一种基于特征向量提取的 FMDP 模型求解方法*

张双民+, 石纯一

(清华大学 计算机科学与技术系,北京 100084)

An Efficient Solution Algorithm for Factored MDP Using Feature Vector Extraction

ZHANG Shuang-Min+, SHI Chun-Yi

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

+ Corresponding author: Phn: +86-10-62785592, E-mail: zsm99@mails.tsinghua.edu.cn, http://www.tsinghua.edu.cn

Received 2004-02-25; Accepted 2004-05-08

Zhang SM, Shi CY. An efficient solution algorithm for factored MDP using feature vector extraction. *Journal of Software*, 2005,16(5):733−743. DOI: 10.1360/jos160733

Abstract: In factored Markov decision process (FMDP) such as Robocup system, the effect to value evaluation of various states is different from each other within state attributes. There are some important state attributes that can determine the whole state value either uniquely, or at least, approximately. Instead of using the relevance among states to reduce the state space, this paper addresses the problem of curse of dimensionality in large FMDP by approximating state value function through feature vector extraction. A key contribution of this paper is that it reduces the computation complexity by constraints reduction in linear programming, speeds up the production of joint strategy by transplanting the value function to the more complex game in reinforcement learning. Experimental results are provided on Robocup free kick, demonstrating a promising indication of the efficiency of the approach and its' ability of transplanting the learning result. Comparing this algorithm to an existing state-of-the-art approach indicates that it can not only improve the learning speed, but also can transplant state value function to the Robocup with more players instead of learning again.

Key words: multi-Agent cooperative problem solving; factored Markov decision process; linear programming; reinforcement learning; curse of dimensionality

摘 要: 在诸如机器人足球赛等典型的可分解马尔可夫决策过程(factored Markov decision process,简称FMDP)模型中,不同状态属性在不同的状态下,对于状态评估的影响程度是不同的,其中存在若干关键状态属性,能够唯一或近似判断当前状态的好坏.为了解决 FMDP 模型中普遍存在的"维数灾"问题,在效用函数非线性的情况下,通过对状态特征向量的提取近似状态效用函数,同时根据对 FMDP 模型的认知程度,从线性规划和再励

* Supported by the National Natural Science Foundation of China under Grant No.60173011 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant Nos.863-317-01-04-99, 2001AA113120 (国家高技术研究发展计划(863))

作者简介: 张双民(1977−),男,湖南岳阳人,博士,主要研究领域为分布式人工智能,多 Agent 系统;石纯一(1935−),男,教授,博士生导师,主要研究领域为分布式人工智能,多 Agent 系统.

学习两种求解角度分别进行约束不等式组的化简和状态效用函数的高维移植,从而达到降低计算复杂度,加快联合策略生成速度的目的.以机器人足球赛任意球战术配合为背景进行实验来验证基于状态特征向量的再励学习算法的有效性和学习结果的可移植性.与传统再励学习算法相比,基于状态特征向量的再励学习算法能够极大地加快策略的学习速度.但更重要的是,还可以将学习到的状态效用函数方便地移植到更高维的 FMDP 模型中,从而直接计算出联合策略而不需要重新进行学习.

关键词: 群体 Agent 合作求解;可分解马尔可夫决策过程;线性规划;再励学习;维数灾

中图法分类号: TP18 **文献标识码**: A

在多 Agent 系统(MAS)中,马尔可夫决策过程(Markov decision process,简称 MDP)是一类有效的群体 Agent 合作求解问题的模型,这是由于,一方面 Agent 知识和能力是有限的,另一方面,由于现实问题的复杂性和分布性,使得用 MDP 模型描述群体 Agent 合作求解模型更加真实和有效.

利用 MDP 模型对群体 Agent 合作求解问题进行策略求解引起了越来越多的关注和重视,一个重要的研究内容是如何降低 MDP 模型中状态空间的划分规模和缩短联合策略的求解时间.众所周知,当 Agent 所处的状态空间呈指数规模增长时,其缓慢的状态空间遍历速度一直是制约群体 Agent 联合策略求解速度的重要因素.因此,利用尽可能小的状态空间学习到有效的联合策略,一直是众多研究人员尽力追求的目标.

根据对 MDP 模型的认知程度,线性规划、策略迭代和再励学习是 3 类常用的研究方向.无论哪类研究方向,利用状态间的相关性压缩状态空间都是加快策略求解速度的常用手段.在对 MDP 模型,包括状态转移概率函数都已知的前提下,运用线性规划和策略迭代方法可以通过理论性计算求解联合策略[1−3].Koller 等人认为,在一类特殊的 MDP 模型——可分解马尔可夫决策过程(factored Markov decision process,简称 FMDP)中,Agent 所处的状态集合是由若干状态属性张成的向量空间,利用状态属性之间的独立性,对状态效用函数作线性化假设,可以将线性规划的约束不等式个数由指数规模化简为多项式规模[4−6].Gestrin 等人首先用相关马尔可夫决策过程(RMDP)模型定义一类特殊的群体 Agent 合作求解问题,并利用同类问题的外部环境相似性,计算通用状态效用函数[7].在对外部环境不了解的情况下,通过再励学习获得状态效用值也是一类求解联合策略的常用方法[8].Tuyls 等人利用 Bayesian 网络将条件概率表近似成决策树,并用决策树代替状态空间[9,10].Uther 将传统的退化树技术应用到再励学习中,将连续状态空间离散化,并搜索和裁减相关状态分支[11].

Gestrin 和 Koller 等人的方法,对于状态属性之间相互影响程度较小,且整体效用分散在个体 Agent 中的群体 Agent 合作求解问题是非常有效的,但正如 Gestrin 在文中承认的那样[4],首先,对于大多数群体 Agent 合作求解问题来说,状态属性之间的相互影响是广泛存在的,而状态效用函数的线性化假设在诸如机器人足球赛等大多数群体 Agent 合作求解问题中也是不成立的.Gestrin 所采用的方法仅仅适用于非常特殊和狭窄的群体 Agent 合作求解问题;其次,Uther 等人的方法虽然能够减少相关状态,加速学习收敛速度,但对于任何群体 Agent 合作求解问题,即使是以往曾经碰到过的类似问题,也需要重新进行学习和计算,而不能借鉴以往的学习成果,做到“举一反三”.

在诸如机器人足球赛等典型的 FMDP 模型中,不同状态属性对于状态评估的影响程度是不一样的,其中存在若干关键属性,能够唯一或近似决定状态的好坏.因此,本文通过提取状态特征向量来近似状态效用值,同时根据对 FMDP 模型,包括状态转移概率函数等的认知程度,从线性规划和再励学习两个求解角度分别进行约束不等式组的化简和状态效用函数的高维移植,从而达到降低求解复杂度、加快联合策略生成速度的目的.本文以机器人足球赛任意球战术配合为背景来验证基于特征向量的再励学习算法的有效性和学习结果的可移植性.与传统再励学习相比,基于特征向量的再励学习算法除了加快策略学习速度以外,还可以将学习到的状态效用函数直接移植到有更多 Agent 参与的比赛中,从而直接计算出联合策略而不需要重新学习.

本文第 1 节简单介绍线性规划方法求解 MDP 模型以及 FMDP 模型的一般步骤.第 2 节和第 3 节从线性规划角度给出基于特征向量的近似求解形式以及相应的约束不等式组的化简过程.第 4 节和第 5 节从再励学习角度给出了基于特征属性的 Agent 再励学习算法,并通过实验验证算法的有效性和高维可移植性,从而扩展了 Gestrin 等人的工作.

1 FMDP 模型和线性规划

众所周知,MDP 是一类应用广泛的随机决策过程,而 FMDP 是 MDP 的一种特殊形式,与普通 MDP 模型相比,FMDP 中的状态集合是以一组相对独立的状态属性张成的向量空间,且状态的转移概率是这些状态属性自身转移概率的乘积[12].例如,机器人足球赛就是一种典型的 FMDP 模型,每个 Agent 的场上位置组成了全局状态,且下一状态的转移概率是每个 Agent 位置变化概率的乘积.

定义 1. 一个简单的 FMDP 模型一般为如下的 5 元组 $\langle \mathbf{X},\mathbf{A},\{P_i\},\mu,R\rangle$*,其中,

$\mathbf{X}$:表示群体 Agent 的环境状态向量空间.令 $\boldsymbol{\mathcal{X}}=\{X_1,X_2,...,X_n\}$ 为 $\mathbf{X}$ 的基,其中 X_i 为环境状态的第 i 个属性,则 $\mathbf{X}$ 是 $\boldsymbol{\mathcal{X}}$ 的扩张空间.令 dom($\boldsymbol{\mathcal{X}}$) 为以 $\boldsymbol{\mathcal{X}}$ 为基向量张成的向量空间,δ为从 $\bigcup_{\mathbf{Z}\in\boldsymbol{\mathcal{X}}^*}\mathrm{dom}(\mathbf{Z})$ 到 $\boldsymbol{\mathcal{X}}^*$ 的映射($\boldsymbol{\mathcal{X}}^*$ 为 $\boldsymbol{\mathcal{X}}$ 的幂集),表示局部状态向量对应的基.

$\mathbf{A}$:表示 Agent 联合动作的集合.

P_i: $\mathrm{dom}(\{X_i\})\times\mathbf{A}\times\mathbf{X}\to\mathcal{R}$,表示第 i 个状态属性的转移概率函数,其中 $\mathcal{R}$ 为实数集合.且 $\forall\mathbf{z}\in\mathrm{dom}(Z)$ ($Z\subset\boldsymbol{\mathcal{X}}$),转移概率函数为 $\mathbf{z}$ 中各个状态属性转移概率函数的连乘积,记做 $P(\mathbf{z}\mid\mathbf{x}',\mathbf{a})=\prod_{i=1}^{|\mathbf{z}|}P_i(z_i\mid\mathbf{x}',\mathbf{a})$, $\mathbf{x}'\in\mathbf{X}$.

μ: $\boldsymbol{\mathcal{X}}\times\mathbf{A}\to\boldsymbol{\mathcal{X}}^*$,表示状态属性的关联函数.例如在机器人足球赛中,无论做什么动作,Agent 下一时刻的位置只和当前时刻的位置有关,因此 Agent 位置的关联集合就是 Agent 位置本身.

R: $\mathbf{X}\times\mathbf{A}\to\mathcal{R}$,表示立即奖励函数.

在 MDP 模型,包括状态转移概率函数都已知的前提下,线性规划是 MDP 的一种理论上的通用求解方法,但由于约束不等式个数随状态属性的增加呈指数规模增长,实际应用并不广泛.然而 FMDP 所具有的特殊性质使得线性规划求解具有了实际可行性.这里,线性规划的变量为 $V(\mathbf{x}_1),\ldots,V(\mathbf{x}_N)$,其中 $V(\mathbf{x}_i)$ 表示状态 $\mathbf{x}_i$ 的效用值变量.FMDP 的标准线性规划求解形式描述如下:

Variable: $V(\mathbf{x}_1),\ldots,V(\mathbf{x}_N)$; ($N=|\mathbf{X}|$,即状态数目)

min: $\sum_{\mathbf{x}\in\mathbf{X}}\alpha(\mathbf{x})V(\mathbf{x})$;

s.t.: $V(\mathbf{x})\ge R(\mathbf{x},\mathbf{a})+\gamma\sum_{\mathbf{x}'\in\mathbf{X}}\prod_{j=1}^{n}P_j(x_j'\mid\mathbf{x},\mathbf{a})V(\mathbf{x}')$, $\forall\mathbf{x}\in\mathbf{X}$, $\forall\mathbf{a}\in\mathbf{A}$.

虽然约束不等式的个数依然为指数规模,但 Gestrin 利用效用函数线性化的前提假设,可以将约束不等式组进行化简[6].然而在诸如机器人足球赛等大多数群体 Agent 合作求解问题中,群体 Agent 的总体收益并不等于个体 Agent 收益的加权线性之和,事实上也不存在个体 Agent 收益,因此也就无法应用 Gestrin 等人提出的化简方法.本文第 2 节给出一种基于状态特征向量的 FMDP 模型的形式化描述,并且在效用函数非线性的情况下,仍然给出基于特征向量的约束不等式组的化简过程.

2 状态特征向量

在 FMDP 模型中,状态由若干相对独立的状态属性组成,但在很多情况下,各属性在不同状态下对于整体状态效用发挥着不同程度的作用,其中存在若干关键属性,对于状态效用的计算起到至关重要的作用.比如在机器人足球比赛中,场上状态由 22 名球员的位置组成,按照传统的状态划分方法,比赛状态应该由 22 名队员的场上位置组成.然而事实上,在每种状态下,只有持球队员的位置是最关键的,基本决定所在球队当前在场上的形势,因此,可以将每种状态下持球队员的位置作为当前状态效用值的重要参考依据,从而近似状态效用函数.状态的特征向量和最小特征向量如定义 2 所示.

---

* 为了阅读方便,若无特殊声明,本文中粗体大写字母代表集合;斜体大写字母代表抽象对象;粗体小写字母代表向量实例;斜体小写字母代表对象实例;花体字母代表对应向量的基.例如,$\mathbf{X}$ 为集合,$\mathbf{x}$ 为 $\mathbf{X}$ 中的一个向量实例,x 为 $\mathbf{x}$ 的某个属性的值,X 为 x 对应的属性,$\mathcal{X}$ 为向量 $\mathbf{x}$ 的基.

定义 2. 在 FMDP 模型中,$\forall \mathbf{x} \in \mathbf{X}$,如果存在 $\mathbf{x}$ 的局部子向量 $\mathbf{y}$,且 $\forall \mathbf{z} \in \mathrm{dom}(\boldsymbol{\mathcal{X}} \setminus \delta(\mathbf{y}))$,$|V(\mathbf{x}) - V(\mathbf{y} \cup \mathbf{z})| < \xi$,称 $\mathbf{y}$ 为 $\mathbf{x}$ 的特征向量.其中分量数目最少的向量,称为 $\mathbf{x}$ 的最小特征向量.这里,$\cup$ 表示两个向量的并操作,ξ 为常数.

定义 2 表明最小特征向量能够基本真实地反应与之相关的所有状态的效用值,这为以后的化简过程奠定了基础.下面的定理将证明最小特征向量的存在性和唯一性.

定理 1. 在 FMDP 模型中,状态的最小特征属性集合是存在且是唯一的,即同一状态不存在两个分量数目一样的最小特征向量.

证明:存在性的证明是显然的,$\forall \mathbf{x} \in \mathbf{X}$,根据定义 2,$\mathbf{x}$ 本身就是一个特征向量.因此必然存在最小特征向量.唯一性的证明用反证法,对于 $\forall \mathbf{x} \in \mathbf{X}$,如果存在两个不同的最小特征属性向量 $\mathbf{y}$,$\mathbf{y}'$.令 $\mathbf{y}'' = \mathbf{y} \cap \mathbf{y}'$,这里,$\cap$ 表示两个向量的交操作.令 $\boldsymbol{\mathcal{Y}} = \delta(\mathbf{y})$,$\boldsymbol{\mathcal{Y}}' = \delta(\mathbf{y}')$.根据定义 2,有

$$\forall \mathbf{z} \in \mathrm{dom}(\boldsymbol{\mathcal{X}} \setminus \boldsymbol{\mathcal{Y}}), |V(\mathbf{x}) - V(\mathbf{y} \cup \mathbf{z})| < \xi \tag{1}$$

$$\forall \mathbf{z} \in \mathrm{dom}(\boldsymbol{\mathcal{X}} \setminus \boldsymbol{\mathcal{Y}}'), |V(\mathbf{x}) - V(\mathbf{y}' \cup \mathbf{z})| < \xi \tag{2}$$

由式(1)、式(2)可知,

$$\forall \mathbf{z} \in \mathrm{dom}(\boldsymbol{\mathcal{X}} \setminus \boldsymbol{\mathcal{Y}} \cap \boldsymbol{\mathcal{Y}}'), |V(\mathbf{x}) - V(\mathbf{y}'' \cup \mathbf{z})| < \xi \tag{3}$$

由式(3)和定义(2)可知,$\mathbf{y}''$ 也是 $\mathbf{x}$ 的一个特征向量.且由于 $\mathbf{y} \neq \mathbf{y}'$,$\mathbf{y}''$ 的分量数目比 $\mathbf{y}$,$\mathbf{y}'$ 都少,这与 $\mathbf{y}$,$\mathbf{y}'$ 是 $\mathbf{x}$ 的最小特征向量矛盾,得证. □

根据最小特征向量的存在性和唯一性,可以定义基于最小特征向量的 FMDP 模型.

定义 3. 一个简单的基于最小特征向量的 FMDP 模型为 6 元组:$\langle \mathbf{X}, \mathbf{A}, K, \{P_i\}, \mu, R \rangle$,其中 $\mathbf{X}, \mathbf{A}, \{P_i\}, \mu, R$ 与定义 1 相同.

K: $\mathbf{X} \to \mathrm{dom}(\boldsymbol{\mathcal{X}}^*)$,表示最小特征属性向量函数.令 $\mathbf{Y}$ 为 $\mathbf{X}$ 的最小特征向量集合,即 $\mathbf{Y} = \{\mathbf{y} \mid \mathbf{y} = K(\mathbf{x}), \mathbf{x} \in \mathbf{X}\}$.

S: $\mathbf{X} \to \mathbf{Y}$,表示最小特征向量的等价类集合函数,$S(\mathbf{y}) = \{\mathbf{x} \mid \mathbf{y} = K(\mathbf{x})\}$,即在同一等价类中的状态,其最小特征向量是一致的.

$\forall \mathbf{y} \in \mathbf{Y}$,令 $f(\mathbf{y}) = \dfrac{\sum_{\mathbf{x} \in S(\mathbf{y})} V(\mathbf{x})}{|S(\mathbf{y})|}$,容易证明,在同一等价类集合中的状态,其效用值近似相等.

推论 1. 对于 $\forall \mathbf{y} \in \mathbf{Y}$,$\forall \mathbf{x} \in S(\mathbf{y})$,$|f(\mathbf{y}) - V(\mathbf{x})| < \xi$.

证明:首先证明,$\forall \mathbf{x}, \mathbf{x}' \in S(\mathbf{y})$,$|V(\mathbf{x}') - V(\mathbf{x})| < \xi$.

由定义 2,

$$\forall \mathbf{z} \in \mathrm{dom}(\boldsymbol{\mathcal{X}} \setminus \delta(\mathbf{y})), \text{满足} |V(\mathbf{x}) - V(\mathbf{y} \cup \mathbf{z})| < \xi \tag{4}$$

由于 $\mathbf{x}' \in S(\mathbf{y})$,因此

$$\exists \mathbf{z}_0 \in \mathrm{dom}(\boldsymbol{\mathcal{X}} \setminus \delta(\mathbf{y})), \mathbf{x}' = \mathbf{y} \cup \mathbf{z}_0 \tag{5}$$

由式(4)和式(5)可得,$|V(\mathbf{x}) - V(\mathbf{x}')| < \xi$.

因此 $|f(\mathbf{y}) - V(\mathbf{x})| \leq \max_{\mathbf{x}' \in S(\mathbf{y})} V(\mathbf{x}') - V(\mathbf{x}) | < \xi$,得证. □

推论 1 表明,如果将 $f(\mathbf{y})$ 近似为 $S(\mathbf{y})$ 中所有状态的效用值,误差均不会超过 ξ.值得一提的是,与以往通过状态间相关性将状态合并的做法不同,近似状态的效用值并不等于将相似状态合并,因为不同状态的转移概率函数也是不一样的,如果合并状态,必然会降低求解结果的准确性.本文近似状态效用函数的目的只是为了化简约束不等式组.在第 4 节会有详细讨论.定义 3 和推论 2 可以将第 1 节给出的线性规划求解形式改写成基于特征向量的近似求解形式.

变量: $f(y_1), \ldots, f(y_M)$

最小化: $\sum_{\mathbf{x} \in \mathbf{X}} \alpha(\mathbf{x}) f(K(\mathbf{x}))$

约束条件: $f(K(\mathbf{x})) \geq R(K(\mathbf{x}), \mathbf{a}) + \gamma \sum_{\mathbf{x}' \in \mathbf{X}} \prod_{j=1}^{n} P_j(x_j' \mid \mathbf{x}, \mathbf{a}) V(\mathbf{x}')$, $\forall \mathbf{x} \in \mathbf{X}$, $\forall \mathbf{a} \in \mathbf{A}$

下一节将详细介绍求解形式的化简过程.

3 线性规划的近似求解

正如上一节所说,FMDP 模型的线性规划求解形式,其约束不等式的个数仍然为指数规模.但如果最小特征向量不随状态属性的增加而改变,则仍然可以进行化简.

3.1 化简不等式

化简过程的第一个重要步骤就是对动作产生的预期效用,即在求解形式中约束不等式右边的部分进行化简.这是必要的,因为很容易看出,由于状态个数是指数规模, $\sum_{\mathbf{x}'\in\mathbf{X}}P(\mathbf{x}'\mid\mathbf{x},\mathbf{a})f(K(\mathbf{x}'))$ 的展开项个数也是指数规模.因此必须首先对预期效用的计算表达式进行化简.

步骤 1.令

$$Q(\mathbf{a},\mathbf{x})=R(K(\mathbf{x}),\mathbf{a})+\gamma\sum_{\mathbf{x}'\in\mathbf{X}}P(\mathbf{x}'\mid\mathbf{x},\mathbf{a})f(K(\mathbf{x}')) \tag{6}$$

令 $g(\mathbf{x})=\sum_{\mathbf{x}'\in\mathbf{X}}P(\mathbf{x}'\mid\mathbf{x},\mathbf{a})f(K(\mathbf{x}'))$,因此 $Q(\mathbf{a},\mathbf{x})=R(K(\mathbf{x}_i),\mathbf{a})+\gamma g(\mathbf{x})$.因为是 FMDP 模型,有

$$g(\mathbf{x})=\sum_{\mathbf{y}\in\mathbf{Y}}\sum_{\mathbf{x}'\in S(\mathbf{y})}P(\mathbf{x}'\mid\mathbf{x},\mathbf{a})f(\mathbf{y})=\sum_{i=1}^{M}P(\mathbf{y}_i\mid\mathbf{x},\mathbf{a})\times f(\mathbf{y}_i)\sum_{\mathbf{u}'\in\mathrm{dom}(\boldsymbol{\mathcal{X}}\setminus\delta(\mathbf{y}_i))}P(\mathbf{u}'\mid\mathbf{x},\mathbf{a}),$$

又因为 $\sum_{\mathbf{u}'\in\mathrm{dom}(\boldsymbol{\mathcal{X}}\setminus\delta(\mathbf{y}_i))}P(\mathbf{u}'\mid\mathbf{x},\mathbf{a})=1$,因此

$$g(\mathbf{x})=\sum_{i=1}^{M}P(\mathbf{y}_i\mid\mathbf{x},\mathbf{a})\times f(\mathbf{y}_i)$$

步骤 2.令 $g_i(\mathbf{x})=P(\mathbf{y}_i\mid\mathbf{x},\mathbf{a})\times f(\mathbf{y}_i)=g_i(\mathbf{x}[\mathbf{Z}_{g_i}])$,其中 $\mathbf{x}[\boldsymbol{\mathcal{Z}}]$ 表示对应 $\boldsymbol{\mathcal{Z}}$ 中属性的 $\mathbf{x}$ 子向量. $\mathbf{Z}_f$ 表示函数 f 的参数空间的基,即参数集合,这里 $\mathbf{Z}_{g_i}=\mu(\delta(\mathbf{y}_i),\mathbf{a})$.因此有

$$g(\mathbf{x})=\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}]) \tag{7}$$

步骤 3.由式(7),

$$Q(\mathbf{a},\mathbf{x})=R(K(\mathbf{x}),\mathbf{a})+\gamma\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}]) \tag{8}$$

由式(6)和式(8),约束不等式组变为

$$f(K(\mathbf{x}))\geq R(K(\mathbf{x}),\mathbf{a})+\gamma\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}]),\ \forall\mathbf{x}\in\mathbf{X} \tag{9}$$

3.2 合并相似不等式

步骤 4.由于 K 为满射,因此 $\forall\mathbf{x}\in\mathbf{X}$, $\exists\mathbf{y}$, $\mathbf{y}=K(\mathbf{x})$,有

$$f(\mathbf{y})\geq R(\mathbf{y},\mathbf{a})+\gamma\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}]),\ \forall\mathbf{y}\in\mathbf{Y},\ \forall\mathbf{x}\in S(\mathbf{y}) \tag{10}$$

步骤 5.从式(10)可以看出,同一等价类中任意两个状态对应的不等式,除了 $\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}])$ 不同以外,其余各项均一致,因此可以将它们进行合并.即

$$f(\mathbf{y})-R(\mathbf{y},\mathbf{a})\geq\gamma\max_{\mathbf{x}\in S(\mathbf{y})}\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}]),\ \forall\mathbf{y}\in\mathbf{Y} \tag{11}$$

令 $T(\mathbf{y})=\max_{\mathbf{x}\in S(\mathbf{y})}\sum_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}])$.

由于 max 函数的存在,式(11)并不是标准形式的约束不等式组,因此需要将式(11)标准化.与 Gestrin 等人的做法类似,这里依然采用消元法,但由于是分类化简,因此过程更为复杂.基本思想是以其中某一状态属性 X_i 为变量,且固定其他状态属性,取最大值,并将其看成以其他状态属性为变量的函数,从而消去 X_i.同时,增加新的变量和新的约束不等式,直到消除所有的状态属性.完整的算法描述见算法 1.

算法 1. 约束不等式的消元算法.

输入:FMDP 模型,变量集合 $\mathbf{C}$,约束不等式集合 $\mathbf{\Omega}$.

1. 初始化

(1) $\mathbf{C}'=\mathbf{C}$; $\mathbf{\Omega}'=\varnothing$; $j=0$

(2) for i=1 to M $\mathbf{Z}_{g_i}=\mu(\delta(\mathbf{y}_i),\mathbf{a})$

//循环 $\forall \mathbf{a}\in\mathbf{A}$

2. R_0=M; $k=1$; $\boldsymbol{\mathcal{X}}=\{X_1,X_2,...,X_n\}$ // $M=|\mathbf{Y}|$

3. If $X_k\in\delta(\mathbf{y}_j)$ $R_k\leftarrow R_{k-1}$; $k\leftarrow k+1$;转 9

//如果 X_k 属性在最小特征属性向量 $\mathbf{y}_j$ 中,则直接消去

4. $T(\mathbf{y}_j)=\max\limits_{x_{k+1},...,x_n}\left\{\sum\limits_{i=1}^{R_k}h_i(\mathbf{x}[\mathbf{Z}_{h_i}]+\max\limits_{x_k}\sum\limits_{i=R_k+1}^{R_{k-1}}h_i(\mathbf{x}[\mathbf{Z}_{h_i}])\right\}$, $h_1,...,h_{R_{k-1}}$ 为 g_1 到 $g_{R_{k-1}}$ 的一个排列,且满足 $\mathbf{x}\in S(\mathbf{y}_j)$

和 $\begin{cases}\bigcup\limits_{i=1}^{R_k}\mathbf{Z}_{h_i}\cap\{X_k\}=\varnothing & (i<=R_{k-1})\\ \mathbf{Z}_{h_i}\cap\{X_k\}\neq\varnothing & (i>R_k)\end{cases}$

5. 令 $e(\mathbf{x}[\mathbf{Z}_e])=\max\limits_{x_k}\sum\limits_{i=R_k+1}^{R_{k-1}}h_i(\mathbf{x}[\mathbf{Z}_{h_i}])$,其中 $\mathbf{Z}_e=\bigcup\limits_{i=R_k+1}^{R_{k-1}}\mathbf{Z}_{h_i}\setminus\{X_k\}$.

6. 新增变量集合 $\mathbf{U}=\{u_{\mathbf{z}}^k\mid\mathbf{z}\in\mathrm{dom}(\mathbf{Z}_e\setminus\delta(\mathbf{y}_j))\}$, $\mathbf{C}'=\mathbf{C}'\cup\mathbf{U}$

7. 新增约束不等式组到 $\mathbf{\Omega}'$ 集合中,即 $u_{\mathbf{z}}^k\geq\sum\limits_{i=R_k+1}^{R_{k-1}}h_i(\mathbf{z}[\mathbf{Z}_{h_i}])$, $\forall\mathbf{z}\in\mathrm{dom}(\mathbf{Z}_e\setminus\delta(\mathbf{y}_j))$, $\forall x_k\in\mathrm{dom}(\{X_k\})$

8. //将 $h_1,...,h_{R_k}$ 函数重新命名,并修正 R_k 和 k 的值

(1) for i=1 to R_k, $g_i\leftarrow h_i$

(2) $g_{R_k+1}\leftarrow e$, $R_k\leftarrow R_k+1$, $k\leftarrow k+1$

9. If ($\boldsymbol{\mathcal{X}}\neq\varnothing$) $\boldsymbol{\mathcal{X}}=\boldsymbol{\mathcal{X}}\setminus\{X_{k-1}\}$,转 3

else if($j\neq M$) $j\leftarrow j+1$;转 2

与 Gestrin 等人得到的结果一样,经过算法 1 的合并和化简,约束不等式组的个数虽然仍然为指数规模,但指数因子却不同程度地减少了,在特殊情况下,还会化简为多项式规模.

3.3 举 例

以机器人足球赛为例(4 人制足球),为了简化起见,不妨设对手的联合策略是静态的,因此只需要考虑己方球员的策略.很显然,这是一个 FMDP 模型.由定义 3,其模型为 $\langle\mathbf{X},\mathbf{A},K,\{P_i\},\mu,R\rangle$,其中,

(1) $\boldsymbol{\mathcal{X}}=\langle X_1,...,X_4,X_5\rangle$, $X_1,...,X_4$ 分别为 4 名队员的位置坐标, X_5 为当前持球队员的编号.不妨设每个队员的位置有 m 个取值,即 $\mathrm{dom}(\{X_i\})=\{z_i^1,...,z_i^m\}$ (i≤4).

(2) $\mathbf{A}$ 为联合动作集合,分别代表传球、射门和奔跑等动作,设有 5 个可选个体动作.

(3) 最小特征向量函数: $K(\mathbf{x})=\langle x_5,x_{x_5}\rangle$,表示状态的最小特征向量为持球队员的编号和持球队员的位置.

(4) $\mu(\{X_i\},\mathbf{a})=\{X_i\}$,表示无论采取什么动作,Agent 下一状态的坐标位置只与自身当前坐标位置有关.

化简过程:

根据化简公式,不妨设 $\mathbf{y}=\langle 1,z_1^1\rangle$,对于 $\mathbf{x}\in S(\mathbf{y})$, $x_5=1$, $x_1=z_1^1$, $T(\mathbf{y})=\max\limits_{\mathbf{x}\in S(\mathbf{y})}\sum\limits_{i=1}^{M}g_i(\mathbf{x}[\mathbf{Z}_{g_i}])$, $\mathbf{Z}_{g_i}\in\{\{X_5,X_1\},\{X_5,X_2\},\{X_5,X_3\},\{X_5,X_4\}\}$,因此可以将具有相同参数基的函数合并,又因为 $x_5=1$ 且 $x_1=1$,则有

$$T(\mathbf{y})=\max_{x_2,x_3,x_4}(g_2'(1,x_2)+g_3'(1,x_3)+g_4'(1,x_4)) \tag{12}$$

根据算法 1,先消去 X_2,式(12)可变为

$$T(\mathbf{y})=\max_{x_3,x_4}(g_3'(1,x_3)+g_4'(1,x_4)+\max_{x_2}g_2'(1,x_2)) \tag{13}$$

令 $e_2=\max_{x_2} g_2'(1,x_2)$,则式(13)可写成

$$T(\mathbf{y})=\max_{x_3,x_4}(g_3'(1,x_3)+g_4'(1,x_4))+e_2 \tag{14}$$

同时增加变量集合 $\mathbf{U}=\{u_1^{e_2}\}$,增加约束不等式组:

$$u_1^{e_2}\geq g_2'(1,x_2)\quad \forall x_2\in \mathrm{dom}(\{X_2\}),$$

同理可以消去 X_3,X_4,增加变量集合 $\{u_1^{e_3},u_1^{e_4}\}$,增加约束不等式组:

$$u_1^{e_3}\geq g_2'(1,x_2)+e_2\quad \forall x_3\in \mathrm{dom}(\{X_3\}),$$

$$u_1^{e_4}\geq g_2'(1,x_2)+e_2+e_3\quad \forall x_4\in \mathrm{dom}(\{X_4\}).$$

由最小特征向量函数的定义可知,最小特征属性向量为 $4m$ 个,因此可以计算出最后在 $\boldsymbol{\Omega}'$ 中约束不等式的个数为 $5\times 4m\times 3m=60m^2$,即 $O(m^2)$,而未经化简在 $\boldsymbol{\Omega}$ 中的初始约束不等式个数为 5×4^m,即 $O(4^m)$.也就是说,计算复杂度会根据对球场的划分按指数规模增长,然而通过化简过程,计算复杂度可以降低为多项式规模,如图 1 所示.

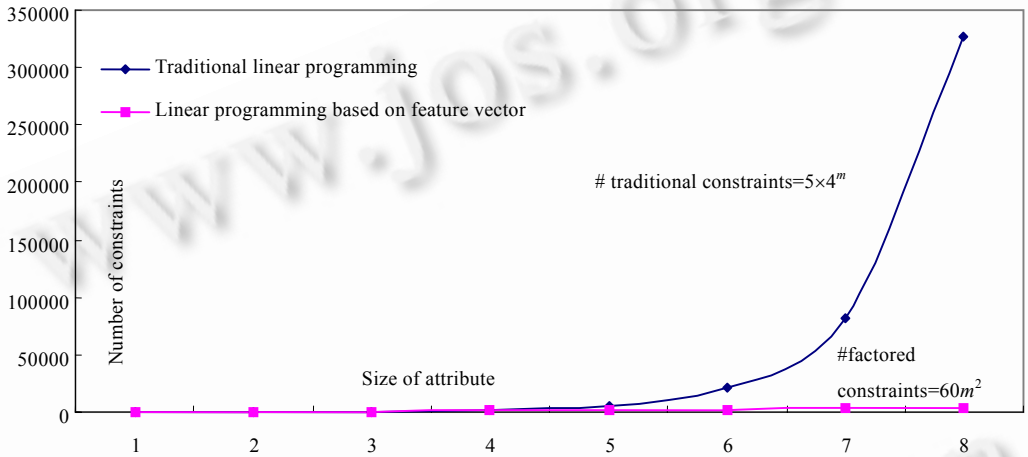

Fig.1 The reduction result of constraints in 4v4 soccer match

图 1 4 人制足球赛的约束不等式组化简结果

4 线性规划的近似求解

4.1 传统的Q学习算法

传统的 Q 学习算法描述如下:

(1) 初始化, $\forall \mathbf{x}$, $\forall \mathbf{a}$,令 $Q(\mathbf{x},\mathbf{a})=Q_0$,设定初始状态 $\mathbf{x}_0$, $k=0$;

(2) 在 k 时刻,当前状态为 $\mathbf{x}$ 时,根据ε-Greedy 算法[14]选择球队合适的联合动作 $\mathbf{a}$,转入到下一个状态 $\mathbf{x}'$;

(3) 修正 Q 值, $Q(\mathbf{x},\mathbf{a})=R(\mathbf{x},\mathbf{a})+\lambda Q(\mathbf{x},\mathbf{a})+\gamma(1-\lambda)\max_{\mathbf{a}} Q(\mathbf{x}',\mathbf{a})$;

(4) $k\leftarrow k+1$, $\mathbf{x}\leftarrow \mathbf{x}'$;

(5) if $k\geq T$ 转(6);else 转(2);

(6) $\forall \mathbf{x}\in \mathbf{X}$,联合策略 $\pi(\mathbf{x})=\arg\max_{\mathbf{a}} Q(\mathbf{x},\mathbf{a})$.

可以看到,对于状态呈指数规模增长的群体 Agent 合作求解问题,需要遍历所有状态的传统 Q 学习算法,使得学习速度异常缓慢,基本不能应用到诸如机器人足球赛等状态空间庞大的群体 Agent 合作求解问题中.

4.2 K-Q再励学习算法

K-Q 算法是在传统再励学习算法的基础上,每次只修正最小特征向量的效用值.但对于不同类型的 FMDP 模型,其特征向量的提取是不一致的,会影响具体算法的流程.因此,这里仅以机器人足球赛为例,在对手采取静态策略的前提下,介绍该算法.

设状态特征向量 $\mathbf{y}=\langle y_1,y_2,y_3\rangle$,其中 y_1 为当前持球队友的编号,y_2 为当前持球队友的坐标,y_3 为当前持球队友附近的对手个数(如果大于 3,则 $y_3=3$).因此,对于任意状态 $\mathbf{x}$,都存在唯一的最小特征向量 $\mathbf{y}=K(\mathbf{x})$.令 $\mathrm{p}_i(\mathbf{x})$ 为在状态 $\mathbf{x}$ 下,第 i 名队友周围的对手个数,令 $f(\mathbf{y},\mathbf{a})=Q(\mathbf{x},\mathbf{a})$.设球队共有 n 名球员,单 Agent 可选动作集合为 $\mathbf{A}$.由于机器人足球比赛的特殊性,Agent 的位置只能通过自身的动作改变,因此 $f(\mathbf{y},\mathbf{a})=f(\mathbf{y},a_k)$,其中 $k=y_1$,即为持球队员的动作.完整算法描述如下:

算法 2. K-Q 算法.

1. 初始化,$\forall\mathbf{y}\in\mathbf{Y}$,$\forall a\in\mathbf{A}$,令 $f(\mathbf{y},a)$=Q_0,设定初始状态 $\mathbf{x}=\mathbf{x}_0$,k =0 //循环
2. 在 k 时刻,$\mathbf{y}\leftarrow K(\mathbf{x})$
3. for i=1 to n
   if ($i\neq y_1$) //非持球队员
   (1) $\mathbf{z}=\langle i,x_i,\mathrm{p}_i(\mathbf{x})\rangle$
   (2) 根据ε-Greedy 算法,对第 i 位队友从 $f(\mathbf{z},a)(a\in\mathbf{A}\setminus\{$传球,射门$\})$ 中选择合适的动作
   (3) $f(\mathbf{y},c_i)=\max\limits_{b\in\mathbf{A}\setminus\{\text{传球,射门}\}}\{f(\mathbf{z},b)\}$ // $c_i\in\mathbf{A}$,表示传球给第 i 名队友
   else
   根据ε-Greedy 算法,对持球队友从 $f(\mathbf{y},a)(a\in\mathbf{A})$ 中选择合适的动作,记作 b.
4. 执行球队的联合动作,进入下一状态 $\mathbf{x}'$,令 $\mathbf{y}'=K(\mathbf{x}')$
5. $f(\mathbf{y},b)=R(\mathbf{y},b)+\lambda f(\mathbf{y},b)+\gamma(1-\lambda)\max\limits_{a\in\mathbf{A}\setminus\{\text{传球,射门}\}}f(\mathbf{y}',a)$ //修正效用值
6. $k\leftarrow k+1$; $\mathbf{x}\leftarrow\mathbf{x}'$
   if $k\geq T$ 转 7;else 转 2
7. for all $\mathbf{x}\in\mathbf{X}$,令 $\mathbf{y}=K(\mathbf{x})$
   for i=1 to n
   (1) $\mathbf{z}_i=\langle i,x_i,\mathrm{p}_i(\mathbf{x})\rangle$;
   (2) $f(\mathbf{y},a_i)=\max\limits_{b\in\mathbf{A}\setminus\{\text{传球,射门}\}}\{f(\mathbf{z}_i,b)\}$

令 $\pi(\mathbf{x})=\mathbf{a}$,其中 $a_i=\begin{cases}\arg\max\limits_{a\in\mathbf{A}\setminus\{\text{传球,射门}\}}f(\mathbf{z}_i,a) & i\neq y_1\\ \arg\max\limits_{a\in\mathbf{A}}f(\mathbf{y},a) & i=y_1\end{cases}$.

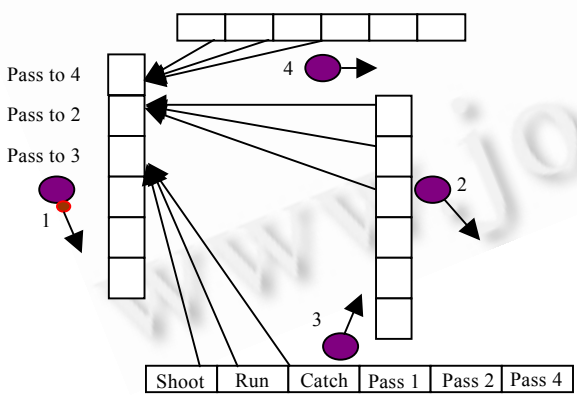

Fig.2 Computing value function of feature vector
图 2 特征向量的效用函数计算

从算法 2 中可以看出,与传统 Q 学习算法不同,在每个状态下,K-Q 算法只考虑特征向量,因此每个特征向量中有关传球的 Q 值并不需要储存,也无法事先得知,而是在比赛过程中动态计算出其他球员的最大 Q 值,并以此代替相应的 Q 值以协助决策.如图 2 所示.

在第 1 节中已经提到,在 FMDP 模型,包括状态转移概率函数都已知的前提下,其线性规划求解形式经过化简过程后,能够有效降低状态空间规模,最终达到加快求解最优联合策略的目的.然而实际复杂和多变的群体 Agent 环境和 Agent 有限的知识无法准确预测 FMDP 的完整结构,只能通过对外部环境的观察和再励学习等动态规划算法,尝试出最优策略.

将特征向量的提取引入到再励学习算法中,可以加快学习速率,从而达到快速收敛的目的.下面首先介绍传统的再励学习算法 Q-learning[13],然后以机器人足球赛为例,给出基于状态特征向量的改进算法.

5 实验结果

前面已经提到,机器人足球赛作为典型的 FMDP 模型,可以采用提取特征向量近似状态效用函数的思想,嵌入到线性规划和再励学习两种求解方法中进行求解.本节以机器人足球赛中任意球战术配合为例,运用基于特征状态属性的再励学习算法,验证其有效性和学习结果的可移植性.

图 3 是机器人足球赛的任意球比赛画面(5 人制),规定如果本方进球,则本回合战术配合成功,本方加 1 分;否则,如果球被对手截下,或球被对方守门员没收,则本回合战术配合失败,给对手加 1 分.最后统计战术配合的成功率,以此作为算法执行效果的唯一判据.假定对手策略是静态、稳定的.比赛过程中,每名队员可以向 8 个方向奔跑,持球队员还有传球给周围队友,或者直接射门等动作.从第 4 节可知,传统的 Q 算法,其所占空间随着队员个数的增加呈指数增长;而 K-Q 算法由于只考虑状态空间中的状态特征向量,实际所占的内存空间只随着队员个数的增加呈线性增长,如图 4 所示.特征向量的提取为解决机器人足球赛中由于大状态空间而学习缓慢的问题奠定了很好的基础.

图 3 为 5 人制任意球战术配合的学习结果图,可以看出,由于状态空间过大,Q 算法基本达不到学习的效果,而 K-Q 算法通过提取状态特征向量来近似状态效用函数,可以在很短时间内达到非常理想的学习效果.图 5 是将从图 3 学习到的最小特征向量的效用函数直接用在 7 人制和 9 人制的任意球战术配合中,得到的测试结果图.从中可以看出,从低维状态空间学习到的状态效用函数可以直接移植到有更多 Agent 参与的比赛中,同样能得到良好的学习效果,从而省去重新学习的巨大开销.

6 结 语

大状态空间的“维数灾”一直是求解 FMDP 模型普遍存在的问题,与以往普遍采用的方法不同,本文面对整体效用无法分散到个体的群体 Agent 合作求解问题,从基于特征向量的 FMDP 模型入手,通过对状态特征向量的提取来近似状态效用函数,并根据对 FMDP 模型,包括状态转移概率函数等的认知程度,从线性规划和再励学习两个求解角度分别进行约束不等式组的化简和状态效用值的高维移植,从而达到降低计算复杂度、加快联合策略生成速度的目的.为了验证 K-Q 算法的有效性和学习结果的可移植性,本文将 K-Q 算法应用到机器人足球赛任意球战术配合中,与传统的再励学习算法相比,K-Q 算法除了能够大幅度加快策略收敛速度之外,还能将学习到的状态效用函数直接移植到更多 Agent 参与的比赛中,直接计算出联合策略并取得良好的学习效果,从而省去重新学习所花费的巨大开销.

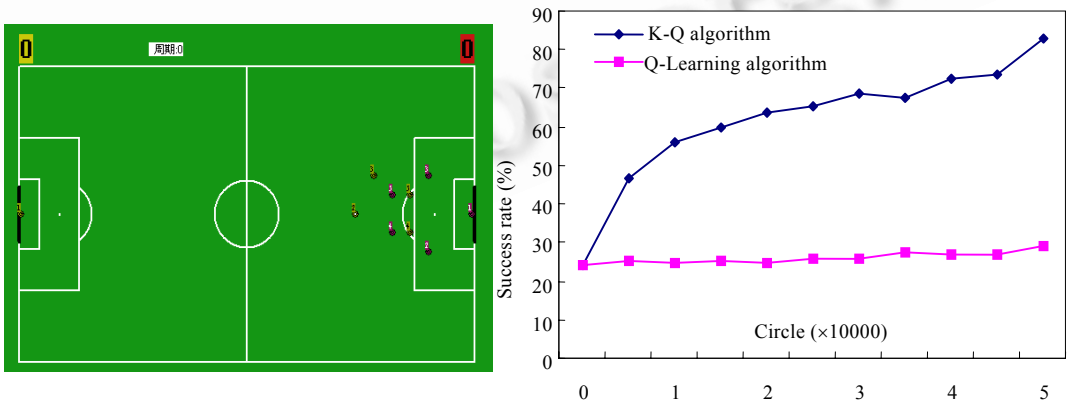

Fig.3 Free kick cooperation experiment of 5v5 soccer match

图 3 5 人制足球赛的任意球战术配合实验

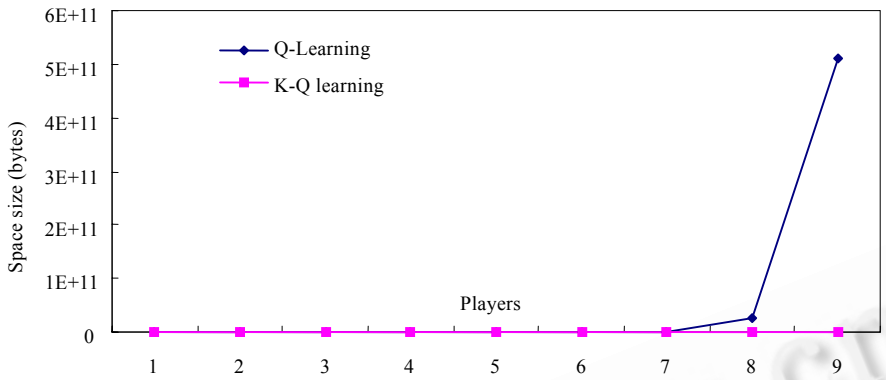

Fig.4 The state space comparison between two algorithms

图 4 两种算法所耗空间的比较

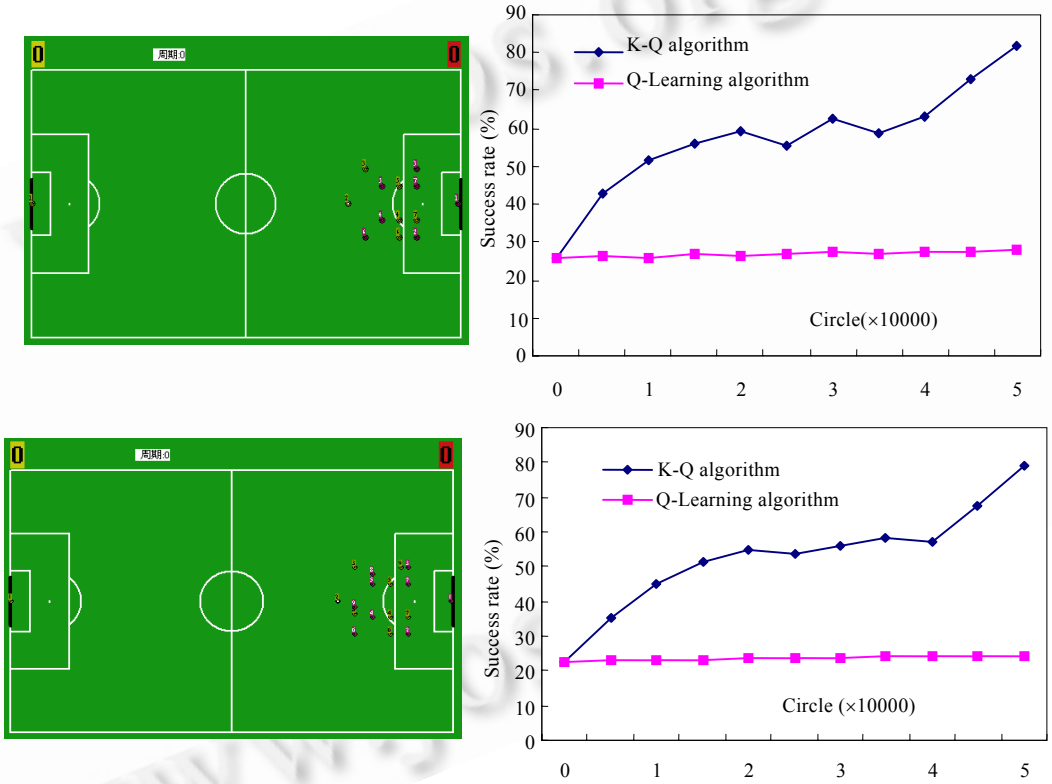

Fig.5 Free kick cooperation experiment of 7v7 and 9v9 soccer match

图 5 7 人制和 9 人制足球赛的任意球战术配合实验

References:

[1] Parr K. Policy iteration for factored MDPs. In: Proc. of the 16th Conf. on Uncertainty in Artificial Intelligence (UAI-00). Stanford, 2000. 326−334. http://ai.stanford.edu/~koller/papers/uai00kp.html

[2] Parr K. Computing factored value functions for policies in structured MDPs. In: Int'l Joint Conf. on Artificial Intelligence (IJCAI'99). Morgan Kaufmann Publishers, 1999. 1332−1339. http://ai.stanford.edu/~koller/papers/ijcai99kp.html

[3] de Farias R. Approximate dynamic programming via linear programming. In: Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2002. http://www.core.org.cn/NR/rdonlyres/Mechanical-Engineering/2-997Spring2004/DF5542A5-BBCC-4BAB-ADBF-41AB0FDA6F95/0/most_uhan_slides.pdf

[4] Guestrin CE, Venkataraman S, Koller D. Context specific multiagent coordination and planning with factored MDPS. In: AAAI-2002 The 18th National Conf. on Artificial Intelligence. Edmonton, 2002. 253−259. http://www-2.cs.cmu.edu/~shobha/research/aaai02.pdf

[5] Guestrin CE, Koller D, Parr R. Efficient solution algorithms for factored MDPs. Journal of Artificial Intelligence Research, 2003, 19:399−468.

[6] Guestrin CE, Koller D, Parr R. Multiagent planning with factored MDPs. In: Advances in Neural Information Processing Systems (NIPS-14). Vancouver, 2001. 1523−1530. http://robotics.stanford.edu/~koller/papers/nips01gkp.html

[7] Guestrin CE, Koller D, Gearhart C, Kanodia N. Generalizing plans to new environments in relational MDPs. In: Int'l Joint Conf. on Artificial Intelligence (IJCAI 2003). Acapulco, 2003. 1003−1010. http://web.engr.oregonstate.edu/~hamann/generalizing_plans_rmdp.pdf

[8] Sallans B. Reinforcement learning for factored Markov decision processes [Ph.D. Thesis]. Toronto: University of Toronto, 2002.

[9] Maes S, Tuyls K, Manderick B. Reinforcement learning in large state spaces: Simulated robotic soccer as a testbed. Lecture Notes in Artificial Intelligence, RoboCup 2002. Fukuoka: Springer-Verlag, 2002. http://como.vub.ac.be:8080/Publications/uploads/1/rl-robo02.ps

[10] Manderick TM. Q-Learning in simulated robotic soccer: Large state spaces and incomplete information. In: Proc. of the ICMLA 2002. Las Vegas, 2002. 226−232. http://como.vub.ac.be:8080/Publications/uploads/1/icmla02.ps

[11] Uther WTB, Veloso MM. Tree based discretization for continuous state space reinforcement learning. In: AAAI 2002 the 18th National Conf. on Artificial Intelligence. Madison, 1998. 769−774. http://www.cs.cmu.edu/~mmv/papers/will-aaai98.pdf

[12] Boutilier C, Dearden R, Goldszmidt M. Stochastic dynamic programming with factored representations. Artificial Intelligence, 2000, 121(1-2):49−107.

[13] Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 1996,4:237−285.

[14] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 1998.

****

2005年全国开放式分布与并行计算学术会议

征 文 通 知

由中国计算机学会开放系统专业委员会主办、上海大学计算机学院承办、上海计算机学会协办的“2005年全国开放式分布与并行计算学术会议”将于2005年10月27日－29日在上海召开，有关信息如下：

一、征文范围(包括但不限于)

开放式分布与并行计算模型及体系结构；

下一代开放式网络、数据通信、网络与信息安全、业务管理技术；

开放式海量数据存储与Internet索引技术，分布与并行数据库及数据/Web挖掘技术；

开放式机群计算、网格计算、Web服务、P2P网络及中间件技术；

开放式移动计算、自组网与移动代理技术；

分布式人工智能、多代理与决策支持技术；

分布与并行计算算法及其在科学与工程中的应用；

开放式虚拟现实技术与分布式仿真；

开放式多媒体技术，包括媒体压缩、内容分送、缓存代理、服务发现与管理技术。

二、征文要求

详见会议主页：http://www.cs.shu.edu.cn/DPCS2005 可查询进一步的会议信息。

三、重要日期

会议时间：2005年10月27～29日　　截稿日期：2005年7月15日　　录用通知：2005年7月30日

四、联系方式

投稿地址：200072 上海延长路149号上海大学计算机学院 缪淮扣 收(请在信封上注明DPCS2005)

电　　话：021-56331669;　　电子邮件: bfzhang@staff.shu.edu.cn (请在邮件主题中注明DPCS2005)