DOI: 10.3969/j.issn.1001 - 3881.2010.20.026

基于强化学习的混合智能控制算法研究与分析

陈玉明，张广明，赵英凯
（南京工业大学自动化与电气工程学院，江苏南京 210009）

摘要： 设计混合智能控制结构，该结构引入强化学习和神经网络，提出基于 BP 神经网络的 Q 学习算法，优化动作的选取，解决传统 Q 学习中 Q 表占用内存空间过大的问题，增强系统的泛化能力。将其应用到 Predator-prey 模型中。实验结果表明，系统无需每次从全部动作中选择，从而大大缩小了状态 - 动作对的数量，节省计算时间，为智能体最优策略的选择提供更大的可能性。

关键词： 智能体；强化学习；神经网络；Markov 决策过程
中图分类号： TP181 **文献标识码：** A **文章编号：** 1001 - 3881（2010）20 - 075 - 3

Q-Learning Reinforcment Learning Algorithm Based on Neural Network

CHEN Yuming, ZHANG Guangming, ZHAO Yingkai
(Institute of Automation and Electric Engineering, Nanjing University of Technology, Nanjing Jiangsu 210009, China)

Abstract: Intelligent mixed control structure with reinforcement learning and neural network was designed. The method, called Q-learning based on BP neural network, optimized action selection. The problem that the Q-table of the traditional Q-learning occupied too much memory was solved. This method also improved the generalization capability of the system. The intelligent mixed structure was applied to the model of predator-prey. The experimental result indicates it need not select action in the whole actions each time, so the quantity of state-action pair is reduced and the time costed to learn is decreased. It supplies more possibility for the selection of optimization policy for agent.

Keywords: Agent; Reinforcement learning; Neural network; Markov decision process

基于行为的 Agent 能直接完成从感知到行为的映射，具有快速执行性和灵活性，已成为人工智能领域的研究热点之一。传统的反应式 Agent 研究方法通常基于具体的环境模型，存在环境知识获取困难、环境模型难以建立、自适应能力差等问题。强化学习具有不依赖于环境模型、不需要先验知识以及鲁棒性强等优点，已成为基于行为的 Agent 研究的一个新的方向[1]。

1 强化学习

学习算法基本上可以分为 3 种类型：非监督学习、监督学习和强化学习。条件反射原理属于非监督学习，属纯开环的学习方法。监督学习规则是一种反馈学习规则，依据理论的输出信号，系统根据输出误差来指导学习过程，无疑好于非监督学习。强化学习，应用了人类适应环境的学习过程，它把学习看成一个试探 - 评价的过程：强化学习系统通过感知，得到环境状态，并采取某一行动作用于环境；环境接受动作并使状态发生变化，同时系统给出强化信号（奖励或惩罚），反馈给强化系统，表示刚才所做动作的评价；系统根据强化信号和环境当前的状态选择下一动作，选择的原则是使受到正强化的概率增大。当然选择的动作影响立即强化值，同时影响下一状态以及最终强化值[2-3]。

2 基于 Q 学习的强化学习理论

在马尔可夫决策过程（MDP）中，Agent 所在的环境描述为状态集合 $S=\{s_i \mid s_i \in S\}$，它可执行的动作集合表示为：$A=\{a_i \mid a_i \in A\}$，Agent 在状态 s_i 下，选择动作 a_i 并执行，此时状态转移到 s_{i+1}，并从环境得到强化信号。强化学习的任务是得到一个控制策略 $\pi: S \rightarrow A$，使状态 - 动作序列的累积回报最大。即：

$$V(S_t)=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots=r_t+\gamma V(S_{t+1}) \quad (1)$$

式中：λ 为折扣因子。Q 学习是一种重要的强化学习算法，它不直接应用上面的值函数，而是利用一个类似的 Q 函数，其表达式如下：

$$Q(s,a) \leftarrow r_t+\gamma V(S_{t+1}) \quad (2)$$

式中：a 是时刻 t 从动作集 A 被选中的动作。由于系

收稿日期：2009 - 10 - 14
基金项目：江苏省自然科学基金项目（BK2006176）
作者简介：陈玉明（1979—），男，讲师，研究方向为智能控制、机器学习。电话：13813994307，E - mail：njutcym@sina.com.cn。

统的目的是使累积奖励值为最大，因此用$\max_{a\in A}Q(s_{i+1},a)$取代式中的$V(S_{t+1})$，得到表达式:

$$Q(s,a)\leftarrow r_t+\gamma\max_{a\in A}Q(s_{i+1},a) \quad (3)$$

在 t 时刻，智能体根据当前所处的状态选择一个动作 a，然后根据以下的表达式来更新 Q 值

$$Q(s_i,a_i)\leftarrow Q(s_i,a_i)+\alpha[r_i+\gamma\max_{a\in A}Q(s_{i+1},a_{i+1})-Q(s_i,a_i)] \quad (4)$$

式中: α 为学习率。式 (4) 使用下一状态的估计来更新 Q 函数，称为一步 Q 学习[4-5]。Q 学习算法的程序表达式如下:

Initialize Q (s, a) arbitrarily
Repeat (for each episode)
Initialize s
Repeat (for each step of episode)
Choose a from s using policy derived from Q (e. g. , ε-greedy)
Take action a, observer r, s´
$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)]$
s←s´
Until s is terminal

3 强化学习与神经网络的混合结构

传统的强化学习算法，如 Q 学习算法利用表格来表示 $Q(s,a)$ 函数，这种方法简单且计算效率高。当状态集合 S、系统动作集合 A 较大时，该方法需要占用大量的内存空间，而且也不具有泛化能力。将强化学习和神经网络相结合，主要是利用神经网络的强大存储能力和函数估计能力。一般来说，神经网络在系统中的工作方式[6]是: 接收外界环境的完全或不完全状态描述，作为神经网络的输入，并通过神经网络进行计算，输出强化学习系统所需的 Q 值，网络的输入对应描述环境的状态。采用神经网络实现 Q 学习克服了传统 Q 学习存在的问题，在较大程度上发挥这两种技术各自特有的优势。作者提出了基于神经网络的 Q 学习，即用神经网络来逼近 Q 函数，从而可以克服图表存储 Q 值所存在的缺陷。

应用一个 3 层的 BP 神经网络，BP 算法是由两部分组成: 信息的正向传递与误差的反向传播。网络的输入为状态矢量 S，网络的输出为每一状态下可选动作的 Q 值，即 $Q(s,a)$。由式 (4) 可知，每次执行一个动作后的 Q 值会更新，其 Q 值的变化为:

$$\Delta Q=\alpha[r_i+\gamma\max_{a_i\in A}Q(s_{i+1},a_{i+1})-Q(s_i,a_i)] \quad (5)$$

而 ΔQ 可以看作是 BP 网络输出层的误差，利用 BP 算法的误差反向传播就可以调整权值，保存调整后的 Q 值，从而实现 Q 值的学习。其框图如图 1 所示。

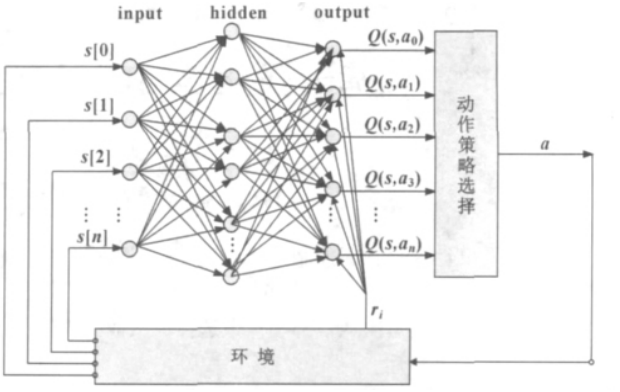

图 1 基于神经网络的 Q 学习算法结构图

4 行为选择策略

在学习的初始阶段，由于其 Q 值是随机初始化的，为了减少不确定的选择，能够探索到所有可能的动作，作者引入模拟退火搜索策略来实现初始阶段对动作的随机选择。设某一个动作被选中的概率为:

$$P(a_i|s_i)=\frac{e^{Q(s_i,a_i)/T}}{\sum_{a_i\in A}e^{Q(s_i,a_i)/T}} \quad (6)$$

式中: A 是在状态 S_i 可用的动作集。温度参数 T 平衡探索和利用，T 越大，探索的概率就越大，在学习过程中，T 逐渐减小。随着学习的进行，Q 值慢慢趋向于所期望的状态 - 动作值，这时候根据贪婪策略来选择动作，即选择最大 Q 值所对应的动作

$$a=\arg\max_{a_i\in A}Q(s_i,a_i) \quad (7)$$

5 仿真实验

用 Predator-prey 模型进行仿真实验[7]，Cheese、Cat 和 Mouse 在 8 × 8 的方格中，其中有一些障碍物，如图 2 所示。

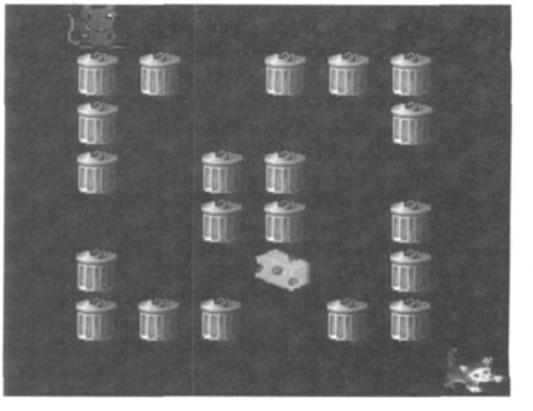

图 2 Predator-prey 模型

学习的目的是，Mouse 通过学习能够最快找到 Cheese，并且躲避 Cat 的捕获。当智能 Mouse 观察当前状态，执行动作，得到 Cheese，接受瞬间激励; 如果被 Cat 捕获，将会受到惩罚。Mouse 有 8 个动作分别为上、下、左、右和斜角的 4 个方向。定义强化信号函数如下:

$$r=\begin{cases}-10 & S_m=S_c\\ 5 & S_m=S_{ch}\\ 1 & \forall S_m,\max(D_{mc}-D_{mch})\end{cases} \quad (8)$$

式中：S_m、S_c、S_{ch}分别代表 Mouse、Cat、Cheese 的状态值（坐标值），$\max(D_{mc}^2-D_{mch}^2)$ 表示在 Mouse 执行动作的下一个状态时，与 Cat 的距离为 D_{mc}，与 Cheese 的距离为 D_{mch}，取状态集中的最大值。

学习算法中的参数选择如下：学习率 $\alpha=0.1$；折扣因子 $\gamma=0.9$；温度参数初始值 $T_0=100$。神经网络的结构是 8-16-8，隐含层激励函数是 Sigmoid 函数，输入输出为线性函数，图 3 为传统的 Q 学习与加入神经网络的混合结构的执行效果比较图。传统 Q 学习中训练的次数为 60 000 次，Mouse 的成功率（成功率 = 激励的次数/（激励的次数 + 惩罚的次数））为 64%，混合结构中 Mouse 的成功率为 78%，通过比较可看出，混合结构的成功率要更高一些。

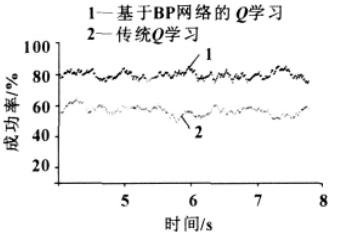

图 3　传统 Q 学习与改进 Q 学习下的 Mouse 成功率曲线

6　结论

在传统 Q 学习的基础上，采用了 BP 神经网络代替 Q 表格，形成混合智能控制，不仅提高了 Q 学习的泛化能力，而且大大缩减了计算量，在一定程度上提高了学习精度，增强了稳定性，并在 Predator-prey 实验中取得了很好的效果，证明了该方法的有效性与可行性。

参考文献：

【1】Wang B N，Cao Y，Chen Z Q，et al. LMRL：a multi-agent reinforcement learning model and algorithm [C]//Proceedings of Third Inter-national Conference on Information Technology and Applications（ICITA'05），2005. 7：303－307.

【2】仲宇，顾国昌，张汝波. 多智能体系统中的分布式强化学习研究现状 [J]. 控制理论与应用，2003，20（3）：317－322.

【3】高阳，陈世福，陆鑫. 强化学习研究综述 [J]. 自动化学报，2004，30（1）：86－100.

【4】周浦城，洪炳镕，黄庆成. 一种新颖的多 agent 强化学习方法 [J]. 电子学报，2006，34（8）：1488－1491.

【5】黄炳强，曹广益，王占全. 强化学习原理、算法及应用 [J]. 河北工业大学学报，2006，35（6）：34－38.

【6】JIANG Ju，Kamel M，CHEN Lei. Rein-forcement learning and aggregation [C]// 2004 IEEE International Conference on Systems，Man and Cybernetics，2004.

【7】http://www.cse.unsw.edu.au/~cs9417ml/RL1/index.htm.

（上接第 71 页）

监测中心；

数据存储。监控数据及时间信息能自动存储到数据库服务器；

数据显示。可以实现实时曲线显示、历史曲线查询、数据报表打印等；

数据查询。数据能根据需要分项显示，按时间查询历史数据、曲线等；

数据分析。可自动统计出年、月、日的累积流量、平均流量、最大流量、最小流量等。

4　结论

针对流量在线监控的要求，将 Zigbee 无线通信平台应用于流量计量监控系统，无需布线、组网灵活、维护方便，避免了人工抄表的繁琐劳动，提高了整个监控系统的自动化水平，为企业节能改造提供了科学的依据[8]，在气动、石油化工等计量企业具有广泛的应用价值。

由于无线传感网络的模块化特性，其可扩展性强，如果终端传感器数量加大，如新加入流量计等；或采集的数据量加大，如还需要采集管道流体温度、压力等，只需增加相应的传感器和无线模块，适当增加软件功能就可在较小的投入下方便地实现网络扩展，实现监控无死角。

参考文献：

【1】蔡茂林，香川利春. 气动系统的能量消耗评价体系及能量损失分析 [J]. 机械工程学报，2007，43（9）：69－74.

【2】凌治浩，周怡颋，郑丽国. Zigbee 无线通信技术及其应用研究 [J]. 华东理工大学学报，2006，32（7）：801－805.

【3】Zigbee Alliance. Zigbee Specification v1.1 [M]. 2006. 11.

【4】朱晓明，赵晓丽. 基于 UART 接口的 Zigbee 传感器网络的设计 [J]. 机床与液压，2008，36（10）：271－273.

【5】李文仲，段朝玉. Zigbee 无线网络技术入门与实战 [M]. 北京：北京航空航天大学出版社，2007.

【6】尹应鹏，李平舟，郭志华. 基于 CC2430 的 ZigBee 无线数传模块的设计和实现 [J]. 电子元器件应用，2008，10（4）：18－20.

【7】LIN Ke，HUANG Tinglei，LI Lifang. Design of temperature and humidity monitoring system based on Zigbee technology [C]//CCDC'09.，2009. 6：3628－3631.

【8】蔡茂林. 现代气动技术理论与实践第十讲：气动系统的节能 [J]. 液压气动与密封，2008（5）：59－62.