УДК: 005.5
DOI: 10.28995/2073-6304-2024-1-38-48

Принятие управленческих решений на основе метода обучения с подкреплением

Татьяна Н. Боровик

МИРЭА – Российский технологический университет
Москва, Россия, borovik@mirea.ru

Роман В. Шамин

МИРЭА – Российский технологический университет
Москва, Россия, roman@shamin.ru

Аннотация. В статье рассматриваются вопросы оценки управляющих решений при повторяющихся задачах. Показано, что использование методов машинного обучения с подкреплением позволяет повысить объективность оценок возможных управленческих решений. Показно, что предлагаемый метод имеет методическую связь с теоретико-игровым подходом и может быть обобщен для различных вариантов при практическом применении.

Ключевые слова: управленческие решения, обучение с подкреплением, теория игр

Для цитирования: Боровик Т.Н., Шамин Р.В. Принятие управленческих решений на основе метода обучения с подкреплением // Вестник РГГУ. Серия «Экономика. Управление. Право». 2024. № 1. С. 38–48. DOI: 10.28995/2073-6304-2024-1-38-48

Management decisions based on reinforcement learning method

Tat'yana N. Borovik

MIREA – Russian Technological University, Moscow, Russia, borovik@mirea.ru

Roman V. Shamin

MIREA – Russian Technological University, Moscow, Russia, roman@shamin.ru

Abstract. The article considers the issues of evaluating control decisions for repetitive tasks. It is shown that using the reinforcement learning method makes it possible to increase the objectivity of assessments of possible management decisions. Authors highlight that the proposed method has a methodological connection with the game-theoretic approach and can be generalized for various options in practical application.

Keywords: management decisions, reinforcement learning, game theory

For citation: Borovik, T.N. and Shamin, R.V. (2024), "Management decisions based on reinforcement learning method", *RSUH/RGGU Bulletin. "Economics. Management. Law" Series*, no. 1, pp. 38-48, DOI: 10.28995/2073-6304-2024-1-38-48

Введение

Современный уровень цифровизации организаций налагает новые требования к механизмам принятия управленческих решений. Принятие решений – это всегда выбор одного варианта из множества. Если решение принимается одним лицом, то рациональное принятие решений основано, как правило, на оценке полезности каждого из возможных вариантов принимаемых решений. В зависимости от характера принимаемых решений используемая оценка может быть в той или иной мере субъективной. В задачах управления работниками субъективная составляющая оценки принимаемого решения во многом определяется субъективным отношением лица, принимающего решение, к своим подчиненным. Более того, если регулярно принимаются однотипные решения относительно одинакового круга подчиненных, то субъективная оценка будет усиливаться. В этой ситуации возникает две проблемы. Во-первых, как определить, что используемая субъективная оценка соответствует объективной оценке? Во-вторых, как найти объективную оценку возможных решений? При этом решение второй задачи включает в себя решение первой задачи.

Настоящая работа посвящена методам повышения объективности оценки возможных решений при решении повторяющихся задач. Разумеется, что в реальности объективной оценки решений может и не быть, поскольку результаты принятия решения не всегда могут быть сравнимы. Так, какое-то решение может быть выгоднее с одной стороны (например, с точки зрения материальных затрат), но менее выгодным с другой стороны (например, с точки зрения времени выполнения). Мы ставим задачу не нахождения оптимального (объективного) решения, а повышения объективности при оценке возможного решения. Отметим, что речь идет только об оценке решения, а не об оптимальности решений.

Метод обучения в подкрепление, относящийся к методам искусственного интеллекта, использует функционал для оцен-

ки возможных действий в заданном состоянии [Саттон, Барто 2020]. При этом метод позволяет находить и уточнять этот функционал по получаемым результатам при выполнении выбранных действий. Этот подход хорошо себя зарекомендовал при решении различных игровых задач и при управлении техническими системами [Николенко 2009]. В настоящей работе мы используем идеи, связанные с методом обучения с подкреплением, для получения объективных оценок возможных решений в задачах менеджмента.

Формальная постановка задачи

Пусть лицо, принимающее решения, должно сделать выбор своего решения из фиксированного множества возможных решений. Это множество обозначим через P. Будем считать, что множество P является конечным множеством. Ситуации, когда множество P может оказаться бесконечным, соответствуют, например, выбору объема финансирования. Далее, мы будем предполагать, что задача выбора решения в рассматриваемой ситуации является повторяющейся.

Пусть у лица, принимающего решение, есть априорная оценка каждого возможного действия из множества P. Таким образом, задана функция $Q_0(p)$ для каждого p из P. Функция $Q_0(p)$ является безразмерной числовой функцией. Мы будем говорить, что решение p_1 хуже решения p_2, если имеет место неравенство

$$Q_0(p_1) < Q_0(p_2).$$

В этой ситуации выбор решения может быть осуществлен следующим образом: необходимо выбрать такое решение p^* из множества P, что для выполняется неравенство

$$Q_0(p_1) \leqslant Q_0(p^*)$$

для всех p из P. Такой выбор имеет называется «жадным» выбором [Кормен и др. 2005]. Жадный выбор является оправданным в случае, если субъективный выбор также является и оптимальным по результатам использования этого решения. Кроме функции $Q_0(p)$ рассмотрим также функцию $R(p)$, определенную на множестве P. Функция $R(p)$ также будет числовой функцией, но эта функция может быть размерной. Функция

$R(p)$ имеет смысл как результат использования решения p. Размерность этой функции может быть рубли, временная шкала или какая-либо еще шкала.

Поскольку значение $R(p)$ становится известным лишь после принятия решения р, то, как правило, лицо, принимающее решение, не может априорно знать значения этой функции, и поэтому не может осуществлять выбор решения на основании функции $R(p)$.

После того, как было принято решение p^*, становится известным значение функции $R(p^*)$. Кроме того, нам известны и результаты предыдущих выбранных решений. На основании этих данных лицо, принимающее решение, может скорректировать свою первоначальную оценку решения. Это изменение выражается в изменении значения или значений функции $Q_0(p)$. Измененную функцию мы будем обозначать через $Q(p)$. Таким образом, мы будем рассматривать следующую процедуру:

$$Q = F(Q_0, R(p)).$$

Процедура оптимизации оценки решений

Рассмотрим подробно процедуру оптимизации оценки повторяющихся решений. Пусть мы имеем априорную или сложившуюся оценку каждого управляющего решения. Если лицо, принимающее решение, будет использовать жадный выбор, то, при прочих равных, он всегда будет выбирать одинаковое решение. Идея обучения с подкреплением состоит в том, что перед очередным принятием решения с определенной вероятностью, которую обозначим через $q > 0$, необходимо выбрать произвольное допустимое решение р из множества P. После реализации выбранного решения следует учесть полученный результат с помощью функции $R(p)$.

Опишем эту процедуру формально. В качестве входных данных мы используем:

P – множество возможных решений;

$Q_0(p)$ – существующая оценка каждого решения;

R_0 – объективный максимальный результат, который был достигнут при решении рассматриваемой задачи;

q – вероятность выбора нового решения.

Шаг 1. Получаем случайное число, равномерно распределенное от 0 до 1, которое обозначим через s.

Шаг 2. Если $s < q$, то переходим к шагу 6.

Шаг 3. Находим такое p^* из множества P, которое удовлетворяет следующему условию

$$Q(p^*) = \max_{p \in P} Q(p).$$

Шаг 4. Применяем решение p^*, найденное на шаге 3.

Шаг 5. Перейти к шагу 9.

Шаг 6. Выбираем случайное решение из множества P. Обозначим это решение через p_1.

Шаг 7. Применяем решение p_1 и получаем объективный результат, который обозначим через R_1.

Шаг 8. Обновляем оценку $Q_0(p_1)$ по следующей формуле:

$$Q_0(p_1) = Q_0(p_1) \cdot \left(1 + \frac{R_1 - R_0}{R_0}\right).$$

Шаг 9. Конец алгоритма.

Заметим, что на шаге 8 коррекция субъективной оценки решения будет в большую сторону, если достигнутый результат окажется лучше, чем ранее наилучший результат, либо в меньшую сторону, если достигнутый вариант окажется хуже наилучшего результата.

Варианты использования процедуры оптимизации оценки

При практическом использовании процедуры оптимизации оценки решений можно использовать следующие вариации описанной выше процедуры.

Использование случайных чисел может быть заменено на периодическое использование выбора случайных решений задачи, например, при каждом n-ом решении задачи. Такой подход позволяет более систематично использовать процедуру коррекции оценки управляющих решений.

В случае, когда в результате использования процедуры оптимизации оценки управляющих решений какое-либо решение получает низкую оценку, то следует такое решение исключить из числа рациональных решений и из множества P. Следовательно, предложенная процедура позволяет не только уточнять субъективную оценку возможных решений, но и осуществлять процедуру отбора рациональных управляющих решений.

Следует учитывать, что объективный результат R_1, как правило, является случайной величиной, которая зависит от очень большого числа факторов. Кроме того, эта величина не всегда является стационарной, поскольку результат управляющего решения может завить от меняющейся внешней ситуации. Таким образом, мы имеем следующую случайную функцию (случайный процесс):

$$R_1 = R_1(t, \omega, p_1),$$

где t – это время, ω – это случайный исход, который отражает случайные факторы, влияющие на результат управляющего решения.

В случае, когда функция R_1 подвержена случайным факторам, которые могут существенно влиять на результат, то в результате величина R_0 также может иметь различные значения в результате случайных факторов. В этом случае вместо R_0 следует использовать среднее значение (математическое ожидание).

Многошаговое управление

Рассмотренный ранее алгоритм для оценки управляющих решений может быть обобщен на случай многошагового управления, когда для решения определенной задачи необходимо принимать ряд управляющих решений. Если исходная задача может быть декомпозирована на отдельные последовательные шаги, то мы можем применять процедуру оценки управляющих решений к каждому шагу, сведя многошаговое управление к последовательности одношаговых управлений. Однако более интересным является случай, когда выбор управляющего решения зависит от ситуации, которая в свою очередь складывается из предыдущих управленческих решений. Общая схема приведена на рис. 1.

В рассматриваем случае много шагового управления мы вводим понятие «состояние» как набор возможных управленческих решений, соответствующих текущему состоянию. В состояние, соответствующему конечному состоянию, мы будем считать множество управляющих решений пустым. Если состояние имеет не пустое множество возможных решений, то, применяя какое-либо возможное решение, мы, во-первых, оказываемся в другом состоянии, а во-вторых, получаем определенный результат (подкрепление) [Уиндер 2023].

Рис. 1. Пример многошагового управления

Для вычисления оптимальной оценки управляющих решений в этой ситуации следует использовать методы обучения с подкреплением (Q-learning, SARSA) [Саттон 2020].

Пример применения процедуры оценки управляющих воздействий

Предположим, что для выполнения заказа на изготовление изделия могут быть применены следующие различные технологии:

1) Технология 1, основанная на использовании токарно-фрезерного станка;
2) Технология 2, основанная на использовании 3D-принтера;
3) Технология 3, основанная на смешанном использовании токарно-фрезерного станка и 3D-принтера.

В этом случае множество возможных решений имеет вид:

$$P = \{\textit{Технология 1}, \textit{Технология 2}, \textit{Технология 3}\},$$

понимая под решением «Технология N» – выбор Технологии N для изготовления изделия. Будем полагать, что аналогичные задачи ставятся регулярно, что означает повторяющиеся решения.

Лицо, принимающее решение, – главный технолог имеет следующие априорные оценки для каждого решения:

Q_0(*Технология 1*) = 0,8;

Q_0(*Технология 2*) = 0,2;

Q_0(*Технология 3*) = 0,5.

Пусть при ранее применявшихся решениях («Технология 1») наилучший объективный результат имел следующее значение:

R_0 = 120,000 руб.

При использовании (согласно процедуре оценки решений) нового решения – «Технология 3» был достигнут объективный результат

R_1 = 150,000 руб.

Согласно процедуре оценки решений, мы должны скорректировать оценку для решения «Технология 3» следующим образом

$$Q_0(\textit{Технология 3}) = 0{,}5 \cdot (1 + (150000 - 1200000)/120000) = 0{,}625.$$

При последующем применении процедуры оценки управляющих решений была получена следующая динамика изменения оценок всех возможных решений, которая представлена на рис. 2.

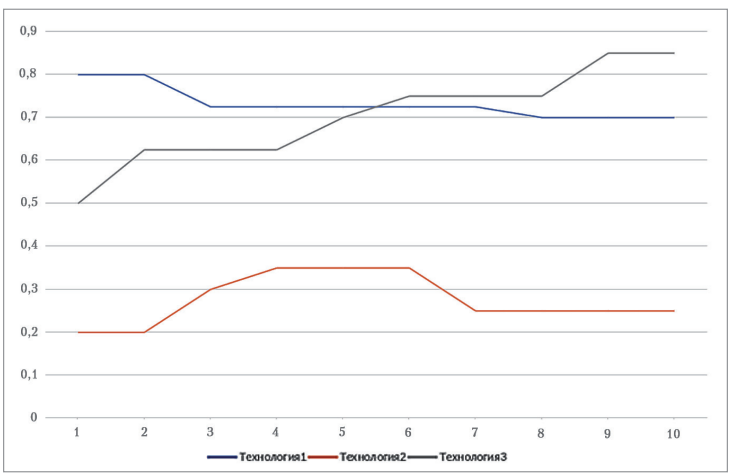

Рис. 2. Динамика изменения оценок решений

В рассматриваемом примере мы видим, что решение «Технология 3» в итоге получила максимальную оценку и должна быть использована.

Связь с теорией игр

Предложенная в работе процедура оценки управляющих решений и последующего выбора решений определенным образом связана с методами теории игр. Одним из основных методов выбора стратегии (решения из множества возможных решений) в теории игр является метод смешанных стратегий [Нейман, Моргенштерн 1970]. Суть метода смешанных стратегий состоит в том, что выбирается не фиксированная стратегия, а стратегия выбирается случайным образом из всех возможных. Разумеется, что стратегия выбирается не произвольным образом, согласно определенным вероятностям.

Пусть множество стратегий состоит из N элементов, которые мы обозначим через

$$s_1, s_2, ..., s_N.$$

С каждой стратегией мы свяжем вероятность выбора этой стратегии:

$$s_1 \to q_1,\ s_2 \to q_2,\ ...,\ s_N \to q_N.$$

Выбор вероятностей рассчитывается согласно платежной матрице, которая задает игру. При выборе вероятностей должно выполняться следующее соотношение:

$$q_1 + q_2 + ... + q_N = 1.$$

После расчета вероятностей мы выбираем стратегию согласно полученным вероятностям.

В нашем случае также можно построить теоретико-игровую интерпретацию случайному выбору управляющего решения таким образом, чтобы используемые вероятности соответствовали вероятностям стратегий в некоторой игре.

Заключение

В статье рассмотрен метод объективной оценки управленческих решений при повторяющихся задачах. Этот метод позволяет уточнять априорную оценку возможных решений и выбирать наиболее эффективное решение. Изменение оценки решений происходит итеративным методом, что позволяет избежать резкого изменения в оценках решений. Это важно, поскольку при оценке решений учитываются случайные факторы, которые могут исказить оценку решения.

В работе рассмотрены различные обобщения процедуры оценки управляющих решений, связанные с задачами, которые требуют многошагового выбора решений, а также учета случайных факторов, которые влияют на результат применения управляющих решений.

Приведен пример использования предложенной процедуры выбора управляющих решений, который демонстрирует применение предложенного метода оценки управляющих решений.

Литература

Кормен и др. 2005 – *Кормен Т., Лейзерсон Ч., Ривест Р., Штайн К.* Алгоритмы: построение и анализ = Introduction to Algorithms. М.: Вильямс, 2005. 894 с.

Нейман, Моргенштерн 1970 – *Нейман Дж. фон, Моргенштерн О.* Теория игр и экономическое поведение. М.: Наука, 1970. 707 с.

Николенко 2009 – *Николенко С.И., Тулупьев А.Л.* Самообучающиеся системы. М.: МЦНМО, 2009. 287 с.

Саттон, Барто 2020 – *Саттон Р.С., Барто Э.Г.* Обучение с подкреплением = Reinforcement Learning. М.: ДМК пресс, 2020. 552 с.

Уиндер 2023 – *Уиндер Ф.* Обучение с подкреплением для реальных задач. Инженерный подход. М.: БХВ-Петербург, 2023.

References

Cormen, T., Leiserson, Ch., Rivest, R. and Stein, K. (2005), *Algoritmy: postroenie i analiz* [Algorithms. Construction and analysis = Introduction to Algorithms], Williams, Moscow, Russia.

Neumann, J. von and Morgenstern, O. (1970), *Teoriya igr i ekonomicheskoe povedenie* [Theory of Games and economic behavior], Nauka, Moscow, Russia.

Nikolenko, S.I. and Tulupyev, A.L. (2009), Self-learning systems, MTsNMO, Moscow, Russia.

Sutton, R.S. and Barto, E.G. (2020), *Obuchenie s podkrepleniem* [Reinforcement Learning], DMK press, Moscow, Russia.

Winder, Ph. (2023), *Obuchenie s podkrepleniem dlya real'nykh zadach. Inzhenernyi podkhod* [Reinforcement learning for actual tasks. Engineering approach], BKhV-Peterburg, Moscow, Russia.

Информация об авторах

Татьяна Н. Боровик, МИРЭА – Российский технологический университет, Москва, Россия; 119454, Россия, Москва, пр. Вернадского, д. 78; borovik@mirea.ru

Роман В. Шамин, доктор физико-математических наук, МИРЭА – Российский технологический университет, Москва, Россия; 119454, Россия, Москва, пр. Вернадского, д. 78; roman@shamin.ru

Information about authors

Tat'yana N. Borovik, MIREA – Russian Technological University, Moscow, Russia; bld. 78, Vernadskii Avenue, Moscow, Russia, 119454; borovik@mirea.ru

Roman V. Shamin, Dr. of Sci. (Physics and Mathematics), MIREA – Russian Technological University, Moscow, Russia; bld. 78, Vernadskii Avenue, Moscow, Russia, 119454; roman@shamin.ru