

Применение стохастических методов в машинном обучении для решения финансовых задач

Милюшков Георгий, гр. 20.Б04-мм

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доцент Шпилёв П. В.
Рецензент: к.ф.-м.н., лектор Пепелышев А. Н.



Санкт-Петербург
2024г.

- ① Биржевой трейдер для 1 компании
- ② Повторение результатов статьи (Hongyang, 2020) для реализации биржевого трейдера для 30 компаний
- ③ Альтернативный подход, основанный на других стохастических алгоритмах
- ④ Результаты по повышению эффективности инвестиционного портфеля

Постановка задачи

Общая постановка задачи

Применение стохастических методов в машинном обучении для решения финансовых задач

Исследование состояло из четырех ключевых частей:

- ① Изучение стохастических методов обучения с подкреплением и разработка программной реализации трейдера, торгующего акциями одной компании на основе алгоритма Q-обучение
- ② Изучение специальной литературы для определения перспективного подхода, основанного на применении стохастических методов глубокого обучения с подкреплением для торговли акциями 30 компаний
- ③ Реализация подхода, предложенного автором (Hongyang, 2020), и разработка модификации данного подхода, основанная на использовании более современных алгоритмов
- ④ Оценка эффективности разработанной программной реализации

Часть 1: Обучение с подкреплением

Обучение с подкреплением (Reinforcement Learning) – область машинного обучения, в которой обучение осуществляется посредством взаимодействия с окружающей средой.



Часть 1: Функция ценности и Q-функция

Функция ценности (Value function)

$$V_{\pi}(s) = \mathbb{E}_{\pi}[G|S_0 = s] = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s \right]$$

R – награда, γ – дисконтирующий коэффициент,
 S_0 – начальное состояние

- Значения данных функций представляются в виде таблицы

Q-функция

$$\begin{aligned} Q_{\pi}(s, a) &= \mathbb{E}_{\pi}[G|S_0 = s, A_0 = a] = \\ &= \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s, A_0 = a \right] \end{aligned}$$

Часть 1: Создание биржевого трейдера

Цель

Максимизация прибыли при торговле акциями 1 компании

- Метод Q-обучение
- Используются исторические данные акции Apple
- Действия(a) — покупка, продажа или удержание
- Состояние(s) — скорректированная цена закрытия
- Дисконтирующий коэффициент(γ) — 0.95
- Коэффициент обучения(α) — 0.8
- Случайное действие(ε) — 30%
- 1 тренировочный цикл

Часть 1: Результаты

- Создан торговый алгоритм который работает в среде с высоким уровнем неопределенности
- Тренировочные данные **до изменений**: прибыльность на 12% лучше, тестовые на 14% хуже
- Улучшаем модель добавляя технические индикаторы: SMA, линии Боллинджера и больше циклов
- Тренировочные данные **после изменений**: прибыльность на 648% лучше, тестовые на 33% лучше

Часть 1: Результаты

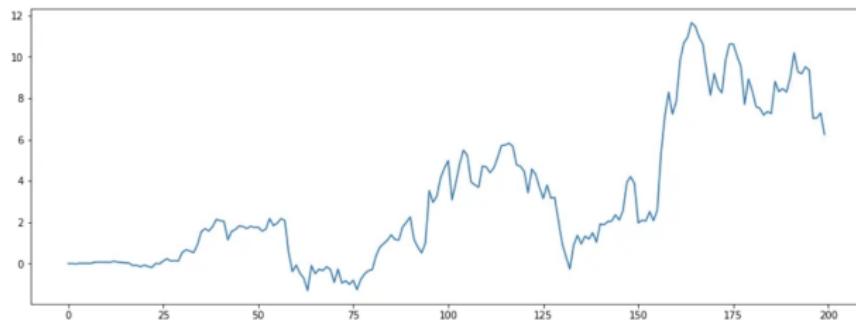


Рис.: График дохода

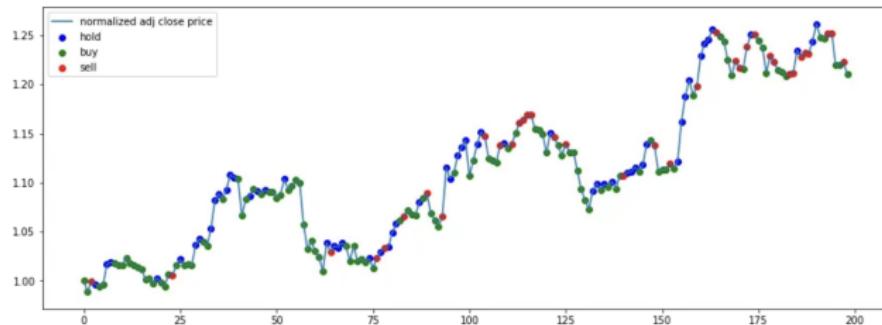
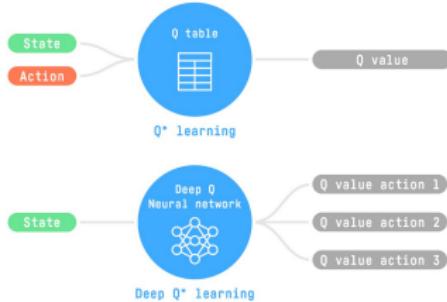


Рис.: График действий агента

Часть 2: Глубокое обучение с подкреплением



- Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy (Hongyang, 2020)
- Методы классического машинного обучения с подкреплением неэффективны для решения задач в сложных и многомерных пространствах
- Цель заключается в разработке биржевого трейдера, на основе стохастических алгоритмов A2C, PPO, DDPG и их сочетания способного проводить торговлю портфелем акций

Часть 2: Актор-Критик

Нейронная сеть актор (policy based)

Основная цель актора — улучшение политики. Принимает на вход текущее состояние среды и генерирует вероятностное распределение по возможным действиям.

Нейронная сеть критик (value based)

Основная цель критика — улучшение точности оценок ценности состояний или действий. Принимает на вход текущее состояние среды и оценивает, насколько хорошо было выбрано действие

Цель — максимизация ожидаемого суммарного вознаграждения

$$G_t = \mathbb{E} \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s \right]$$

$\gamma \in [0, 1)$ — коэффициент дисконтирования, R_{t+k+1} — вознаграждение, S_t — состояние на шаге t

Часть 2: Используемые алгоритмы

Ключевые различия			
	Advantage Actor Critic (A2C)	Proximal policy optimization (PPO)	Deep Deterministic Policy Gradient (DDPG)
Особенности обновления параметров в стохастическом градиентном спуске параметров сети	Параметры обновляются синхронно, т.е. все агенты выполняют свои шаги параллельно, после чего происходит общее обновление.	Параметры обновляются с использованием функции срезания (clipping) вероятностных коэффициентов.	Применяет целевые сети и повторное воспроизведение опыта
Количество нейронных сетей	1 (2)	1 (2)	4
Отличительная черта	Простота реализации, использование функции преимущества для уменьшения дисперсии градиента	Стабильность, благодаря функции срезания, что предотвращает слишком большие обновления параметров.	Быстро адаптируется к изменяющимся рыночным условиям. Использование отдельных сетей для актора и критика.

Коэффициент Шарпа

$$S = \frac{\mathbb{E}(R - R_f)}{\sigma},$$

где R — доходность за данный период, R_f — доходность без риска, то есть прибыль, если бы ничего не делалось, σ — стандартное отклонение доходности портфеля.

- Тренировочные данные с 2009 по 2015. Тестовые данные с 2016 по 2020
- Состояние из 4 финансовых технических индикаторов
- 3 действия для каждой компании:
 - Покупка
 - Продажа
 - Удержание
- Стохастический выбор действий

Часть 3: Результаты

- В точности повторены результаты задачи стохастической оптимизации



- Наиболее прибыльный алгоритм — PPO, наиболее стабильный — алгоритм сочетания

Часть 4: Альтернативный подход

Мотивация:

Оригинальная программа реализована неоптимально.

Бесконфликтная работа программы возможна только при соблюдении большого списка требований к системе компьютера.

Ключевые различия		
	Soft Actor-Critic (SAC)	Twin Delayed DDPG (TD3)
Подход к обновлению параметров	Целевая функция для критиков основана на минимуме из двух сетей для уменьшения переоценки. Актор обновляется с учетом энтропийной регуляризации.	Параметры критиков обновляются путем минимизации ошибки между предсказанным Q-значением и целевым значением. Параметры актора обновляются реже, чем параметры критиков, чтобы избежать нестабильности.
Количество нейронных сетей	3	3
Отличительная черта	Включает энтропийный термин в целевую функцию для поощрения исследования. Надстройка DDPG. Стабильный с малой вероятности переоценки.	Задержка в обновлении актора. Добавление шума в целевую функцию. Устойчивый, стабильный и менее склонный к резким изменениям оценок

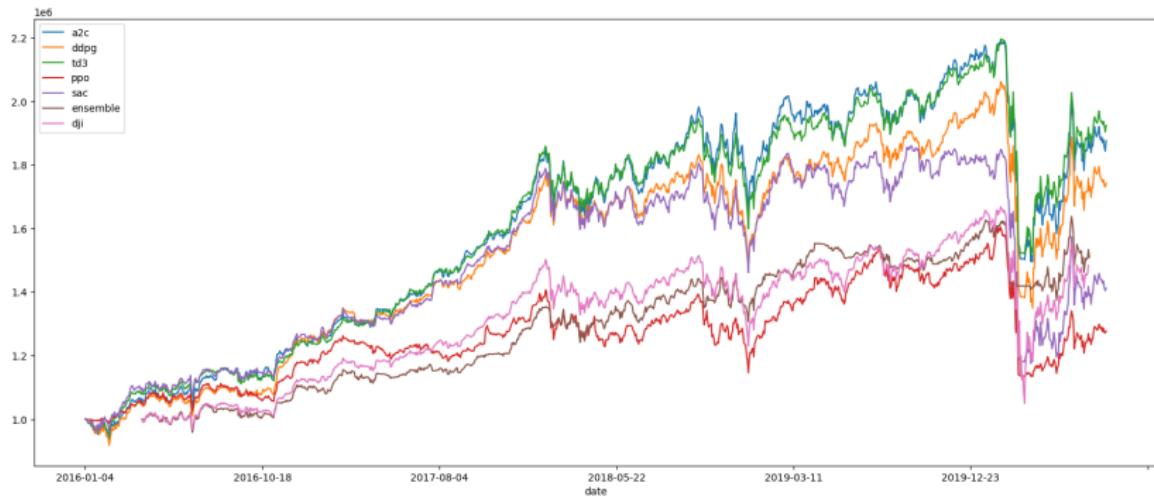
Часть 4: Альтернативная реализация

Были выполнены следующие модификации:

- ❶ Реализованы более современные алгоритмы: SAC и TD3
- ❷ Произведена модификация алгоритмов A2C и PPO, основанная на включении в целевую функцию дополнительного слагаемого энтропии, для удержания алгоритмов от ранней конвергенции к подоптимальным решениям
- ❸ Модифицированы используемые параметры
- ❹ Добавлены 4 новых технических индикатора
- ❺ Исправлены ошибки в данных
- ❻ Реализована возможность запуска программы в системе Google Colab
- ❼ Программа реализуется в одном файле при использовании библиотеки FinRL

Часть 4: Результаты

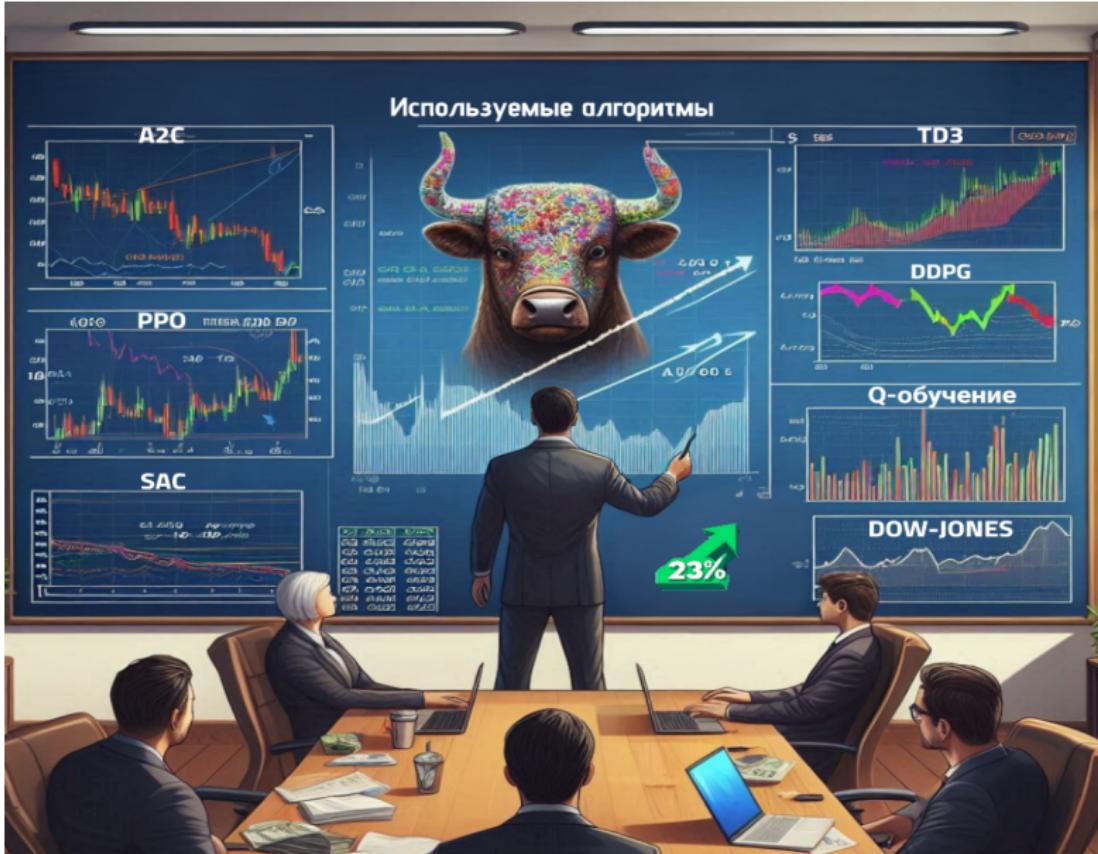
- Реализованы более доступным и эффективным путем: A2C, PPO, DDPG, алгоритм сочетания, SAC, TD3
- Прибыльность увеличена на 23% за период торговли с 2016 по 2020
- TD3 самый прибыльный



Мною были выполнены следующие задачи:

- ❶ Разработана программная реализация для торговли акциями одной компании, применяя стохастический алгоритм Q-обучения
- ❷ Для программной реализации трейдера работающего с акциями 30 компаний изучена специальная литература в частности посвященная алгоритмам A2C, PPO, DDPG и их сочетании
- ❸ Повторены результаты авторов и предложен более эффективный подход, основанный на использовании более современных алгоритмов для торговли портфелем акций 30 компаний
- ❹ Разработана программная реализация предложенного подхода, что позволило увеличить эффективность программы и прибыльность на 23%

Спасибо за внимание



- Состояние: 181 размерный вектор ($30 \cdot 6 + 1$)
 - Скорректированная цена закрытия
 - Объем имеющихся акций компании
 - MACD (Moving Average Convergence/Divergence) – позволяет оценивать силу тренда
 - RSI (Relative strength index) – соотношение положительных и отрицательных изменений
 - CCI (Commodity Channel Index) – измеряет отклонение цены инструмента от его среднестатистической цены
 - ADX (Average Directional Index) – используется для определения наличия или отсутствия тренда, его направления и силы
- Пространство действий для одной компании $\{-k, \dots, -1, 0, 1, \dots, k\}$, следовательно всего $(2k + 1)^{30}$
- Добавлены линии Болленаджера, DX (направление и сила), и SMA (60 и 30)

Технический слайд 2

	Коэффициенты в моей программе	Коэффициенты в статье[26]
A2C	Шаги в среде: 5 Энтропия: 0.01 Обучение (α): 0.0007 Выборка обучения: 50 000	Шаги в среде: 5 Обучение (α): 0.0005 Выборка обучения: 10 000
PPO	Шаги в среде: 2048 Энтропия: 0.01 Обучение (α): 0.00025 Размер пакета (Batch size): 128 Выборка обучения: 50 000	Шаги в среде: 2048 Обучение (α): 0.0005 Размер пакета (Batch size): 128 Выборка обучения: 10 000
DDPG	Размер буфера: 50 000 Обучение(α): 0.001 Размер пакета (Batch size): 128 Выборка обучения: 50 000	Размер буфера: 50 000 Обучение(α): 0.005 Размер пакета (Batch size): 128 Выборка обучения: 50 000
TD3	Размер буфера: 100 000 Энтропия: auto_0.1 Обучение(α): 0.001 Размер пакета (Batch size): 100 Выборка обучения: 50 000	x
SAC	Максимальный размер буфера: 100 000 Коэффициент обучения(α): 0.0001 Размер пакета (Batch size): 128 Кол. эпи. в начальной фазе: 100 Выборка обучения: 50 000	x

Технический слайд 3 — ключевые формулы

Градиент целевой функции обновляющий значение

актора(A2C): $\nabla J_\theta(\theta) = \mathbb{E} \left[\sum_{t=1}^T \nabla_\theta \log \pi_\theta(a_t | s_t) A(s_t, a_t) \right]$

Функция преимущества:

$$A(s_t, a_t) = r(s_t, a_t, s_{t+1}) + \gamma V(s_{t+1}) - V(s_t)$$

Обновление критика: $\varphi \leftarrow \varphi - \alpha \nabla_\varphi (R + \gamma V_\varphi(s') - V_\varphi(s))^2$

Обновление критика(DDPG) минимизируя функцию потерь:

$$L_Q(\phi) = \mathbb{E}_{s_t, a_t, r_t, s_{t+1} \sim R} \left[(Q_\phi(s_t, a_t) - q_t)^2 \right]$$

Обновление актора:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_t \sim R} \left[\nabla_\theta \mu_\theta(s_t) \nabla_a Q_\phi(s_t, a) \Big|_{a=\mu_\theta(s_t)} \right]$$

Функция срезания:

$$J^{CLIP}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) A(s_t | a_t), \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon) A(s_t, a_t) \right) \right]$$

SAC: $\pi^* = \arg \max_\pi \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot | s_t)) \right) \right]$

- Основано на Q-функции и обновляется по следующей формуле:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha \left(R_{t+1} + \gamma \max_a Q_{t+1}(s_{t+1}, a) \right)$$

α – коэффициент обучения,

$R_{t+1} = R(s_{t+1}, a_{t+1})$ – награда, t – момент времени

ε -жадная стратегия

$$a_t = \begin{cases} \max_a Q(s_t, a), & \text{с вероятностью } 1 - \varepsilon \\ \text{любой } a_t & \text{с вероятностью } \varepsilon \end{cases}$$

Марковский процесс принятия решений

- Задается множеством состояний \mathcal{S} и множеством действий \mathcal{A}
- $P_a(s, s') = P(\mathcal{S}_{t+1} = s' | \mathcal{S}_t = s, \mathcal{A}_t = a)$ – вероятность
- $R_a(s, s')$ – награда
- t – момент времени

Политика π – определяет поведение агента в среде

- $\pi : \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$
- $\pi(a, s) = P(\mathcal{A}_t = a | \mathcal{S}_t = s)$