Модели адаптивного поведения

В.Г. Редько

Институт оптико-нейронных технологий РАН
vgredko@gmail.com

Аннотация

Характеризуются работы отечественных и зарубежных исследователей адаптивного поведения. Излагаются метод обучения с подкреплением и оригинальная модель эволюции самообучающихся агентов, посвященная анализу взаимодействия между обучением и эволюцией в популяции адаптивных агентов.

1. Введение

В процессе биологической эволюции возникли чрезвычайно сложные и вместе с тем удивительно эффективно функционирующие живые организмы. Эффективность, гармоничность и согласованность работы "компонент" живых существ обеспечивается биологическими управляющими системами. Но каковы эти управляющие системы? Как и почему они эволюционно возникли? Какие информационные процессы обеспечивают работу этих управляющих систем? Как животные познают внешний мир и используют это познание для управления своим поведением? Как эволюционное развитие биологических информационных систем и познавательных способностей животных привело к возникновению интеллекта человека? Какие уроки из знаний о биологических информационных системах можно извлечь для разработки новых информационных технологий?

Есть ли современные исследования на стыке биологии и информатики, которые направлены на изучение этих интригующих проблем? Оказывается, что да, есть. Сравнительно недавно, в начале 1990-х годов сформировалось направление исследований "Адаптивное поведение" [1,2], которое можно рассматривать как задел в этой области и обзору которого посвящена настоящая лекция.

2. Направление исследований "Адаптивное поведение"

Основной подход направления "Адаптивное поведение" – 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) или "агентами".

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

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 – проанализировать эволюцию когнитивных способностей животных и эволюционное происхождение интеллекта человека.

В исследованиях адаптивного поведения используется ряд нетривиальных компьютерных методов [2], таких как:

- нейронные сети,
- генетический алгоритм и другие методы эволюционной оптимизации,
- обучение с подкреплением (Reinforcement Learning) [3].

В лекции характеризуются работы отечественных и зарубежных исследователей адаптивного поведения и в качестве конкретных примеров исследований излагаются метод обучения с подкреплением и оригинальная модель эволюции самообучающихся агентов, посвященная анализу взаимодействия между обучением и эволюцией.

3. Метод обучения с подкреплением

Обучение с подкреплением происходит в результате получения аниматом наград и наказаний, поступающих из внешней среды (рис. 1) [3].

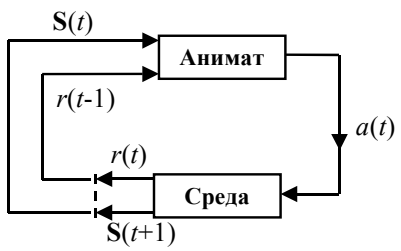

Рис. 1. Схема обучения с подкреплением.

Proceedings of the 1th International Conference on Mathematical Biology and Bioinformatics, Pushchino, Russia, 2006

В текущей ситуации **S**(t) анимат выполняет действие $a(t)$, получает подкрепление $r(t)$ и попадает в следующую ситуацию **S**(t+1), t = 1,2,…. Подкрепление $r(t)$ может быть положительным (награда) или отрицательным (наказание).

Цель анимата – максимизировать суммарную награду $U(t)$, которую можно получить в будущем. Величина $U(t)$ оценивается как

$$U(t) = \sum_k \gamma^k r(t+k) \quad , \qquad k = 0,1,2,\ldots,$$

где γ – коэффициент забывания ($0 < \gamma < 1$), учитывающий, что чем дальше анимат "заглядывает" в будущее, тем меньше у него уверенность в оценке ожидаемой награды.

В процессе обучения анимат формирует свои "субъективные" оценки суммарной награды $U(t)$, соответствующие разным ситуациям **S**(t) и действиям $a(t)$, и на основе этих оценок выбирает текущие действия. Метод обучения с подкреплением имеет серьезную теоретическую основу, он базируется на теории динамического программирования Беллмана и теории марковских процессов.

4. Модель эволюции популяции самообучающихся агентов

В нашей модели [4] исследовалось взаимодействие между обучением и эволюцией в популяции агентов, которые имели достаточно "интеллектуальную" нейросетевую систему управления, состоящую из двух нейронных сетей: Модели и Критика. Роль Модели – прогноз будущей ситуации **S**(t+1) для каждого из возможных действий. Роль Критика – оценка ожидаемой суммарной награды U для текущей и прогнозируемых ситуаций. На основе оценок величины U для прогнозируемых ситуаций производился выбор действий. Веса синапсов нейронных сетей агентов могли адаптироваться посредством 1) обучения с подкреплением и 2) дарвиновской эволюции. Модель отрабатывалась на примере очень простых агентов – агентов-брокеров.

Было проведено сравнение трех вариантов модели, в которые включены 1) либо обучение и эволюция одновременно, 2) либо отдельно эволюция, 3) либо отдельно обучение. Результаты компьютерного моделирования продемонстрировали, что в исследованной модели эволюция обеспечивает более эффективную стратегию поведения, чем обучение. Тем не менее, оказалось, что хотя обучение и несовершенно, оно способствует более быстрому нахождению оптимальной стратегии поведения по сравнению со случаем чистой эволюции.

Влияние обучения на эволюционный процесс может также проявляться в форме известного эффекта Болдуина – генетической ассимиляции приобретенных навыков в дарвиновской эволюции [5]. Этот эффект наблюдался в ряде наших компьютерных экспериментов. В исследованной модели этот эффект связан с эволюционным дрейфом начальных весов синапсов нейронных сетей, получаемых агентами при рождении, к таким весам, которые обеспечивают удачную стратегию поведения.

В случае чистой эволюции наблюдалось интересное поведение агента на начальных этапах эволюционного процесса. Стратегия поведения агента представляла собой реакцию только на общие изменения в окружающей среде, т.е. игнорировались мелкие и кратковременные вариации в изменении среды. Кроме того, в компьютерных экспериментах наблюдалась определенная инерционность в переключении между поведенческими тактиками агента.

Подобное инерционное, "усредняющее" поведение наблюдается у простейших животных, например, у личинок ручейников, осуществляющих поиск частиц подходящего размера для своего чехла-домика [6]. Процесс инерционного переключения между поведенческими тактиками позволяет животному использовать только общие, крупномасштабные свойства окружающего мира и игнорировать мелкие случайные детали.

Благодарность

Работа выполнена при финансовой поддержке программы Президиума РАН "Интеллектуальные компьютерные системы" (проект 2-45) и РФФИ (проект № 04-01-00179).

Литература

[1] J.-A. Meyer, S. W. Wilson, Editors, *From Animals to Animats. Proc. of the First International Conference on Simulation of Adaptive Behavior*. The MIT Press: Cambridge, Massachusetts, London, England, 1990.

[2] *От моделей поведения к искусственному интеллекту* (под ред. В.Г. Редько). М.: УРСС, 2006.

[3] R. Sutton, A. Barto. *Reinforcement Learning: An Introduction.* Cambridge: MIT Press, 1998.

[4] V.G. Red'ko, O.P. Mosalov, D.V. Prokhorov. A model of evolution and learning // *Neural Networks*, volume 18 (5-6), pages 738-745, 2005.

[5] P. Turney, D. Whitley, R. Anderson, Editors, *Evolution, Learning, and Instinct: 100 Years of the Baldwin Effect. Special Issue of Evolutionary Computation on the Baldwin Effect*, volume 4 (3), 1996.

[6] В.А. Непомнящих. Модели автономного поискового поведения // В книге [2], с. 200-242.