В.Б. КОТОВ

Научно-исследовательский институт системных исследований Российской академии наук, Москва

vkotov@ru.ru

К ВОПРОСУ О ФОРМИРОВАНИИ ЦЕЛЕЙ

На основе простых моделей модификации карты ценности состояний рассмотрены возможности формирования целей поведения разумного существа. Показано, что при подходящей стратегии выбора маршрутов в пространстве состояний возможно формирование целей, отличных от первичных целей, связанных с непосредственным получением награды.

**Ключевые слова:** *формирование целей, пространство состояний, ценность состояний, маршруты в пространстве состояний.*

Введение

Одной из главных особенностей нервной системы разумных существ, в частности, человека является способность формировать новые цели деятельности по мере освоения окружающего мира [1,2]. Даже у новорождённого ребёнка имеются врождённые потребности и инстинкты. Удовлетворение таких потребностей и реализацию таких инстинктов можно считать первичными врождёнными целями. По мере взросления в результате взаимодействия с природой и обществом в мозге человека возникают новые потребности и установки – идёт формирование вторичных целей, не связанных напрямую с удовлетворением биологических потребностей. Достижение сформированных целей является движущей силой деятельности человека.

Искусственная интеллектуальная система, претендующая на замену нервной системы человека, должна обладать развитым аппаратом работы с целями. Представляется естественным строить такой аппарат на основе функции ценности, отображающей пространство состояний существа в окружающем мире в пространство ценностей [3-6]. Ценность состояния – это число (или набор чисел), характеризующее данное состояние с точки зрения получения определённых выгод – наград. Отрицательная ценность связана с некоторыми издержками или вредными воздействиями («наказанием»). У живых существ ценность состояния выражается в «эмоциональной» оценке состояния (приятно – неприятно и т.п.).

В качестве целей деятельности естественно рассматривать достижение состояний с максимальной ценностью, т.е. максимума (глобального или хотя бы локального) функции ценности. Это соответствует стремлению животного достичь состояния с высокой «эмоциональной» оценкой. Таким образом, множество целей – это множество локальных максимумов функции ценности.

Функция ценности (или карта ценности для двумерного пространства состояний) формируется в процессе движения в пространстве состояний. Движение по некоторому маршруту в пространстве состояний приводит к получению наград и наказаний, в зависимости от которых происходит изменение ценности состояний маршрута. В результате исследования пространства состояний формируется функция ценности, рельеф которой определяет целенаправленное поведение.

Как правило, максимумы функции ценности находятся в точках (состояниях), где можно получить награду, т.е. где удовлетворяются некоторые потребности. Возникает вопрос. Возможно ли формирование максимумов, не связанных с непосредственным получением награды? И каков смысл соответствующих целей? Чтобы ответить на вопрос, обратимся к простой модели.

Модель формирования функции ценности

Ограничимся рассмотрением двумерного пространства состояний. Такой случай реализуется, в частности, если состояния – это места в физическом двумерном пространстве, при этом маршруты в пространстве состояний – это траектории движения на плоскости. В некоторых точках (дискретных) пространства состояний имеются источники награды. Для одного типа награды расположение и интенсивность источников награды можно задать функцией $Pr\left(x,y\right)$, равной нулю вне источников награды. Координаты *x*,*y* нумеруют точки пространства состояний.

В простейшем случае функция ценности – это скалярная функция $Val\left(x,y\right)$. Формирование этой функции происходит при движении в пространстве состояний. Пусть при движении по некоторому маршруту *x*(*t*),*y*(*t*) в момент времени *t*=*t*0 достигается точка с источником награды: $Pr\left(x\left(t\_{0}\right),y\left(t\_{0}\right)\right)\ne 0$. Тогда в данной точке и нескольких предыдущих точках маршрута происходит модификация значений ценности:

 $∆Val\left(x\left(t\_{0}-∆t\right),y\left(t\_{0}-∆t\right)\right)=k\_{pr}\left(∆t\right)Pr\left(x\left(t\_{0}\right),y\left(t\_{0}\right)\right)$, (1)

где $∆t=0,1,…,∆t\_{max}$, *kpr* – коэффициент модификации, зависящий от удалённости (по времени) точки маршрута от точки с наградой. Ясно, что чем больше *t*, тем меньше коэффициент модификации (дисконтирование). В частности, можно взять экспоненциальную зависимость

 $k\_{pr}\left(∆t\right)=k\_{pr0}exp\left(-α∆t\right)$, (2)

где *kpr*0 – коэффициент модификации для текущего состояния, *α* – скорость дисконтирования.

В реальных условиях имеется несколько видов наград, связанных с удовлетворением разных потребностей. Каждому типу награды соответствует своя функция ценности. При этом ценность состояния задаётся набором чисел, который называем вектором ценности состояния $\vec{v}$. Из компонент векторной функции $\vec{v}\left(x,y\right)$ нужно составить скалярную функцию *Val*(*x*,*y*), которая и управляет целенаправленным поведением. Сделать это можно разными способами. Простейший способ — суммирование компонент с весами:

 $Val\left(x,y\right)=\sum\_{j}^{}a\_{j}v\_{j}\left(x,y\right)$. (3)

Существенно, что веса *aj* зависят от насущных потребностей организма (его внутреннего состояния). Так, для организма, испытывающего голод, наиболее существенна награда в виде пищи — наибольший вес имеет компонента вектора ценности, связанная с пищевой наградой.

Конечно, описание ценности на основе формулы (3) может оказаться неадекватным при описании реальных организмов. Это связано с тем, что разные модальности наград и наказаний обрабатываются разными отделами мозга (при использовании разных методов обработки) [3-5]. В частности, отсутствует симметрия между положительными наградами и наказаниями (отрицательными наградами) — подчас высокая награда не способна преодолеть страх наказания. Достоинство формулы (3) — её простота и ясный смысл.

Что касается формулы (1), описывающей модификацию функции ценности (или компонент векторной функции ценности), то она может быть усовершенствована в нескольких направлениях. Во-первых, интенсивность модификации функции ценности в общем случае зависит от внутреннего состояния организма. (Так, сытое животное может не заметить источник пищи.) Этот факт легко учесть, считая коэффициент *kpr* зависящим от внутреннего состояния организма.

Во-вторых, из формулы (1) следует неограниченный рост значений ценности состояний по мере многократного прохождения через эти состояния. Чтобы устранить такой «нефизический» рост, необходимо ввести те или иные механизмы ограничения. Можно, например, ввести тотальное затухание на каждом временном шаге:

 $∆Val\left(x,y\right)=-k\_{t}Val\left(x,y\right)$, (4)

где *kt* – малый коэффициент затухания. Такой механизм ограничения слишком грубый – он может приводить к стиранию информации о ценности состояний некоторой области пространства, если эта область долго не посещается.

Другой, более правдоподобный с биологической точки зрения способ ограничения значений функции ценности, основан на вычитании из интенсивности источника награды *Pr*(*x*,*y*) в правой части формулы (1) некоторого ожидаемого значения этой величины *Pre*(*x*,*y*):

Рис. 1. Максимум функции ценности, сформированный в результате «прямолинейного» движения

$∆Val\left(x\left(t\_{0}-∆t\right),y\left(t\_{0}-∆t\right)\right)=k\_{pr}\left(∆t\right)\left(Pr\left(x\left(t\_{0}\right),y\left(t\_{0}\right)\right)- -Pr\_{e}\left(x\left(t\_{0}\right),y\left(t\_{0}\right)\right)\right)$. (5)

Ожидаемая интенсивность награды пропорциональна значению ценности в данной точке: *Pre*(*x*,*y*)=*CVal*(*x*,*y*), где *C* – нормировочный коэффициент. У уравнения (5) имеется стационарные решения, для которых *Val*(*x*,*y*)=*Pr*(*x*,*y*)/*C* в точках с наградой. При больших временах движения по пространству состояний (с заходом во все точки-состояния) функция ценности стремится к одному из стационарных решений. Значения ценности состояний без источников награды определяются выбором маршрутов движения. Заметим, что на начальных этапах исследования пространства состояний значения функции ценности малы, при этом можно отбросить член с *Pre* в правой части уравнения (5), что приводит к формуле (1).

В-третьих, может возникнуть необходимость учёта дополнительных источников модификации функции ценности. Дело в том, что живое существо может испытывать приятные ощущения не только от получения награды, но также находясь в состоянии с высокой ценностью (без источника награды) – от «предвкушения награды». Таким образом, высокую ценность данного состояния можно рассматривать как своеобразную награду, которая может служить источником модификации функции ценности. Для учёта этого в правую часть уравнения (1) (или (5)) необходимо добавить члены, зависящие от ценности состояний маршрута. Сделать это можно различными способами. Соответствующие выражения не приводим из-за недостатка места.

Рис. 2. Максимум функции ценности, сформированный в результате хаотического движения

Маршруты движения в пространстве состояний

и рельеф функции ценности

Процесс формирования функции ценности существенно зависит от выбора маршрутов движения в пространстве состояний. Например, при движении по «прямолинейному» маршруту от начальной точки-состояния до источника награды формируется рельеф функции ценности с сильно анизотропным максимумом (рис. 1). Если же движение хаотическое (случайный маршрут), то после многократного прохождения состояния с наградой формируется более или менее изотропный максимум с центром в источнике награды (рис. 2). Рисунки 1, 2 получены с использованием уравнения (1) при *t*max=6, *α*=0,3. Ширина максимума для случайного маршрута существенно меньше ширины максимума для «прямолинейного» маршрута в направлении движения. 

Относительно большая ширина максимума и возможность управлять направлением оси максимума путём изменения направления движения позволяют формировать довольно интересные рельефы функции ценности, которые содержат дополнительные, не привязанные к источнику награды, локальные максимумы. Так, при наличии двух источников награды и использовании пересекающихся «прямолинейных» маршрутов для подхода к этим источникам, можно получить дополнительный максимум в точке пересечения маршрутов (рис. 3). Появление такого максимума можно рассматривать как формирование новой промежуточной цели, достижение которой позволяет легко достичь любого из двух источников награды. Заметим, что при том же, что и на рис. 3, расположении источников награды использование случайных маршрутов не приводит к формированию дополнительного максимума — новой цели (рис. 4).

Рис. 3. Дополнительный максимум в точке пересечения маршрутов

Насколько реалистично использование заданных повторяющихся маршрутов при формировании функции ценности? Наблюдение за живыми существами (мышами, человеком) показывает, что даже при первоначальном исследовании новой обстановки используемые маршруты далеки от случайных [7]. Общая картина движения довольно предсказуема, хотя нет полной повторяемости маршрутов. Полностью хаотическое движения очень неэкономично как из-за большого числа поворотов, так и из-за слишком медленного освоения пространства состояний. И для искусственных существ нет причин пользоваться полностью случайными маршрутами. По мере освоения обстановки движения существа становятся все более целенаправленными, так что по большей части происходят перемещения между максимумами функции ценности. В результате таких упорядоченных движений могут формироваться новые максимумы (цели).

Рис. 4. Отсутствие дополнительного максимума функции ценности при использовании случайных маршрутов

Но даже при использовании практически неинтересных случайных маршрутов формирование новых максимумов функции ценности может происходить за счёт специальной геометрии пространства состояний. Наличие границ и препятствий специального вида приводит к уменьшению хаотичности движения и возможности формирования дополнительных максимумов функции ценности. На рис. 5 изображена формирующаяся функция ценности для разделённого на два отсека пространства состояния, причём отсеки соединяются только узким проходом. Два источника награды находятся в дальнем отсеке, а маршруты начинаются в ближнем отсеке. Несмотря на случайную процедуру выбора каждого шага маршрута около прохода между отсеками формируется максимум, не связанный непосредственно с наградой. Появление такого максимума соответствует формированию промежуточной цели (прохода между отсеками), достижение которой необходимо для достижения одной из целей — источников награды. Заметим, что если положения источников награды известно (из предыдущего исследования пространства), можно использовать кратчайшие маршруты между проходом и источниками награды; это позволяет увеличить интенсивность максимума функции ценности около прохода.

Рис. 5. Формирование максимума функции ценности в проходе между отсеками

При наличии нескольких типов наград возможности формирования дополнительных максимумов скалярной функции ценности и их содержательной интерпретации многократно возрастают. Напомним, что рельеф этой функции (в частности, её максимумы) существенно зависит от внутреннего состояния организма. При изменении насущных потребностей происходит изменение максимумов-целей. При этом целенаправленное поведение сводится к движению между появляющимися последовательно целями. Перемещение между ними происходит по наиболее удобным траекториям; обычно это кратчайшие траектории. В местах пересечения траекторий создаются условия для формирования локальных максимумов скалярной функции ценности. Реализация таких максимумов зависит от насущных потребностей организма – новая цель проявляется, когда желательными оказываются два или больше типов наград.

1

4

2

5

3

Рис. 6. Формирование новой цели на пересечении путей между старыми целями

На рис. 6 показан пример с четырьмя разными источниками наград: 1, 2, 3, 4. Стрелками показаны некоторые часто используемые пути между этими источниками. В точке (состоянии) 5, находящейся на пересечении путей, оказываются положительными компоненты векторной ценности, соответствующие четырём типам наград. Когда внутреннее состояние существа таково, что одинаково полезными оказываются несколько типов наград (от 2 до 4), скалярная функция ценности (2) имеет в точке 5 локальный максимум, соответствующий дополнительной цели. Целевое состояние 5 можно рассматривать как состояние ожидания, где можно дождаться изменения внутреннего состояния, в результате которого одна из целей-наград станет доминирующей.

Если реализован механизм модификации ценности за счёт особенностей функции ценности, а не только источников наград, то оказывается возможным формировать новые цели на основе вторичных ранее сформированных целей. Здесь возможности формирования целей ограничиваются в основном сложностью пространства состояний. Заметим, что для богатой системы ценностей двумерного пространства состояний явно недостаточно.

Заключение

Таким образом, даже при использовании простейшей модели формирования функции ценности возможно формирование максимумов функции ценности, не связанных непосредственно с источниками награды. Формирование таких дополнительных максимумов естественно рассматривать как формирование новых целей, достижение которых может давать определённые выгоды.

В представленных вариантах модели формирование новых максимумов происходит в местах пересечения «хвостов» от других максимумов функции ценности. Отсюда следует важность выбора маршрутов движения в пространстве состояний. Приведённые примеры показывают, что предпочтительными являются маршруты, имеющие признаки закономерности и повторяемости. Хаотическое, полностью случайное движение в пространстве состояний непригодно для формирования новых целей.

Задание подходящих маршрутов возможно при запрограммированном поведении (реализации инстинктов) или при целенаправленном поведении, управляемом уже имеющимися целями. Заданию подходящих маршрутов движения может способствовать наличие геометрических ограничений (препятствий, границ) в пространстве состояний.

*Список литературы*

1. Котов В.Б. Мозгостроение для дилетантов. Москва: ООО «Центр Инновационных Технологий». 2015.
2. Саттон Р.С., Барто Э.Г. Обучение с подкреплением. М.: Лаборатория знаний. 2014.
3. Baxter M.G., Murray E.A. The Amygdala and Reward. //Nature Reviews. Neuroscience. Vol. 3. July 2002. P. 563-573.
4. Wilson R.C, Takahashi Y.K., Schoenbaum G., Niv Y. Orbitofrontal Cortex as a Cognitive Map of Task Space. //Neuron 81 (2014), January 22. P. 267–279.
5. Strait C.E., Blanchard T.C., Hayden B.Y. Reward Value Comparison via Mutual Inhibition in Ventromedial Prefrontal Cortex. //Neuron 82 (2014), June 18. P.1–10.
6. Murray E.A., Rudebeck P.H. The drive to strive: goal generation based on current needs. //Frontiers in Neuroscience. Decision Neuroscience. June 2013. V. 7. Article112.
7. Хайнд Р. Поведение животных. Синтез этологии и сравнительной психологии. М.: Мир. 1975