一种基于划分和集成思想的多智能体强化学习

王 云, 韩 伟

(南京财经大学 信息工程学院, 江苏 南京 210046)

[摘要] 针对 Q 学习状态空间非常大, 导致收敛速度非常慢的问题, 利用智能体在不同样本上分类性能不同, 提出了基于样本的学习误差对样本空间进行划分, 充分发掘了样本和智能体的匹配关系. 以带障碍物的格子世界作为仿真环境, 表明该算法提高了在线学习性能.

[关键词] 多智能体系统, 强化学习, 状态空间划分

[中图分类号] TP 301 [文献标识码] A [文章编号] 1672-1292(2008)04-0059-04

An Multiagent Reinforcement Learning Based on Partition and Integration

Wang Yun, Han Wei

(Information Science and Engineering College, Nanjing University of Financial and Economics, Nanjing 210046, China)

Abstract To counter for the problem of slowly convergence of Q leaning when comeing to large state-space, the paper puts forward an algorithm which divide the states space according to learning errors. The basic idea of our algorithm is to discover the matching relationship between agents and the sub-space of states space. The simulations in grids with blocks indicate that the algorithm performs better when comeing to on-line learning.

Key words multiagent system, reinforcement learning, state-space partition

自治智能体的学习问题是多智能体系统一个重要研究内容. Q 强化学习将智能体与环境的交互看作一个 Markov决策过程, 是一类具有广泛应用价值的学习算法. 许多工作基于 Q 强化学习提出相应的扩展算法, 刘海涛等提出不确定环境下基于进化算法的 Q 强化学习方法[1]; 韩伟等提出基于内省推理的 Q 学习方法[2], 并针对电子市场提出基于 Q 强化学习的智能定价算法[3]. 单个 agent的 Q 强化学习要求 agent遍历整个状态 - 动作空间才能使估计 Q 值收敛到真实 Q 值. 但是这个收敛条件在实际复杂的情形中几乎无法满足, 通常情况下, 状态空间是很大的, 因为随着环境特征数量的增加, 可能出现的状态的数量也会成指数地增加. 这导致在涉及大的特征空间的随机环境中, 强化学习的收敛速度非常慢, 若采用单个 agent学习, 需要在大量样本上进行训练才能得到较满意的 Q 估计值. 这限制了 Q 学习的实际应用价值.

由 Q 学习的更新规则 $Q(s, a) \leftarrow r(s, a) + \gamma \cdot \max_{a'} Q(\delta(s, a), a')$ 可以看出, 更新后的 Q 值既取决于训练样本 (s, a), 又取决于样本 (s, a) 附近样本的 Q 值. 比如, 若格子世界中有两个智能体, 一个从目标状态附近开始学习, 另一个从左上角的位置开始学习, 经过同样数量的样本的训练 (比如为 5), 得到的 Q 估计值与实际值的误差并不相同. 这启发我们为了得到较精确的 Q 值, 必须将正确的样本分配给在该样本附近的样本上具有较正确 Q 值的智能体进行训练 (假设环境中有多个智能体), 即采用 Q 学习的智能体的学习能力在样本空间上具有一定的分块特性, 在某个区域上比较强, 而在某些区域上比较弱. 我们据此提出基于状态空间划分的多智能体合作学习方法, 通过训练一个两层的学习网络 (内层由每个单独的具有学习能力的智能体组成, 外层则根据误差平方和最小的原则调整每个智能体的决策权重) 得到每个区域的样本在智能体上的权重信息, 以此作为先验信息来预测新状态的 Q 值来提高在线学习性能.

收稿日期: 2008-06-18
基金项目: 国家自然科学基金 (70802025)资助项目.
通讯联系人: 王云, 讲师, 研究方向: 电子商务、人工智能. E-mail: dallashw@gmail.com

1　多智能体 Q 表集成

设环境中存在多个采用 Q 学习算法的智能体, 每个智能体单独维护一张 Q 表. 由于每个智能体的训练序列不同, 因此即使是在相同的学习步内得到的 Q 表项也不一致. 文献 [4] 提出集成不同 Q 表项的方法来预测新的 Q 表值, 智能体在每个样本具有相同的预测权重. 事实上, 由于状态空间具有分块特性, 使得智能体某个状态子空间上预测能力有强有弱. 所以仅以智能体在样本上的平均表现来决定其权重还是不够的, 我们有必要发掘智能体与状态子空间的匹配关系, 据此将状态空间划分成若干子空间并计算每个智能体在各个子空间上的预测权重. 即经过训练之后, 对于特定的状态 – 动作转换, 我们能够得到每个智能体在这个转换上的决策权重, 并以此作为依据对新的样本决定输出. 假定环境中有 n 个反应式智能体 $A_1, \cdots A_n$, 对于给定状态 – 动作转换 $x = (s, a)$, $A_i (i = 1 \cdots n)$ 的输出为 $a_i(x)$, 定义误差函数

$$\text{error}' = \sum_x \text{error}'(x) = \sum_x \sum_{i=1}^{n} w_i(x)(y(x) - a_i(x))^2 = \sum_x \left(y(x) - \sum_{i=1}^{n} w_i(x) a_i(x) \right)^2 + \sum_x \sum_{i=1}^{n} \left(a_i(x) - \sum_{i=1}^{n} w_i(x) a_i(x) \right)^2, \quad (1)$$

(1) 式定义的误差函数可以看作两部分误差组成, 前半部分表示全局误差, 后半部分表示局部误差. 若将每个智能体看作反应式的, 为了降低决策误差, 进一步可以用梯度下降法对每个智能体的决策权重进行优化.

$$\Delta w_i(x) \propto \frac{\partial \sum_x \sum_{i=1}^{n} w_i(x)(y(x) - a_i(x)^2}{\partial w_i(x)}, \quad (2)$$

若我们理想的假定智能体的权重与每个特定样本相关, 则整个样本集上的误差为

$$\text{error} = \sum_{i=1}^{n} \text{error}_i = \sum_{i=1}^{n} \sum_{x \in S} w_i(x)(y(x) - a_i(x))^2, \quad (3)$$

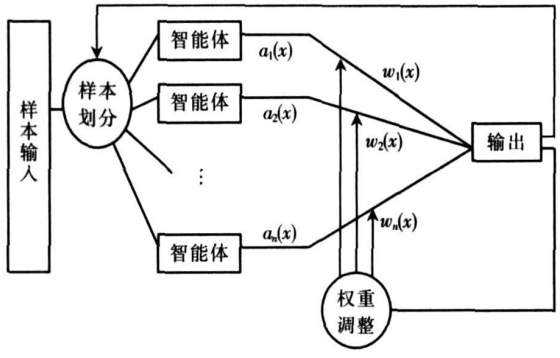

图 1　基于样本空间划分的多智能体学习系统

Fig.1　Multi-agent learning system based on sample space patition

若将训练样本集 S 划分为 m 个等价类, 即 $\bigcup_j S_j = S$, 且 $S_i \cap S_j = \Phi (i \neq j)$. 且对于每个等价类上的任意两个样本, 每个智能体的权重大体相等, 即对 $\forall x_1, x_2 \in S_j (j = 1 \cdots m)$, $w_i(x_1) \approx w_i(x_2)$ $(i = 1 \cdots n)$. 于是, 样本划分后, 整个训练样本集上的误差为

$$\text{error} = \sum_{i=1}^{n} \text{error}_i = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{x \in S_j} w_i(x)(y(x) - a_i(x))^2 \approx \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{x \in S_j} w_{ij}(y(x) - a_i(x))^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} \sum_{x \in S_j} (y(x) - a_i(x))^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij} e_{i, s_j}, \quad (4)$$

首先对状态空间进行合理的离散化处理 (在每个维度上选定若干可能的划分点, 最简单的办法是等分该维度上的值域), 然后根据最小化总误差 $\sum_{i=1}^{n} \sum_{j=1}^{m_p} w_{ij} e_{i, s_j}$ 的原则对状态空间进行划分.

训练后得到的外部决策权重可以看作是每个智能体在整个训练样本集上的平均表现, 即样本在智能

体上的匹配信息. 而样本划分可以看作是单个智能体在样本空间上的表现差异, 即智能体在样本上的匹配信息.

2　多智能体的 Q 学习

若样本子空间的数目太大, 则样本划分后每个子空间上的训练样本数量太小导致外部权重和内部权重调整不大, 划分就失去了意义. 因此, 对于特定问题, 我们总是控制子空间的数目, 一般与环境中的智能体数目相当. 另外, 为了充分利用样本的全局信息, 在每个子空间上的训练开始之前, 总是保留在上一级空间上的外部权重和内部权重作为初始权重. 对于 Q 学习, 我们选取误差函数为 Bellman误差, 即 $error(s, a) = |Q^t(s, a) - Q^{t-1}(s, a)|$.

表 1　基于 Q学习的多智能体离线学习(MLPR)

Table 1　Off-line multiagent learning based on partition and integration

将整个训练样本集在所有智能体上进行训练, 同时调整外部决策权重, 最终得到权重向量 $w^{(R)}$

(1)重复.

① for每个子空间 R_p

for每个维度 v_q

for每个该维度上的有效值 x_k

{ i将空间 R_p 以超平面 $v_q = x_k$ 划分, 分别用划分后的子空间上的样本训练所有智能体, 同时调整外部决策权重, 得到智能体 i 在各个子空间下的权重 w_{ij};

ii根据式(4)计算总误差;

}

② 根据最小误差, 确定子空间 R_p^*、维度 v_q^*、有效值 x_k^*. 根据超平面 $x_i = x_k^*$ 对样本子空间 R_p^* 划分.

(2)直到划分得到的子空间达到一定数目或者误差变化小于一定阈值.

经过上述算法, 我们得到与每个子空间 R_j 相联系的特定权向量 $w^{(R_j)}$. 对于训练后的新样本 (s, a), 首先判断其所处的状态子空间 R_j, 然后采用 $w^{(R_j)}$ 集成各个智能体的 Q 表项. 一个实用的方法是将在线学习和离线学习相结合, 每个智能体的内部权重更新可以在线进行. 在经过一段时间的在线学习之后, 用累积存储的样本重新对样本空间进行划分, 得到每个子空间及其权向量.

3　仿真实验与结果分析

实验环境为图 2所示的格子世界, 其中 $L(s)$ 为状态 s 到目标状态的街区距离. 随机选取 100个状态 – 动作对作为训练样本, 用来训练 3个智能体直到划分的子空间数目超过 3 训练完毕后, 得到图 2不同颜色所示样本空间的一个划分及权重向量. 对 5个子空间进行合并 (按照 3个智能体权重大小), 得到阴影所示 3个子空间: 在下斜线表示的阴影区域, 智能体 1的权重最大, 权重向量为 (0.91, 0.08, 0.01); 在上斜线表示的阴影区域, 智能体 2的权重最大, 权重向量为 (0.83, 0.15, 0.02); 在平线表示的阴影区域, 智能体 3的权重最大, 权重向量为 (0.01, 0.09, 0.90).

训练结束后, 根据样本所处子空间决策权重, 得到 1条从 S处开始的决策路径 (如图 2虚线所示). 实验测得的平均收益为 73.35, 10次中有 4次智能体到达了目标状态.

若去除智能体的权重信息, MLPR退化为无权重的 Q表集成方法, 此时测的 10次决策平均收益 56.17, 10次仿真智能体均未到达目标状态.

4　结语

Q 强化学习将智能体与环境的交互看作一个 Markov决策过程, 是一类具有广泛应用价值的学习算法. 但是许多应用中, Q 学习状态空间非常大, 导致收敛速度非常慢的问题, 这限制了强化学习的实际应用. 本文利用智能体在不同样本上分类性能不同, 提出基于样本的学习误差对样本空间进行划分, 充分发掘了样本和智能体的匹配关系.

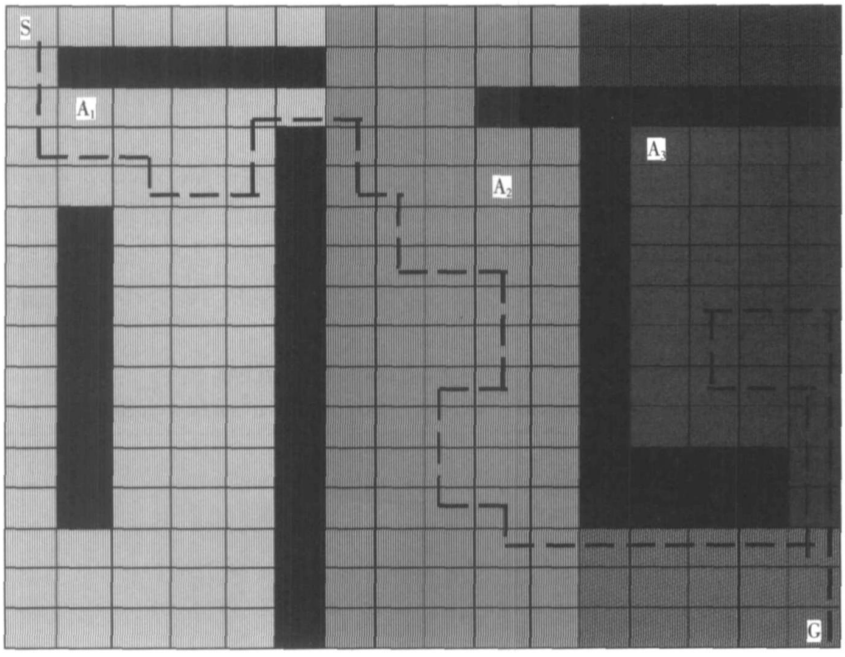

图 2　离线学习得到的样本空间的划分情况,虚线表示从 s 处开始的一条决策路径

Fig.2　A partition of sample space, dot line presents an decision path from s

[参考文献](References)

[1] 刘海涛, 洪炳熔, 朴松昊, 等. 不确定环境下基于进化算法的强化学习[J]. 电子学报, 2006, 7(34): 1356-1360.
Liu Haitao, Hong Bingrong, Pu Songhao, et al. Evolutionary algorithm based reinforcement learning in the uncertain environments[J]. Acta Electronica Sinica, 2006(34)7: 1356-1360. (in Chinese)

[2] 韩伟, 陈优广, 姜昌华. 基于内省推理的多 agent在线学习新方法[J]. 模式识别与人工智能, 2007, 20(2): 254-260.
Han Wei, Chen Youguang, Jiang Changhua. An Internal- Inference Based Multiagent Learning Method[J]. Pattern Recognition & Artificial Intelligence, 2007, 20(2): 254-260. (in Chinese)

[3] 韩伟. 基于情节序列训练的电子市场智能定价算法[J]. 计算机工程与应用, 2007, 43(6): 17-19.
Han Wei. Intelligent pricing algorithm based on multiagent learning[J]. Computer Engineering and Applications, 2007(43)6: 17-19. (in Chinese)

[4] 文益民, 杨旸, 吕宝粮. 集成学习算法在增量学习中的应用研究[J]. 计算机研究与发展, 2005, 42: 222-227.
Wen Yimin, Yang Yang, Lu Baoliang. Rescearch of the application ensemble learning algorithms to incremental learning[J]. Journal of Computer Research and Development, 2005, 42: 222-227. (in Chinese)

[责任编辑: 孙德泉]