УДК 004.8:623

А.С. ГОРСКИЙ, кандидат технических наук
В.М. ПОЛУШКИН, кандидат технических наук
Р.И. КНЯЗЕВ, кандидат технических наук

РАСПОЗНАВАНИЕ ОБРАЗОВ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ

В статье рассматривается подход к решению задачи распознавания образов, объединяющий метод обучения с подкреплением на основе временных различий с архитектурой «исполнитель-критик» в виде глубоких искусственных нейронных сетей и использованием модели окружающей среды. Предложен методический аппарат, который может быть использован при разработке алгоритмов искусственного интеллекта (ИИ) на основе обучения с подкреплением как для решения задачи распознавания образов, так и для других задач ИИ в условиях, когда обучение с учителем (без учителя) имеет организационные сложности или требует больших затрат времени и ресурсов.

***Ключевые слова**: искусственный интеллект; машинное обучение с подкреплением; метод на основе временных различий; метод «исполнитель-критик»; распознавание образов; функция ценности.*

Введение

В современных условиях развития передовых информационных технологий одной из основных мировых тенденций является применение технологий искусственного интеллекта (ИИ) в различных сферах человеческой деятельности. Потребность в решении прикладных задач с помощью ИИ вызывает необходимость переосмысления многих аспектов создания новых технических систем как в России, так и за рубежом.

Развитие технологий ИИ тесно связано с таким направлением, как машинное обучение. В работе [1] рассматривается подход, который предполагает решение типовой задачи распознавания образов объектов с применением искусственных нейронных сетей (ИНС) преимущественно на основе обучения с учителем.

Метод обучения с учителем широко освещен во многих научных работах, достаточно прост в понимании, но для его эффективного применения необходим достаточно большой набор обучающих данных, подготовленный квалифицированным учителем (человеком). Этот набор представляет собой совокупность прецедентов – пар «объект-ответ». Каждый прецедент имеет признаковое описание (вектор признаков) объекта и эталонный отклик, который соотносит объект с заранее определенным классом. Целью такого обучения является способность обобщать полученный опыт на новые объекты, которые не были представлены в обучающем наборе [2; 3].

Другим методом машинного обучения является обучение без учителя. В отличие от обучения с учителем он предполагает обнаружение структуры, признаков, скрытых в обучающем наборе неразмеченных данных. Далее на основе обобщения выявленных признаков происходит разделение объектов на кластеры.

Обучение с учителем также, как и обучение без учителя – несомненно важные методы для решения задач распознавания образов, кластеризации, обработки естественного языка, синтеза речи и других задач ИИ. Но в условиях, когда подготовка необходимого количества обучающих данных и само обучение требует больших затрат времени и ресурсов, актуальность приобретает еще один вид машинного обучения – обучение с подкреплением.

1. Основные понятия и элементы системы обучения с подкреплением

Отличительной особенностью обучения с подкреплением является взаимодействие с окружающей средой, которое заключается в восприятии ее состояний, их отображение на действия и получение сигналов обратной связи в виде вознаграждений. Целью обучения с подкреплением является максимизация вознаграждений.

переходов и соответствующих им вознаграждений. Модель выборки гораздо проще получить, и она больше всего подходит для многих алгоритмов обучения с подкреплением.

Таким образом, обучающийся с подкреплением агент взаимодействует с окружающей Сторона, которая обучается и выполняет действия, называется агентом. Агент во многих источниках интерпретируется как техническое устройство, способное воспринимать окружающую среду с помощью датчиков и совершать действия с помощью исполнительных механизмов. Интеллектуальность агента состоит в разработке компьютерной программы, адекватно отображающей последовательность восприятия состояний окружающей среды на действия. Эта программа работает на основе входящего в состав агента вычислительного модуля, который связан с датчиками и исполнительными механизмами [4]. В дальнейшем под агентом будем понимать его интеллектуальную часть в виде компьютерной программы, реализующей соответствующие методы (алгоритмы) ИИ.

Сторона, с которой агент взаимодействует, включающая в себя все, что находится вне агента, называется окружающей средой. Помимо агента и окружающей среды в системе обучения с подкреплением обязательны такие элементы, как стратегия, функция ценности, сигнал вознаграждения и, при необходимости, модель окружающей среды.

Стратегия определяет правило отображения множества состояний среды на действия, выбираемые агентом в этих состояниях. В простом случае стратегия может быть представлена в виде функции или таблицы соответствия, в более сложных задачах используются поиск и аппроксимация неизвестных зависимостей.

Сигнал вознаграждения – это оценка совершенного агентом действия, имеющая либо положительное значение при правильном действии, либо отрицательное – при ошибочном. В общем случае может быть выражена стохастической или постоянной функцией состояния среды и предпринятого действия.

Если сигнал вознаграждения оценивает совершенное действие в текущий момент, то функция ценности дает оценку всех возможных действий агента (состояний среды) из текущего состояния, которые с большой вероятностью встретятся в будущем, и вознаграждений в этих состояниях.

Под моделью окружающей среды будем понимать компьютерную программу, имитирующую поведение окружающей среды в зависимости от действий агента. Существует два основных вида моделей – модель распределения и модель выборки. Модель распределения содержит вероятности переходов всех возможных состояний среды и вознаграждений за возможные действия в этих состояниях. Модель выборки формируется на полученных в ходе обучения статистических данных и позволяет определять и обновлять вероятности одиночных средой или ее моделью непрерывно: агент выбирает действия, а среда реагирует на эти действия, оценивает их и предлагает агенту новые состояния для последующего выбора. Среда оценивает действия агента сигналами вознаграждения, которые агент стремится со временем максимизировать, выбирая наилучшие действия (рисунок 1).

Рисунок 1 – Схема взаимодействия между агентом и окружающей средой

2. Анализ и выбор методов обучения с подкреплением для решения задачи распознавания образов

Множество существующих методов обучения с подкреплением можно классифицировать по трем основным признакам: точности, наличию модели окружающей среды и характеру задач обучения (рисунок 2).

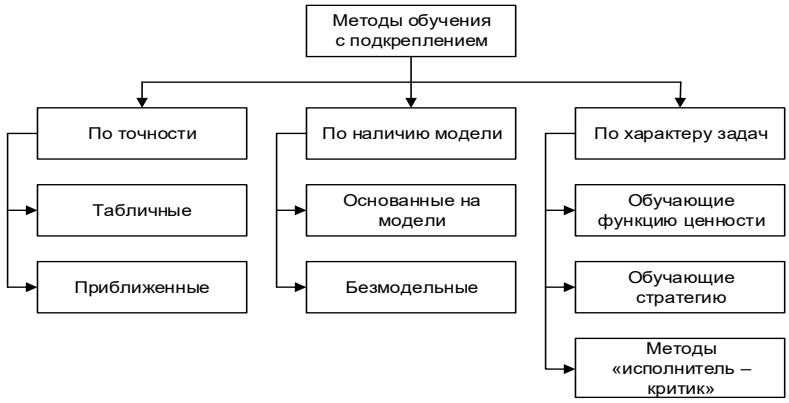

Рисунок 2 – Классификация методов обучения с подкреплением

По точности методы подразделяются на табличные и приближенные. Применение табличных методов предполагает, что пространства состояний внешней среды и действий агента определены и достаточно малы, соответственно функцию ценности можно представить в виде таблиц или массивов данных, что в большинстве случаев позволяет оптимизировать функцию ценности и стратегию.

Во многих задачах обучения с подкреплением пространство состояний может быть неограниченно большим и зачастую неизвестным. Это означает, что учесть все возможные состояния среды, как и найти оптимальное решение в каждом из них, невозможно. В этих условиях и при наличии ограниченных вычислительных ресурсов главной целью является поиск приближенного решения.

Методы обучения с подкреплением, которые предполагают использование модели окружающей среды и планирование действий на ее основе, называются основанными на модели. Другие, более простые методы, в которых обучаемый агент взаимодействует с реальной окружающей средой методом проб и ошибок, являются безмодельными.

Современные тенденции обучения с подкреплением предлагают комбинированное использование указанных методов, что позволяет добиться более надежных результатов обучения. Особенно это проявляется, когда прямое взаимодействие агента с реальной средой может повлечь нежелательные, опасные или даже необратимые последствия. В этом случае обучение производится на имитированном опыте, а реальная среда используется для экспериментального подтверждения.

По характеру задач обучения можно выделить три категории методов:
методы, обучающие функцию ценности состояний предсказывать вознаграждение для выбранной стратегии (задача предсказания);
методы, обучающие стратегию выбирать действия с наибольшей оценкой функции ценности (задача управления);
методы, представляющие собой сочетание первых двух (по типу «исполнитель-критик»).

Проведенный анализ методов обучения с подкреплением показал, что наиболее частым выбором исследователей и разработчиков является обучение на основе временных различий (temporal-difference – TD), которое универсально для решения задач любой сложности и размерности. TD-обучение сочетает в себе свойства, заложенные в методах Монте-Карло и динамическом программировании. Как и методы Монте-Карло, они позволяют обучаться непосредственно на опыте, не требуя модели динамики окружающей среды. Как и динамическое программирование, TD-методы обновляют оценки текущего состояния на основе знания возможных будущих состояний и их вероятностей.

Принимая во внимание результаты анализа и тот факт, что методы «исполнитель-критик», по мнению многих ученых, обладают близким сходством с обучением нейронов головного мозга и показывают достаточно высокие результаты в компьютерных приложениях, для решения задачи распознавания образов целесообразно рассмотреть применение методов «исполнитель-критик» на основе TD-обучения [2; 4].

3. Решение задачи распознавания образов с использованием методов «исполнитель – критик» на основе TD-обучения

При использовании методов «исполнитель-критик» в структуре агента можно выделить два концептуальных компонента. Исполнителем называется компонент, который обучается стратегии, а критиком – компонент, который оценивает выбранную исполнителем стратегию, чтобы «критиковать» и улучшать ее. Критик использует алгоритм TD-обучения, чтобы обучить функцию ценности состояний для текущей стратегии исполнителя. Для этого функция ценности формирует и посылает исполнителю сигнал TD-ошибки. Положительная ошибка сигнализирует о правильном действии, потому что привело в состояние, ценность которого лучше ожидаемой, отрицательная – об ошибочном. Ориентируясь на получаемые критические сигналы, исполнитель постоянно улучшает свою стратегию.

На рисунке 3 показана возможная структура системы обучения «исполнитель-критик» для решения задачи распознавания образов. В данном случае обучение проводится на имитированном опыте, а реальная среда может быть использована для экспериментального подтверждения результатов, полученных с помощью модели. В качестве модели предлагается разрабатывать алгоритмы, позволяющие формировать случайную выборку обучающих примеров $S_t \in S$, имитирующих состояние окружающей среды на каждом временном шаге t. Для генерации сигналов вознаграждения R_t по каждому обучающему примеру алгоритм должен иметь эталонные отклики по типу «объект-ответ». Чем ближе к эталонному полученный от исполнителя отклик, тем выше сигнал вознаграждения.

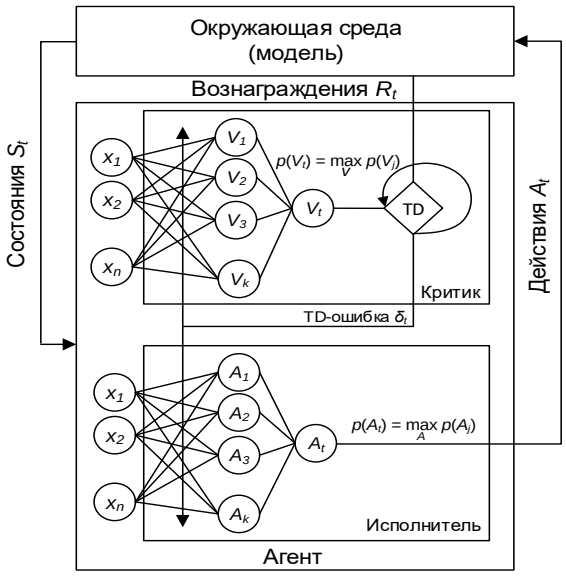

Рисунок 3 – Структура системы обучения «исполнитель-критик»

В рассматриваемом варианте исполнитель и критик представлены в виде глубоких ИНС (показаны как однослойные ИНС, чтобы не усложнять рисунок), которые лучше справляются с неструктурированными данными (изображения, тексты, звуки, речь) и классификацией большого количества объектов. Оба компонента получают одни и те же входные данные в виде вектора признаков $x(S_t) = (x_1(S_t), x_2(S_t), \dots, x_n(S_t))$, характеризующих состояние окружающей среды (модели). С каждой связью любого признака с нейронами скрытого слоя критика и исполнителя ассоциирован вес, представляющий силу синапса. Веса в сети критика параметризируют функцию ценности, а веса исполнителя – стратегию. Сети обучаются по мере того, как эти веса изменяются в соответствии с правилами обучения. TD-ошибка, порождаемая сетью критика, является сигналом подкрепления для изменения весов в сетях критика и исполнителя, что показано на рисунке 3 стрелкой с пометкой «TD-ошибка δ_t».

Количество и метки классов для обеих сетей определяются однообразно в соответствии с количеством категорий объектов в обучающей выборке. Для этого выходные слои ИНС критика и исполнителя должны содержать одинаковое количество k нейронов, помеченных V_j и A_j, $j = 1,2,\dots,k$. Выходом каждого нейрона является элемент k-мерного вектора, который целесообразно представить в виде $softmax$-распределения (мягкого максимума):

$$softmax(V_j) = \frac{e^{V_j}}{\sum_{j=1}^{k} e^{V_j}}. \quad (1)$$

Данная функция позволяет получать вероятности, сумма которых равна 1. В этом случае критик и исполнитель стремятся к выбору наибольшей вероятности $p(V_t) = \max_V p(V_j)$ и $p(A_t) = \max_A p(A_j)$ соответственно.

Помимо этого, критик включает элемент в виде ромба с меткой TD, который на основе объединения сигналов вознаграждения и предыдущих ценностей состояний среды (на что указывает петля, соединяющая выход со входом) вычисляет TD-ошибку δ_t. Эта ошибка сообщает критику направление и абсолютную величину изменения параметров функции ценности, которое приведет к повышению ее предсказательной точности. Критик должен стремиться уменьшить величину $\delta_t = R_{t+1} + V(S_{t+1}, w) - V(S_t, w)$. Иначе говоря, целью обучения критика является вектор весов, минимизирующий значение TD-ошибки:

$$w^* = \operatorname*{argmin}_w (R_{t+1} + V(S_{t+1}, w) - V(S_t, w)).$$

В противоположность критику обучение исполнителя заключается в том, чтобы значение δ_t было положительно и как можно больше по абсолютной величине. Учитывая, что выходные значения ИНС исполнителя интерпретируются как вероятности действий агента, то в качестве функции потерь в большинстве приложений глубокого обучения выбирается отрицательное логарифмическое правдоподобие [5; 6]. В этом случае необходимо найти такой вектор весов θ^*, который, с одной стороны, максимизирует вероятность выбираемого действия, а с другой, минимизирует сумму отрицательных логарифмических правдоподобий N примеров, в которых это действие выбиралось:

$$\theta^* = \operatorname*{argmin}_\theta - \sum_{t=1}^{N} \ln p\,(A_t|S_t, \theta),$$

где $p(A_t|S_t, \theta)$ – параметры стратегии исполнителя, характеризующие вероятность того, что действие A_t выбрано в момент t при условии, что в этот момент окружающая среда (модель) находится в состоянии S_t, а вектор весов равен θ. В каждом состоянии выбираются действия с наибольшей вероятностью согласно распределению $softmax$ (1).

Правила обучения критика и исполнителя используют один и тот же сигнал TD-ошибки δ_t, но его влияние на обучение каждого из них различно. Различие заключается в использовании так называемых следов приемлемости.

При каждом переходе окружающей среды из состояния S_t в состояние S_{t+1} с выбором действия A_t и получением вознаграждения R_{t+1} алгоритм вычисляет TD-ошибку δ_t, а затем обновляет векторы следов приемлемости (z_t^w и z_t^θ) и параметры критика и исполнителя (w и θ) по следующим правилам:

$$\begin{gathered} \delta_t = R_{t+1} + \gamma[V(S_{t+1}, w) - V(S_t, w)]; \\ z_t^w = \lambda^w z_{t-1}^w + \nabla V(S_t, w); \\ z_t^\theta = \lambda^\theta z_{t-1}^\theta + \nabla \ln p\,(A_t|S_t, \theta); \\ w \leftarrow w + \alpha^w \delta_t z_t^w; \\ \theta \leftarrow \theta + \alpha^\theta \delta_t z_t^\theta, \end{gathered} \quad (2)$$

где $\gamma \in [0; 1)$ – коэффициент обесценивания прогнозных оценок функции ценности, $\lambda^w \in [0; 1]$ и $\lambda^\theta \in [0; 1]$ – параметры затухания следов приемлемости, α^w и α^θ – параметры скорости обучения для критика и исполнителя соответственно.

Следы приемлемости z_t^w и z_t^θ – это временные векторы в памяти ИНС критика и исполнителя, которые дополняют постоянные векторы весов w и θ. На каждом временном шаге t в вычислении оценок V_t и A_t учитываются не только векторы весов w и θ, но и соответствующие им z_t^w и z_t^θ, которые сначала резко увеличиваются, а затем медленно уменьшаются пропорционально коэффициентам спадания следа λ^w и λ^θ до нуля. Эффект памяти в ИНС достигается за счет наличия циклических связей в вычислительном графе, когда входные данные, полученные на более ранних этапах, оказывают влияние на отклик сети на текущие данные. В этом случае принято говорить об использовании слоев с рекуррентными соединениями, которые имеют внутреннее состояние или память [2; 5]. Именно такой слой может применяться в скрытых слоях критика и исполнителя, а также в блоке, вычисляющем TD-ошибку.

Вычисление градиентов функции ценности критика $\nabla V(S_t, w)$ и стратегии исполнителя $\nabla \ln p\,(A_t|S_t, \theta)$ в следах приемлемости (2) осуществляется с использованием алгоритма обратного распространения ошибки δ_t и методов стохастического градиентного спуска [1; 2; 7].

После того, как определены правила обновления параметров ИНС критика и исполнителя, в алгоритме обучения с подкреплением необходимо указать количество обучающих примеров и эпох. Параметры ИНС критика и исполнителя, а также количество примеров и эпох могут изменяться в зависимости от результатов обучения. Необходимо отметить, что на первой эпохе обучения значения функции ценности V_t будут нулевыми, пока по всем обучающим примерам рекуррентный блок TD-ошибки не накопит опыт получения вознаграждений. До начала второй эпохи исполнитель обучается без участия критика, напрямую работая с сигналами вознаграждения, т.к. $\delta_t = R_{t+1} + 0 - 0$. Затем значения V_t будут ненулевыми, а предсказательная точность критика с каждой следующей эпохой будет расти. Предпочтения отдаются тем значениям V_t, по которым среднее арифметическое от всех предыдущих вознаграждений больше. Для этого последовательность обучающих примеров должна быть строго одинаковой по всем эпохам.

Для проверки результатов обучения, как и при обучении с учителем, используются проверочные и, при необходимости, подтверждающие данные, отличные по содержанию от обучающей выборки и друг от друга [1; 8]. При обеспечении необходимых условий безопасности для проверки может использоваться определенное количество взаимодействий с реальной (физической) окружающей средой.

В предлагаемом решении задачи распознавания образов в качестве критика и исполнителя были выбраны глубокие ИНС, потому что они способны отображать неразмеченные данные на входе непосредственно в выходы. Что позволяет в значительной мере избежать сложностей разработки признакового описания, как в случае с использованием традиционных алгоритмов машинного обучения.

Заключение

В статье предложен подход к решению задачи распознавания образов, объединяющий метод обучения с подкреплением на основе временных различий с архитектурой «исполнитель-критик» в виде глубоких ИНС и использованием модели окружающей среды.

Данный подход позволяет исполнителю выбирать наилучшие действия с учетом прогноза критика о состоянии окружающей среды, с одной стороны, а критику повышать точность оценки ожидаемого вознаграждения при использовании исполнителем текущей стратегии, с другой. Таким образом, происходит взаимная тонкая подстройка весовых коэффициентов сетей исполнителя и критика, при которой оба обучаются не только на своих ошибках, но и на ошибках друг друга.

Предложенный методический аппарат может быть использован при разработке алгоритмов ИИ на основе обучения с подкреплением как для решения задачи распознавания образов, так и для других задач ИИ в условиях, когда обучение с учителем (без учителя) имеет организационные сложности или требует больших затрат времени и ресурсов.

Вместе с тем важно понимать, что предлагаемый методический подход, как и любой другой в области решения задач на основе глубокого машинного обучения, может быть реализован на практике только экспериментальным путем. Результат практической реализации в каждом отдельном случае будет зависеть от большого количества параметров искусственной нейронной сети, технологии обучения, адекватности модели внешней среды, а также имеющегося научно-технического задела и возможностей разработчика. В то же время обучение с подкреплением может продолжаться и в ходе эксплуатации предобученных систем искусственного интеллекта. Возможных способов практического применения данного подхода может быть неограниченное множество, и получаемые результаты могут использоваться только для подтверждения или опровержения какого-либо из конкретных способов, но не самого подхода. В связи с этим варианты практической реализации не являются предметом настоящих исследований и не рассматриваются в данной статье.

Несмотря на впечатляющие успехи, глубокое обучение с подкреплением в настоящее время редко применяется в коммерческих целях по причинам, связанным со сложностями использования физической среды или разработки ее модели, непосредственно влияющими на качество обучения. Тем не менее, учитывая, что развитие обучения с подкреплением тесно связано с нейронаукой, изучающей механизмы обучения в головном мозге, это одна из самых перспективных и активно исследуемых междисциплинарных областей машинного обучения [2; 9; 10].

Список использованных источников

1. Горский А.С., Никоноров В.И. Распознавание образов на поле боя на основе применения искусственных нейронных сетей // Стратегическая стабильность. 2021. №3(96). – С. 63-66.

2. Саттон Р.С., Барто Э.Дж. Обучение с подкреплением: Введение. 2-е изд. М.: ДМК Пресс, 2020. – 552 с.

3. Жиленков А.А., Силкин А.А., Серебряков М.Ю., Колесова С.В. Сравнительный анализ систем глубокого обучения с подкреплением и систем обучения с учителем // Известия Тульского государственного университета. Технические науки. 2022. №10. – С. 109-112.

4. Рассел С., Норвиг П. Искусственный интеллект: современный подход. 4-е изд. Т.1: Решение проблем: знания и рассуждения. СПб.: Диалектика, 2021. – 704 с.

5. Рассел С., Норвиг П. Искусственный интеллект: современный подход. 4-е изд. Т.3: Обучение, восприятие и действие. СПб.: Диалектика, 2022. – 640 с.

6. Толстых А.А., Ступников Д.С., Малюков С.В., Лукьянов А.С., Лунев Ю.С. Применение метода обучения с подкреплением в роботизированных и автоматизированных системах лесной промышленности // Лесотехнический журнал. 2020. Т.10. №1(37). – С. 256-265.

7. Горелик А.Л., Скрипкин В.А. Методы распознавания. 4-е изд. М.: Высшая школа, 2004. – 262 с.

8. Хайкин С. Нейронные сети. 2-е изд. М.: Вильямс, 2019. – 1104 с.

9. Намиот Д.Е., Ильюшин Е.А., Чижов И.В. Военные применения машинного обучения // International Journal of Open Information Technologies. 2022. Т.10. №1. – С. 69-76.

10. Ведяхин А.А. Сильный искусственный интеллект: на подступах к сверхразуму. М.: Интеллектуальная Литература, 2021. – 232 с.