А. Е. Лебедев, А. А. Жданов
ОАО "Институт точной механики и вычислительной техники имени С.Лебедева", Москва
aazhdanov@ipmce.ru

ДИНАМИЧЕСКАЯ СЕГМЕНТАЦИЯ ПРОСТРАНСТВА ПРИЗНАКОВ ДЛЯ СИСТЕМ АВТОНОМНОГО АДАПТИВНОГО УПРАВЛЕНИЯ И СИСТЕМ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

В данной статье рассматривается алгоритм динамической сегментации пространства входных параметров на основе бинарных деревьев. Данный алгоритм может быть использован как в системах автономного адаптивного управления, так и в системах обучения с подкреплением, и предназначен для повышения эффективности и скорости обучения этих систем.

Введение

Метод «Автономного Адаптивного Управления» разработан на основе концептуальной модели нервной системы живых организмов [1]. Обучение агента с системой управления на основе данного метода происходит непосредственно в процессе его взаимодействия со средой. При этом информация о совершенных действиях и их результате накапливается в базе знаний. Впоследствии этот опыт используется для повышения качества управления (агент стремится максимизировать свою «эмоциональную оценку», отражающую качество текущего состояния).

Концепция обучения с подкреплением также была инспирирована естественными системами адаптации [2]. В соответствие с данной парадигмой обучение агента также производится методом «проб и ошибок». Агент может совершать различные действия, которые переводят его в различные состояния и приводят к получению подкрепления – вознаграждения или наказания.

Рассмотренные классы управляющих систем не совпадают полностью, однако имеют значительное пересечение. В общем случае методика обучения с подкреплением предполагает, что величины вознаграждения, а также вероятности перехода между состояниями среды известны. Однако существуют способы (основанные, в основном, на алгоритме «Q-learning» [3]) которые позволяют обходиться только полученными в процессе обучения данными. Такие алгоритмы можно рассматривать как разновид-

ность метода «автономного адаптивного управления». С другой стороны, если система автономного адаптивного управления построена так, что каждому возможному состоянию среды соответствует единственный образ, то можно построить эквивалентную систему обучения с подкреплением. При этом различным образам в терминологии метода автономного адаптивного управления будут соответствовать различные состояния в терминологии обучения с подкреплением. В дальнейшем мы будем рассматривать именно это пересечение, поэтому все рассуждения в настоящей статье одинаково применимы как к системам автономного адаптивного управления, так и к системам обучения с подкреплением.

В обеих концепциях, как правило, текущее состояние агента определяется набором параметров. Например, в качестве таких параметров могут выступать текущие показания различных датчиков робота. Параметры могут принимать значения как из дискретного, так и из непрерывного множества. Совокупность всех возможных значений параметров и есть пространство параметров. Для обеспечения работы алгоритма обучения это пространство должно отображаться на множество образов (состояний). Самый простой способ такого отображения – «геометрический», при котором множество значений каждого параметра разбивается на несколько интервалов. При этом получается многомерная сетка-таблица (её размерность совпадает с количеством параметров), индексами которой выступают номера соответствующих интервалов, а элементами – образы (состояния). Выбор размера интервалов всегда связан с нахождением компромисса. С одной стороны, чем мельче будет размер ячеек (и чем их будет больше), тем большей точности управления можно достигнуть. С другой стороны, увеличение числа образов не только увеличивает затраты памяти, но и требует большего времени для обучения (так как для каждого образа требуется накопление эмпирических данных). Это нежелательно не только с точки зрения вычислительных затрат, но и с точки зрения экономии физических ресурсов в случае обучения на реальной модели (например, с помощью мобильного робота). Более того, при табличном подходе число образов растет экспоненциально с увеличением числа параметров. Поэтому данный подход неприменим для решения многих задач с большим числом параметров даже при обучении на виртуальной модели.

Наиболее популярное решение этой проблемы для обучения с подкреплением – использование многослойного персептрона для осуществления отображения [4]. При этом текущее состояние задается неявно: вычисляются лишь величины ожидаемого вознаграждения при совершении определенного действия (Q-фактор). Этого достаточно для выбора оптималь-

ного действия в каждый момент времени. Обобщение эмпирических данных достигается за счет аппроксимации функций ожидаемой награды в нейросетевом базисе.

Для этой же цели могут быть использованы также деревья решений. Например, в [5] рассматривается алгоритм, позволяющий свести задачу обучения с подкреплением к серии классических задач обучения с учителем. Он был опробован с использованием нескольких типов регрессионных деревьев решений. Однако обучение в данном случае происходит не одновременно с выполнением действий, а в «пакетном режиме», т.е. сначала данные только накапливаются, а затем сформированное обучающее множество поступает на вход классификатора. Таким образом, нет возможности сразу же использовать накопленный опыт. Кроме того, редуцируя проблему аппроксимации Q-фактора к классическим алгоритмам, не учитываются некоторые особенности обучения с подкреплением. Например, критерий качества работы здесь должен быть несколько иной - главное не получение точного значения для всех параметров, а установление лучшего действия.

В отличие от сказанного, метод «автономного адаптивного управления» способен обучаться непосредственно в процессе управления». В соответствие с этим, в настоящей работе рассматривается алгоритм построения дерева, осуществляющего адаптивную динамическую сегментацию пространства признаков, т.е. одновременно с процессом обучения и с учетом предварительно накопленных данных.

Постановка задачи и описание алгоритма

Разработанный алгоритм предполагает решение задачи автономного адаптивного управления или обучения с подкреплением. Агент управления обучается действовать в среде, априорная информация о свойствах которой минимальна. При управлении агент использует только значения входных параметров и конечный набор действий. Принятие решения происходит дискретно, на каждом шаге агент может выбрать одно из доступных действий. Эффект от выполнения действия может как вычисляться при компьютерном моделировании среды, так и быть связанным с физическим совершением действия. При этом произошедшие изменения могут как отражаться на наблюдаемом значении входных параметров агента, так и не отражаться. Предполагается, тем не менее, что значения входных параметров как-то (возможно неявно) связаны с состоянием среды, иначе их использование для управления невозможно. Кроме того, при этом мо-

жет измениться величина «эмоциональной» оценки состояния агента (что соответствует получению вознаграждения или наказания). Задача агента – максимизировать среднюю оценку своего состояния.

В соответствие с методологией автономного адаптивного управления, набор входных параметров отображается в множество образов. В данной работе предполагается, что это отображение дискретно и каждому образу соответствует определенный интервал значений для каждого из параметров, т.е. образы разделяют пространство признаков на многомерные прямоугольники. Для каждого образа накапливается статистика поведения системы при распознавании этого образа. Для каждого действия вычисляется средняя оценка результата выполнения этого действия. Эта оценка позволяет судить, насколько выполнение данного действия в данных условиях способствует достижению хорошего (по определенному критерию) качества работы системы в целом, и используется при управлении для выбора действий. В простом случае в качестве такой оценки может выступать изменение эмоциональной оценки на следующем шаге или полученная непосредственно после совершения действия награда (наказание). В более сложном случае эта оценка учитывает подкрепление, которое может быть получено через несколько шагов. Например, Q-фактор также может выступать в роли такой оценки.

В основе предлагаемого метода лежит идея поэтапного разбиения пространства признаков. В начале производится грубое разбиение, которое позволяет получить приближенную версию закона управления. Как правило, это способствует определенному улучшению качества управления системой. При этом частота распознавания различных образов может измениться по сравнению с управлением по случайному закону (обычно обучение приводит к более частому посещению состояний с более высокой оценкой качества). Иногда без предварительного обучения достижение определенных состояний практически невозможно. Это обстоятельство позволяет надеяться на более быстрое дальнейшее обучение.

В соответствие с рассматриваемым алгоритмом, все множество образов организуется в иерархическую древовидную структуру. Изначально все пространство признаков соответствует одному образу (впрочем, при необходимости, можно задать любое начальное приближение). По мере обучения может происходить разделение некоторых образов на два и более мелких образа. В рассматриваемой версии алгоритма разделение производится только по одному из параметров, то есть образы сохраняют форму многомерных прямоугольников. При разделении образа происходит рост дерева: к узлу, соответствующему разделяемому образу, добавляются два

потомка, соответствующих новым образам. Множество всех листьев дерева образует множество активных на текущий момент образов. Для каждого образа накапливается как общая статистика оценок результатов действий, так и отдельная для каждого из доступных действий. Информацию об оценке результата совершения определенного действия при нахождении системы в состоянии, соответствующем определенному образу, будем называть правилом. Каждому образу может соответствовать столько правил, сколько действий может совершить агент в данной ситуации. Однако для какого-то из правил может быть недостаточно прецедентов для накопления статистики. Это может произойти, если какое-то действие совершалось редко или вообще не совершалось при условии распознавания данного образа. В этом случае для принятия решения может быть использовано правило, соответствующее вышестоящему (в дереве) образу, то есть такому образу, который содержит рассматриваемый листовой образ как часть.

Основными подзадачами, которые необходимо было решить для обеспечения работы алгоритма, являлись способ выбора параметров разделения очередного образа, а также способ определения готовности к проведению разделения.

Выбор направления разделения означает выбор параметра, по которому будет производиться очередное разделение. Основной фактор, учитываемый при выборе разделения – максимальное значение оценки результата действий для одного из дочерних образов. Чтобы выявить это значение в процессе обучения, для каждого неразделенного еще образа (листа дерева) создается по два «потенциальных» потомка на каждый параметр, разделенные по соответствующему параметру. Иными словами, производятся все возможные предварительные разделения. Для каждого из получившихся «потенциальных» образов накапливается статистика, аналогично тому, как это делается для активных образов. При выборе направления разделения находится пара потенциальных новых образов, один из которой содержит правило с максимальной (для рассматриваемого активного образа) оценкой результата действия. Если выбранное разделение в итоге производится, и рассматриваемые потенциальные образы переводятся в разряд активных, то упомянутое правило с максимальной оценкой будет соответствовать лучшему для данного образа действию. Если это действие не совпадает с лучшим действием для родительского образа, то закон управления становится более детальным. Если же лучшие действия для обоих новых образов совпадают с лучшим действием для их родителя,

то закон управления не меняется, однако уменьшается дисперсия оценки результата действия, что способствует более точному прогнозированию.

Разбиение определяется не только направлением, но и величиной параметра, которая станет граничной. Эту величину следует выбирать таким образом, чтобы для обеих частей образа количество прецедентов было по возможности одинаковым. В противном случае могут возникнуть образы, слишком редко возникающие в процессе работы системы, что замедляет обучение. Кроме того, близость частот возникновения образа желательна при экстраполяции статистики обобщающего родительского образа на его потомков (в противном случае статистика для родительского образа будет отражать не усредненную ситуацию для его частей, а отвечать лишь одной из них). Лучше всего для получения разбиения на равные (по числу прецедентов) части подошло бы использование медианы значения разделяемого параметра всех прецедентов. Однако постоянное её вычисление затруднительно в режиме «on-line» обучения. Поэтому на практике был применен более грубый способ. Для каждого параметра изначально создается несколько вариантов разбиения по рассматриваемому параметру с различными разделяющими значениями. Для каждого варианта разбиения в процессе обучения накапливается статистика. Затем выбирается та пара новых образов, у которой распределение прецедентов наименее отличается от равного. Если вариантов несколько, то выбирается тот, при котором дочерний образ, содержащий меньшее количество прецедентов, имеет большую среднюю оценку качества состояния системы, чем тот, который содержит большее количество прецедентов. Это связано с предположением, что в результате обучения агент чаще должен попадать в ситуации с более высокой оценкой качества состояния системы, т.е. ожидается, что распределение прецедентов в выбранном разделении изменится в сторону выравнивания.

Другой основной подзадачей, решение которой необходимо для успешной работы алгоритма, является выбор способа определения готовности к проведению разделения. С одной стороны, при слишком медленном темпе проведения разделений процесс обучения будет замедляться. С другой стороны, если темп будет слишком быстрый, то обучение не будет поспевать за ростом количества деталей, и поведение агента может превратиться в хаотическое. Кроме того, при поспешном проведении разделения повышается шанс выбора неправильного направления, что также может отрицательно сказаться на качестве управления. Поэтому для определения готовности к разделению используется множество различных факторов. Например, наиболее простым необходимым условием проведения разде-

ления является наличие у обоих дочерних образов не менее определенного (эмпирически подобранного) количества прецедентов. При этом прецеденты должны соответствовать хотя бы двум различным действиям, иначе нет смысла в уточнении образа. Кроме того, дисперсия оценки результатов действия не должна возрастать при уточнении образа, а лучше, если она убывает (чем сильнее она убывает, тем вероятнее проведение разделения). Помимо этого, оценивается надежность выбора направления разделения. Например, если этот выбор остается постоянным в течение определенного количества последних шагов, то повышается вероятность того, что он не был произведен случайно.

Эксперименты и выводы

Описанный алгоритм применялся для обучения управления компьютерной и физической моделью наноспутника. Внешний вид физической модели наноспутника представлен на Рис. 1.

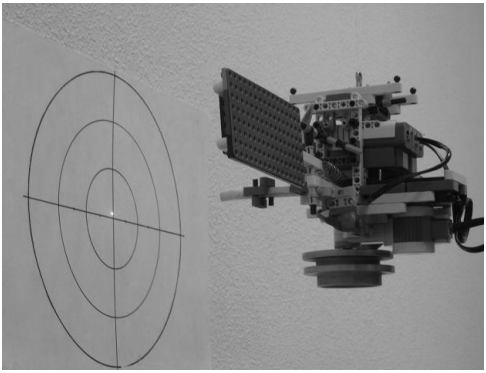

Рис. 1. Физическая действующая модель наноспутника.

В обоих случаях целью системы управления является стабилизация углового положения спутника в одном измерении. Оценкой текущего состояния для такой системы является максимум модулей углового отклонения от цели и угловой скорости. Соответственно, критерием качества работы системы является средняя величина этой оценки за определенный промежуток времени. Для обеспечения управляющего воздействия спутнику может быть сообщено угловое ускорение различной силы и направления.

На каждой из этим моделях испытывалась работа системы автономного адаптивного управления как с использованием, так и без использования описанного алгоритма. Во втором случае каждый из параметров был разделен на 8 интервалов. В качестве входных параметров использовались угловое отклонение от цели и угловая скорость. Выяснилось, что при использовании алгоритма динамической сегментации пространства признаков обучение происходит быстрее. В среднем, уровень качества 0.9 достигался за вдвое меньшее время.

В следующей серии экспериментов к двум использовавшимся признакам был добавлен также номер предыдущего совершенного действия (этот признак может быть полезным для управления, если двигатели спутника обладают инерционностью). На Рис. 2 и на Рис. 3. представлены примеры графиков изменения качества работы системы с использованием алгоритма динамической сегментации и, соответственно без его использования. Из графиков видно, что качество управления при использовании предложенного алгоритма возрастает быстрее. Время обучения при этом у обеих систем выросло (по сравнению со случаем с 2 параметрами), однако у обычной системы адаптивного управления оно выросло более значительно. В среднем уровень качества 0.9 достигался за втрое меньшее время. Результаты экспериментов подтверждают эффективность разработанного алгоритма.

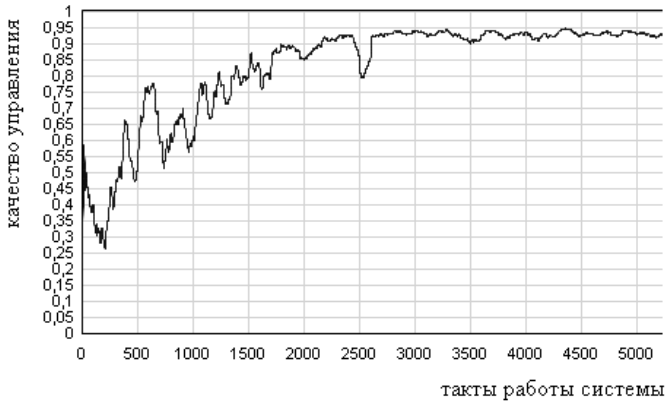

Рис. 2. Пример графика обучения системы при использовании алгоритма сегментации (для обучения используются 3 параметра).

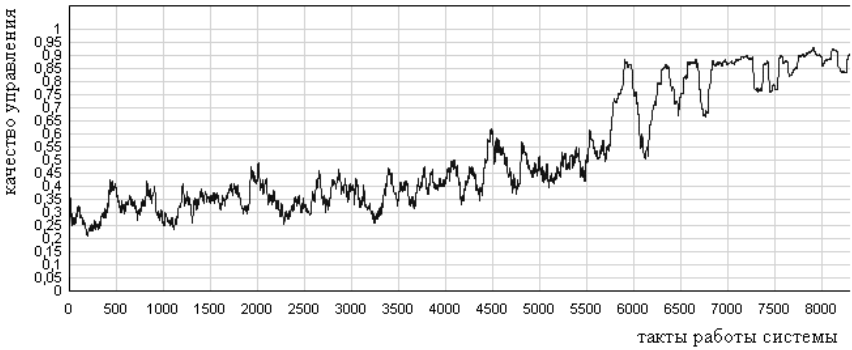

Рис. 3. Пример графика обучения системы без использования алгоритма сегментации (для обучения используются 3 параметра).

Список литературы

1. Жданов А.А. Метод автономного адаптивного управления //Известия РАН. Теория и системы управления. №5. 1999. С. 127-134.
2. Sutton R.S., Barto A.G. Reinforcement Learning, an Introduction //MIT Press, 1998.
3. Watkins C.J.C.H., Dayan P. «Q-Learning» //Machine Learning, 1992, vol. 8, p. 279-292.
4. Bertsekas D.P., Tsitsiklis J.N. Neuro-Dynamic Programming //Belmont, MA: Athenas Scientific, 1996
5. Ernst D., Geurts P., Wehenkel L. Tree-Based Batch Mode Reinforcement Learning //Journal of Machine Learning Research 6 (2005) 503–556