МОДЕЛИ АДАПТИВНОГО ПОВЕДЕНИЯ [*]

Редько В.Г.

Институт оптико-нейронных технологий РАН, Москва, Россия, E-mail: redko@iont.ru

В докладе характеризуется направление исследований «Адаптивное поведение».

1. Естественный путь к искусственному интеллекту. С начала 1990-х годов активно развивается направление исследований «Адаптивное поведение» [14]. Основной подход этого направления – конструирование и исследование искусственных (в виде компьютерной программы или робота) «организмов», способных приспосабливаться к внешней среде. Эти организмы называются «*аниматами*» (от англ. animal + robot = animat). Также часто используются термины «агент», «автономный агент».

Поведение аниматов имитирует поведение животных. Исследователи адаптивного поведения стараются строить именно такие модели, которые применимы к описанию поведения *как реального животного, так и искусственного анимата* [3].

Программа-минимум направления «Адаптивное поведение» – исследовать архитектуры и принципы функционирования, которые позволяют животным или роботам жить и действовать в переменной внешней среде.

Программа-максимум этого направления – попытаться проанализировать эволюцию когнитивных способностей животных и эволюционное происхождение человеческого интеллекта [12].

При этом данное направление исследований рассматривается как бионический подход к разработке систем искусственного интеллекта [21].

2. Исследователи адаптивного поведения. Исследования по адаптивному поведению ведутся в ряде университетов и лабораторий, таких как:

- AnimatLab (Париж, руководитель – один из инициаторов данного направления Жан-Аркадий Мейер) [5,12,14]. В этой лаборатории ведется широкий спектр исследований адаптивных роботов и адаптивного поведения животных. Подход AnimatLab предполагает, что система управления анимата может формироваться и модифицироваться посредством 1) *обучения*, 2) индивидуального *развития* (онтогенеза) и 3) *эволюции*.

[*] Работа выполнена при финансовой поддержке РАН (Программа "Интеллектуальные компьютерные системы", проект 2-45) и РФФИ (проект 04-01-00179).

- Лаборатория искусственного интеллекта в университете Цюриха (руководитель Рольф Пфейфер) [4,17]. Основной подход этой лаборатории – познание природы интеллекта путем его создания («understanding by building»). Он включает в себя 1) построение моделей биологических систем, 2) исследование общих принципов естественного интеллекта животных и человека, 3) использование этих принципов при конструировании роботов и других искусственных интеллектуальных систем.
- Лаборатория искусственной жизни и роботики в Институте когнитивных наук и технологий (Рим, руководитель Стефано Нолфи) [6,16], ведущая исследования в области эволюционной роботики и принципов формирования адаптивного поведения.
- Лаборатория информатики и искусственного интеллекта в Массачусетском технологическом институте (руководитель Родни Брукс) [7,11], которая ведет исследования широкого спектра интеллектуальных и адаптивных систем, включая разработки интеллектуальных роботов.
- Институт нейронаук Дж. Эдельмана, где ведутся разработки поколений моделей работы мозга (Darwin I, Darwin II, …) и исследования поведения искусственного организма NOMAD (Neurally Organized Mobile Adaptive Device), построенного на базе этих моделей [8,13].

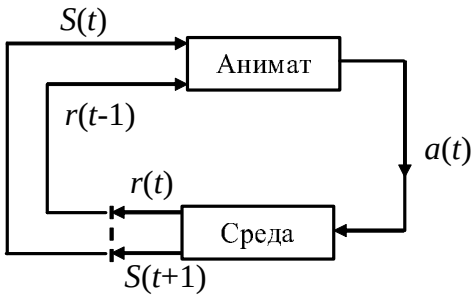
В России исследования адаптивного поведения пока ведутся скромными усилиями ученых-энтузиастов, среди этих работ следует отметить:

- модели поискового адаптивного поведения на основе спонтанной активности, приводящей к переключениям между разными тактиками поведения [15] (В.А. Непомнящих, Институт биологии внутренних вод им. И.Д. Папанина РАН);
- концепции и модели автономного адаптивного управления на основе аппарата эмоций [2] (А.А. Жданов, Институт системного программирования РАН);
- разработку принципов построения систем управления антропоморфных и гуманоидных роботов [10] (Л.А. Станкевич, Санкт-Петербургский политехнический университет);
- разработку нейросетевых моделей поведения роботов и робототехнических устройств [9] (А.А. Самарин, НИИ нейрокибернетики им. А.Б. Когана РГУ);
- модели адаптивного поведения на основе эволюционных и нейросетевых методов [1,18,19] (В.Г. Редько, М.С. Бурцев, О.П. Мосалов, Институт оптико-нейронных технологий РАН, Институт прикладной математики им. М.В. Келдыша РАН).

3. Обучение с подкреплением – ключевой метод самообучения аниматов. Один базовых методов обучения аниматов – обучение с подкреплением, которое

обеспечивает *самообучение* аниматов, в результате случайного поиска и получения поощрений и наказаний из внешней среды. Теория обучения с подкреплением (reinforcement learning) была развита в работах Р. Саттона и Э. Барто [20].

В обучении с подкреплением (рисунок) рассматривается анимат, взаимодействующий с внешней средой. Время предполагается дискретным: $t = 1,2,\ldots$ В текущей ситуации анимат $S(t)$ выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию $S(t+1)$. Подкрепление может быть положительным (награда) или отрицательным (наказание).



Общая схема обучения с подкреплением

Цель анимата – максимизировать суммарную награду, которую можно получить в будущем в течение длительного периода времени. Подразумевается, что анимат имеет свои внутренние субъективные оценки суммарной награды и в процессе обучения постоянно совершенствует эти оценки. Формула для оценок имеет вид:

$$U(t) = \sum_{k=0}^{\infty} \gamma^k \, r(t+k) \,, \qquad (1)$$

где $U(t)$ – оценка суммарной награды, ожидаемой после момента времени t, γ – коэффициент забывания, $0 < \gamma < 1$. Коэффициент забывания учитывает, что чем дальше анимат «заглядывает» в будущее, тем меньше у него уверенность в оценке награды («рубль сегодня стоит больше, чем рубль завтра»).

В процессе обучения анимат формирует *политику*. Политика определяет выбор действия в зависимости от ситуации. Если множество возможных ситуаций $\{S_i\}$ и действий $\{a_j\}$ конечно, то существует простой метод обучения SARSA, каждый шаг которого соответствует цепочке событий $S(t) \rightarrow a(t) \rightarrow r(t) \rightarrow S(t+1) \rightarrow a(t+1)$.

Кратко изложим метод SARSA. В этом методе итеративно формируются оценки величины суммарной награды $Q(S(t), a(t))$, которую получит анимат, если в ситуации $S(t)$ он выполнит действие $a(t)$. Математическое ожидание награды равно:

$$Q(S(t), a(t)) = \mathrm{E}\,\{r(t) + \gamma\, r(t+1) + \gamma^2\, r(t+2) + \ldots\} \mid S = S(t),\, a = a(t). \qquad (2)$$

Из (1) и (2) следует $Q(S(t), a(t)) = E\,[r(t) + \gamma\, Q(S(t+1), a(t+1))]$. Ошибку в оценке величины $Q(S(t), a(t))$ естественно определить так (см. также [20]):

$$\delta(t) = r(t) + \gamma\, Q(S(t+1), a(t+1)) - Q(S(t), a(t)). \qquad (3)$$

Величина $\delta(t)$ называется ошибкой временной разности.

В методе SARSA каждый такт времени происходит как выбор действия, так и обучение анимата.

Выбор действия происходит так:

- в момент t с вероятностью $1 - \varepsilon$ выбирается действие, соответствующее максимальному значению $Q(S(t), a_j)$: $a(t) = \arg\max_j \{Q(S(t), a_j)\}$
- с вероятностью ε выбирается произвольное действие, $0 < \varepsilon << 1$. Такую схему выбора действия называют «ε–жадным правилом».

При обучении к величине $Q(S(t), a(t))$ добавляется величина, пропорциональная ошибке временной разности $\delta(t)$:

$$\Delta Q(S(t), a(t)) = \alpha\, \delta(t) = \alpha\, [r(t) + \gamma\, Q(S(t+1), a(t+1)) - Q(S(t), a(t))], \qquad (4)$$

где α – параметр скорости обучения.

Метод обучения с подкреплением идейно связан с методом динамического программирования, и в том и другом случае общая оптимизация многошагового процесса принятия решения происходит путем упорядоченной процедуры одношаговых оптимизирующих итераций, причем оценки эффективности тех или иных решений, соответствующие предыдущим шагам процесса, переоцениваются с учетом знаний о возможных будущих шагах. Например, при решении задачи поиска оптимального маршрута в лабиринте от стартовой точки к определенной целевой точке сначала находится конечный участок маршрута, непосредственно приводящий к цели, а затем ищутся пути, приводящие к конечному участку, и т.д. В результате постепенно прокладывается трасса маршрута от его конца к началу. Обучение с подкреплением, адаптивные критики (см. ниже) и подобные методы часто называют приближенным динамическим программированием.

Важное достоинство метода обучения с подкреплением – его естественность. Анимат получает от учителя или из внешней среды только сигналы подкрепления $r(t)$. Здесь учитель поступает с обучаемым объектом примитивно: «бьет кнутом» (если действия объекта ему не нравятся, $r(t) < 0$), либо «дает пряник» (в противоположном случае, $r(t) > 0$), не объясняя обучаемому объекту, как именно нужно действовать. Это радикально отличает этот метод от традиционного в теории нейронных сетей метода

обратного распространения ошибок, для которого учитель точно определяет, что должно быть на выходе нейронной сети при заданном входе.

4. Нейросетевые адаптивные критики. Конструкции нейросетевых адаптивных критиков можно рассматривать как развитие моделей обучения с подкреплением на случай, когда как ситуации, так и действия задаются векторами **S** и **A,** и изложенная выше схема итеративного формирования матрицы $Q(S(t), a(t))$ не работает. В этом случае характеристики системы управления целесообразно представить с помощью аппроксимирующих нейронных сетей, а обучение проводить подстройкой весов синапсов нейронов путем минимизации ошибки временной разности.

5. Проект «Мозг Анимата» и модели эволюции адаптивных агентов. В докладе также будут охарактеризованы конкретные разработки на базе нейросетевых адаптивных критиков: 1) проект «Мозг Анимата» [18], который основан на теории функциональных систем П.К. Анохина и нацелен на формирование общей «платформы» для систематического построения моделей адаптивного поведения, и 2) модель эволюции автономных самообучающихся агентов, система управления которых обеспечивает прогноз будущих ситуаций во внешней среде и принятие решений на основе этого прогноза [19].

Литература

1. Бурцев М.С., Гусарев Р.В., Редько В.Г. Исследование механизмов целенаправленного адаптивного управления // Изв. РАН. Теория и системы управления. 2002. N.6. С.55-62.

2. Жданов А.А. Метод автономного адаптивного управления // Изв. РАН. Теория и системы управления. 1999. N. 5. С. 127-134.

3. Непомнящих В.А. Поиск общих принципов адаптивного поведения живых организмов и аниматов // Новости искусственного интеллекта. 2002. N. 2. С. 48-53.

4. Сайт AI Laboratory of Zurich University: http://www.ifi.unizh.ch/groups/ailab/

5. Сайт AnimatLab: http://animatlab.lip6.fr/index.en.html

6. Сайт Laboratory of Artificial Life and Robotics: http://gral.ip.rm.cnr.it/

7. Сайт MIT Computer Science and Artificial Intelligence Laboratory: http://www.csail.mit.edu/index.php

8. Сайт Neuroscience Institute: http://www.nsi.edu/

9. Самарин А.И. Модель адаптивного поведения мобильного робота, реализованная с использованием идей самоорганизации нейронных структур // IV Всероссийская

научно-техническая конференция «Нейроинформатика-2002». Материалы дискуссии «Проблемы интеллектуального управления – общесистемные, эволюционные и нейросетевые аспекты». – М.: МИФИ, 2003. С. 106-120.

10. Станкевич Л.А. Нейрологические средства систем управления интеллектуальных роботов // VI Всероссийская научно-техническая конференция «Нейроинформатика-2004». Лекции по нейроинформатике. Часть 2. – М.: МИФИ, 2004. С. 57-110.

11. Brooks R.A. Cambrian Intelligence: The Early History of the New AI. – MIT Press, 1999.

12. Donnart J.Y., Meyer J.A. Learning reactive and planning rules in a motivationally autonomous animat // IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, 1996. V. 26. N. 3. PP.381-395. See also: http://animatlab.lip6.fr/index.en.html

13. Krichmar J.L., Edelman G.M. Machine psychology: autonomous behavior, perceptual categorization and conditioning in a brain-based device // Cerebral Cortex, 2002, V. 12. PP. 818-830.

14. Meyer J.-A., Wilson S. W. (Eds.) From Animals to Animats. Proceedings of the First International Conference on Simulation of Adaptive Behavior**.** – The MIT Press: Cambridge, Massachusetts, London, England. 1990.

15. Nepomnyashchikh V.A., Podgornyj K.A. Emergence of adaptive searching rules from the dynamics of a simple nonlinear system // Adaptive Behavior. 2003. V.11. N.4. PP. 245-265.

16. Nolfi S., Floreano D. Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines. – Cambridge, MA: MIT Press/Bradford Books, 2000. 384 p.

17. Pfeifer R., Scheier C., Understanding Intelligence. – MIT Press, 1999.

18. Red'ko V.G., Prokhorov D.V., Burtsev M.S. Theory of functional systems, adaptive critics and neural networks // International Joint Conference on Neural Networks, IJCNN-2004, Budapest, 2004. PP. 1787-1792.

19. Red'ko V.G., Mosalov O.P., Prokhorov D.V. A model of Baldwin effect in populations of self-learning agents // International Joint Conference on Neural Networks, IJCNN-2005, Montreal, 2005.

20. Sutton R., Barto A. Reinforcement Learning: An Introduction. – Cambridge: MIT Press, 1998. See also: http://www.cs.ualberta.ca/~sutton/book/the-book.html

21. Wilson S.W. The animat path to AI // In: [14]. PP. 15-21.