XXXII Неделя науки СПбГПУ. Материалы межвузовской научно-технической конференции. Ч.V : С.157-158

УДК 681.3.016; 621.311.075

С.В.Серебряков (6 курс, каф. САиУ), Л.А.Станкевич, доц.

ОБУЧЕНИЕ ПОВЕДЕНИЮ ДИНАМИЧЕСКИХ ОБЪЕКТОВ

Управление командной работой динамических объектов является одной из актуальных проблем развития многоагентных систем. Оно широко применяется для создания систем реального времени управления роботами, функционирующими в группе и выполняющими единую работу; сообществами интеллектуальных агентов, функционирующих в Internet; командами исполнителей, реализующих бизнес-процессы; организационно-техническими объектами, требующими корпоративного управления; группами автономных космических и военных объектов и пр.

В данной работе решается задача создания программных агентов, управляющих командой динамических объектов (роботов или автономных аппаратов) при ограничениях реального времени и необходимости взаимодействия агентов в плане кооперации или противодействия. Задача управления – на основе текущей информации о внешнем окружении и цели выбрать необходимые действия.

Обучение с подкреплением позволяет строить функции, которые отображают ситуации в действия так, чтобы максимизировать некоторый критерий. Обучающемуся не говорят, какие действия необходимо выполнить; наоборот, он должен сам определить, какие действия приводят к наилучшему результату. В более сложном случае выбор действия может затрагивать не только моментальный эффект, но также и все последующие. В этом случае обучающийся учится "жертвовать" моментальным результатам ради долгосрочной цели.

Приведём общий алгоритм обучения, известный как $Sarsa(\lambda)$ [1].

Введём следующие обозначения:

s – текущая ситуация; a – действие объекта; $Q(s,a)$ - функция, определяющая ценность действия a в ситуации s; $e(s,a)$ - вектор траекторий значений $Q(s,a)$; r – значение критерия, в соответствии с которым происходит обучение; α - шаг обучения; λ - параметр обновления траекторий.

Инициализировать $Q(s,a)$ произвольно и $e(s,a)=0$ для всех s,a.

Повторять для каждого эпизода: инициализировать s,a.

Повторять для каждого шага эпизода: выполнить действие a, определить $r,s^{/}$.

Выбрать в состоянии $s^{/}$ действие $a^{/}$, используя Q.

$$\delta \leftarrow r + \gamma Q(s^{/},a^{/}) - Q(s,a).$$

$$\begin{cases} 1.\text{ или } e(s,a) \leftarrow e(s,a)+1 \text{ - либо накапливаем} \\ 2.\text{ или} \\ \text{if } a == a^{/} \\ e(s,a) \leftarrow 1 \\ \text{else - либо замещаем} \\ e(s,a) \leftarrow 0 \end{cases}$$

Для всех s,a:

$$Q(s,a) \leftarrow Q(s,a) + \alpha\delta e(s,a)$$

$e(s,a) \leftarrow \lambda e(s,a)$

$s \leftarrow s^{/}$, $a \leftarrow a^{/}$.

до тех пор, пока $s^{/}$ - терминальное состояние.

Приведённый алгоритм был взят за основу реализации модуля обучения поведению агента в среде футбольного сервера RoboCup Soccer Server. Структура и интерфейс модуля были выбраны такими, чтобы была возможность встраивать этот модуль в различные сценарные заготовки. Это позволило унифицировать алгоритм обучения агента в различных сценариях. Об эффективности использования обучения при реализации агентов говорит тот факт, что практически все команды, достигшие серьезного успеха на соревнованиях, используют технику обучения

ЛИТЕРАТУРА:

1. Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An introduction. MIT Press, Cambridge, MA 1998.

2. Peter Stone and Richard S. Sutton. Scaling reinforcement learning toward RoboCup Soccer. In proceedings of the Eighteenth International Conference on Machine Learning, P. 537-544. Morgan Kaufmann, San Francisco, CA, 2001.