行预测, 实现了智能体团队的协作. 使用提出的多智能体强化学习方法来实现足球机器人动态角色分配策略的学习, 每个机器人作为一个独立的学习个体, 它们学习共同的决策策略, 构成分布式多智能体强化学习系统. 本文的方法可以在不需要专家知识的情况下, 使足球机器人在比赛实践中逐渐掌握决策能力和行为能力.

参考文献 (References):

[1] KIM J H, VADAKEPAT P. Multi-agent systems: a survey from the robot-soccer perspective[J]. *International Journal of Intelligent Automation and Soft Computing*, 2000, 6(1) : 3 – 17.

[2] STONE P, VELOSO M. Multiagent systems: a survey from a machine learning perspective[J]. *Autonomous Robots*, 2000, 8(3) : 345 – 383.

[3] ERFU Y, DONGBING G. *Multiagent reinforcement learning for multi-robot systems: a survey*[R]. Technical Report CSM-404, Department of Computer Science, University of Essex, 2004.

[4] LITTMAN M L. Markov games as a framework for multiagent learning[C] // *Proceeding of the 11th International Conference on Machine Learning*. San Francisco: IEEE , 1994, 157 – 163.

[5] HU J L, WELLMAN M P. Multiagent reinforcement learning: theoretical framework and an algorithm[C] // *Proceeding of the 15th International Conference of Machine Learning*. San Francisco: IEEE , 1998, 115 – 122.

[6] SUTTON R S, BATRO A G. *Reinforcement Learning: An Introduction*[M]. Cambridge, Massachusetts: MIT , 1998.

[7] 李晓毅, 徐兆棣. 增量式贝叶斯分类的原理和算法[J]. 沈阳工业大学学报, 2006, 28(4): 422 – 425.
(LI Xiaoyi, XU Zhaodi. Principle and algorithm of incremental bayes classification[J]. *Journal of Shenyang University of Technology*, 2006, 28(4): 422 – 425.)

[8] DOMINGOS P, PAZZANI M. On the optimality of the simple bayesian classifier under zero-one loss[J]. *Machine Learning*, 1997, 29(2/3): 103 –130.

[9] 赵红, 李雅菊, 宋涛. 基于贝叶斯网络的工程项目风险管理[J]. 沈阳工业大学学报, 2008, 1(3): 439 – 444.
(ZHAO Hong, LI Yaju, SONG Tao. Study on engineering project rist management based on bayesian network[J]. *Journal of Shenyang University of Technology*, 2008, 1(3): 439 – 444.)

[10] JOUFFE L. Fuzzy inference system learning by reinforcement methods[J]. *IEEE Transaction on Systems, Man, and Cybernetics*, 1998, 28(3): 338 – 355.

作者简介:

段 勇 (1978—), 男, 讲师, 博士, 目前研究方向为智能机器人、机器学习等, E-mail: duanyong0607@126.com;

崔宝侠 (1962—), 女, 教授, 博士, 目前研究方向为工业过程控制、管理信息系统等, E-mail: cuibx88@126.com;

徐心和 (1940—), 男, 教授, 博士, 目前研究方向为计算机博弈、智能机器人等, E-mail:xuxh@gmail.com.