DOI 10.31891/2307-5732-2023-327-5-132-138
УДК 004.93, 004.8

ХАВАЛКО ВІКТОР
Національний університет "Львівська політехніка"
https://orcid.org/0000-0002-9585-3078
e-mail: viktor.m.khavalko@lpnu.ua
ДОХНЯК БОГДАН-ОЛЕКСАНДР
Національний університет "Львівська політехніка"
https://orcid.org/0000-0003-4911-8950
e-mail: bohdan-oleksandr.o.dokhniak@lpnu.ua
СЛАВА ЛЮБОМИР
Національний університет "Львівська політехніка"
e-mail: liubomyr.slava.mknssh.2022@lpnu.ua

ЗАСТОСУВАННЯ НЕЙРОМЕРЕЖЕВИХ ПІДХОДІВ ДО ВИРІШЕННЯ ЗАДАЧІ ПРО БАГАТОРУКОГО БАНДИТА

Основна проблема більшості людей – незнання того, як зробити перший крок в сфері інвестування власних коштів на фінансовому. Люди, які хочуть почати займатись інвестуванням зазвичай не знають з чого почати та акції яких компаній можна вигідно перепродати. В статті проведено аналіз та порівняння восьми базових алгоритмів для вирішення задачі про багаторукого бандита. Для цього було спроектовано та розроблено відповідне середовище для досліджень, яке дозволило спостерігати за поведінкою алгоритмів впродовж семи років. Середовище, максимально наближене до реального і це дало можливість проаналізувати поведінку агентів в симуляції та зробити відповідні висновки щодо їхньої ефективності.

Створена нова модифікація жадібного агента, який замість власних оцінок використовує передбачення, сформовані рекурентними нейронними мережами (запропоновано підхід, який поєднує в собі можливості штучного інтелекту та традиційних алгоритмів для вирішення задачі про багаторукого бандита). Проаналізовано ефективність використання кожного з алгоритмів та доцільність їхнього використання для визначення інвестиційної привабливості. Результати експериментів представлені в чіткому та зрозумілому аналітичному вигляді.

Ключові слова: рекурентна нейронна мережа, багаторукий бандит, прогнозування, ефективність алгоритму.

KHAVALKO VIKTOR, DOKHNIAK BOHDAN-OLEKSANDR, SLAVA LIUBOMYR
Lviv Polytechnic National University

APPLICATION OF NEURAL NETWORK APPROACHES TO SOLVE THE MULTI-ARMED BANDIT PROBLEM

The primary challenge for many individuals is the lack of knowledge on how to take the first step into the realm of investing their finances. People aspiring to delve into investing typically lack guidance on where to begin and which stocks of companies can be lucratively traded. This article conducts an analysis and comparison of eight fundamental algorithms for solving the multi-armed bandit problem. To achieve this, a corresponding research environment was designed and developed, allowing observation of algorithm behavior over a simulated period of seven years. The environment closely resembles real-world conditions, enabling the analysis of agent behavior in the simulation and drawing pertinent conclusions regarding their effectiveness.

A new modification of the greedy agent was created, which, instead of using its own evaluations, utilizes predictions formed by recurrent neural networks. The proposed approach combines the capabilities of artificial intelligence and traditional algorithms to address the multi-armed bandit problem. The effectiveness of each algorithm and the appropriateness of their use in determining investment attractiveness were analyzed. The results of the experiments are presented in a clear and understandable analytical format.

Two best algorithms from each domain were chosen: UCB and the greedy agent, whose evaluations are formed by a recurrent neural network based on GRU. The results of using other algorithms, which do not require prior knowledge of the environment while providing a decent profit, were also analyzed.

The best results were obtained when using UCB and the greedy agent, whose evaluations are formed by a recurrent neural network based on GRU. Although the profit obtained using UCB was three times greater than the profit obtained by the GRU agent, it is worth noting that the probability of the correct selection of the trust parameter in UCB is very low. Therefore, depending on the needs of potential users, one of these approaches can be chosen, keeping in mind the risk of using UCB.

Keywords: recurrent neural network, multi-armed bandit, prediction, algorithm effectiveness.

Постановка проблеми

Багато початківців стикаються з проблемою незнання базових правил вкладення коштів та незнанням мінімального аналізу фінансового ринку. Всі ці фактори зупиняють людину від вкладень, або, втративши перші гроші через брак досвіду людина перестає цікавитись даною сферою заробітку. Отож, головна проблема більшості людей – незнання того, як зробити перший крок в даному вельми прибутковому середовищі. Люди, які хочуть почати займатись інвестуванням зазвичай не знають з чого почати та акції яких компаній можна вигідно перепродати. Щоб уникнути цієї проблеми запропоновано використовувати різні алгоритми для вирішення задачі про багаторукого бандита, такі як: Випадковий алгоритм, Епсилон-жадібний, Оптимістичний жадібний алгоритм, UCB. Результати цих методів будуть порівнюватись з алгоритмами передбачення: RNN з використанням різних типів мереж: звичайна, LSTM, GRU, задля визначення найефективнішого підходу до вирішення проблеми. Окрім того, для підвищення ефективності прогнозування інвестиційної привабливості запропоновано підхід поєднання засобів штучного інтелекту та задачі про багаторукого бандита.

В ролі стохастичного середовища будуть виступати датасети, які містять в собі динаміку змін цін на акції компаній з 2010 по 2017 роки. Таким чином, будемо мати змогу подивитись на роботу алгоритмів в реальному середовищі та оцінити їхню ефективність. Стохастичні середовища дозволяють створювати нескінченну кількість різних умов для аналізу поведінки алгоритмів. Вони широко використовуються для оцінки ефективності різних алгоритмів. Стохастична природа середовища дозволяє отримати загальну картину поведінки алгоритмів. Для нашого дослідження воно повинно складатись з абстрактних автоматів, які при виборі відповідним агентом повинні повертати винагороду та змінювати її при кожній ітерації.

У роботі буде розглянуто задачі: 1) розробки середовища симуляції функціонування біржі та препроцесинг даних для його наповнення; 2) програмна реалізація досліджуваних алгоритмів та проведення їх порівняльного аналізу; 3) обґрунтування вибору найефективнішого алгоритму після аналізу основних їх параметрів.

Аналіз останніх джерел

В [1] дослідженні запропонована напівпараметрична графічна модель, яка вивчає стохастичні функції на основі методом апроксимації вихідних даних функції з обмеженими вибірками даних за допомогою байєсівської оптимізації. Ця стаття надає додаткові варіанти покращення моделі багаторукого бандита, що може доповнити мої дослідження. Автори з [2] розглядали модель «обмеженого багаторукого бандита», яка додає обмеження до доцільності дій. Проаналізувавши це дослідження я зможу проаналізувати його ефективність не тільки в запропонованому автором епсилон-жадібному алгоритмі.

В [3] запропоновано створення механізму передбачення та оцінки акцій. Проаналізувавши це дослідження можна почерпнути для себе досить цікаві результати емпіричного дослідження ринку акцій. Також, наявне порівняння з іншими еталонними моделями. В статті [4] описана реалізація алгоритмів машинного навчання для генерування інвестиційних ознак, враховуючи бразильський сценарій. Було реалізовано три техніки штучного інтелекту, а саме: багатошаровий перцептрон, логістична регресія та дерево рішень, які виконували класифікацію інвестицій. Була досягнута точність 77 %. У роботі [5] показано дослідження різних сценаріїв використання глибокого навчання на фінансових ринках, особливо на фондовому. Більшість досліджень зосереджено на торговій стратегії, прогнозуванні цін і управлінні портфелем, причому обмежена кількість досліджень розглядає симуляцію ринку, вибір акцій, стратегію хеджування та управління ризиками.

Роботу [6] присвячено дослідженню механізму переривання. Існує проблема балансу експлуатації та дослідження. Дана робота доказує, що розроблений механізм переривання покращує винагороду з часом. З мінусів – дослідження цієї модифікації на алгоритмі верхньої межі, що в собі теж має механізм переривання. В [7] досліджували одночасне використання декількох багаторуких бандитів, які можуть спілкуватись одне з одним для мінімізації власного параметру жалю. Також вони самі обирають з ким спілкуватися. Один багаторукий бандит може досить сильно мінімізувати параметр жалю, якщо використовує ефективні алгоритми, але декілька бандитів, які допомагають один-одному та обмінюються інформацією здатні на більше.

Метою роботи є створення середовища, максимально наближеного до реального та порівняння в ньому алгоритмів для розв’язування задачі про багаторукого бандита та алгоритмів передбачення задля визначення найефективнішого підходу.

Постановка задачі

Розглядувану задачу можна поділити на чотири частини. В першій частині реалізовано аналіз даних, який полягає в графічному представленні динаміки зміни цін, аналізу різких змін на ціни, період різких коливань, причини, тощо. Також для аналізу слід знайти середнє, максимальне, мінімальне значення, смугу Боллінджера. Проаналізувавши наявну інформацію зможемо ефективніше оцінювати результати роботи алгоритмів.

Далі в процесі вирішення цієї задачі потрібно створити експериментальне середовище. Для цього необхідно систематизувати та поєднати дані так, щоб забезпечити відповідність між кожною ітерацією та конкретним періодом часу, коли ціни на акції відповідних компаній були актуальні. У цьому контексті, агент буде емулювати дії особи, яка здійснює торгівлю на фінансовому ринку, торгуючи акціями різних компаній. Створене середовище повинне містити можливість інвестування коштів агентом, який представляє відповідну стратегію, в акції десяти різних компаній, як зображено на рис. 1.

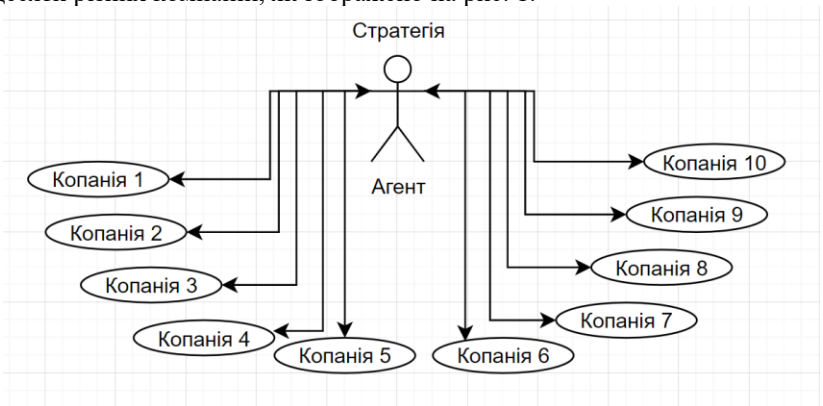

Рис. 1. Агент в середовищі, яке складається з десяти компаній

Як проілюстровано вище, взаємодія в середовищі відбувається в двосторонньому керунку. Після інвестування в одну з компаній, від неї агент отримає винагороду. Оскільки вкладення в компанії може привести до збитків, наш агент може отримати не тільки додатну, а ще й від'ємну суму своєї винагороди.

Наступною частиною задачі буде створення агентів, результати вкладень яких будуть аналізуватись в подальшому. Задля досягнення цієї цілі потрібно, в залежності від обраних алгоритмів, програмно реалізувати відповідні агенти. Також, потрібно створити та натренувати рекурентні нейронні мережі для подальшої модифікації жадібного агенту з їхньою допомогою. Обов'язковою частиною даного етапу є програмна реалізація запису, зберігання та графічного подання результатів роботи кожного з них. Графіки, які ілюструватимуть результати роботи кожного з агентів повинні аналітично зрозуміло представляти динаміку отримуваної винагороди агенту та кількість вкладень у відповідні компанії.

Завершальним етапом задачі буде порівняльний аналіз даних, отриманих в результаті проведення експериментів. Порівнюючи отримані винагороди від вкладень у відповідні компанії, зможемо оцінити ефективність та придатність агентів для практичного використання.

Виклад основного матеріалу

Пропонована система складатиметься з трьох основних компонент: *експериментального середовища*, *агентів*, які реалізовуватимуть відповідні алгоритми та *інструментів для аналізу їхньої ефективності*.

Експериментальне середовище є абстрактною репрезентацією фінансової біржі, воно буде складатись з десяти компаній, матиме такий функціонал:

- При виборі компанії для інвестування, агент отримуватиме відповідну винагороду.
- Автоматичне оновлення винагород від вкладень в компанії при переході на наступну ітерацію.
- Можливість зміни джерела вхідних даних та автоматична генерація винагород при його зміні.
- При кожній ітерації середовище зберігає винагороду, отриману агентом від перепродажу цінних паперів відповідної компанії, а також, оновлювати дані щодо кількості вкладень у обрану компанію для подальшого аналізу поведінки.
- Зміна кількості компаній в середовищі.
- Запуск експерименту необмежену кількість разів задля об'єктивної оцінки стохастичних алгоритмів.

Агент, який імплеметовує конкретну стратегію, матиме такий функціонал:

- Агент реалізовує стратегію відповідного алгоритму, на якому він базується.
- Після ініціалізації, агент входить в експериментальне середовище. При цьому він повинен сформувати початкові оцінки для наявних в середовищі компаній. Початкові оцінки компаній не будуть формуватись при використанні агентом алгоритму на основі рекурентних нейронних мереж.
- Після кожної ітерації агент оновлює оцінки компаній, відповідно до алгоритму, окрім випадків, коли в алгоритмі використовуються оцінки на основі передбачень рекурентних нейронних мереж.
- Після кожної ітерації агент повинен зробити вибір між розвідкою, щоб отримати нові дані про компанії, та експлуатацією знань, які він отримав протягом дослідження середовища. (окрім агентів на основі рекурентної нейронної мережі).

Засоби для обробки результатів включатимуть в себе набір функцій для створення графіків та розрахунку кумулятивної суми. Вони слугуватимуть для аналітично зрозумілого подання даних, які були отримані в процесі дослідження, з метою подальшого аналітичного вивчення. Розрахунок кумулятивної суми потребуватиме винагороду, отриману у відповідний часовий проміжок та, відповідно, цей часовий відрізок. Після розрахунку кумулятивної суми, матимемо можливість спостерігати, як змінювався наш прибуток протягом семи років. Також, задля розуміння поведінки відповідних агентів, засоби обробки результатів вимагатимуть суму кількісті вкладень у відповідні компанії. Після отримання та обробки всіх цих даних, отримаємо графічну репрезентацію результатів експериментів, що дозволить нам проаналізувати та оцінити ефективність кожного з алгоритмів в зручній для аналізу формі.

У результаті аналізу предметної області «Середовище симуляції біржі для дослідження ефективності методів для розв'язання задачі про багаторукого бандита та алгоритмів передбачень» отримано перелік об'єктів: Біржа; Компанія; Акції компанії (акції Amazon, акції GOOG та інші); Експериментальне середовище; Агент (багаторукий агент, агент на основі рекурентної нейронної мережі); Багаторукий агент (агент на основі алгоритмів передбачень, жадібний, епсилон-жадібний, LSTM та ін.); Результати (результати взаємодії агентів з середовищем).

Для цього переліку інформаційних об'єктів було отримано такі реквізити:

- Дані про акції для певного періоду (ціна на початок та кінець дня, потенційна винагорода, попит);
- Характеристики агентів (алгоритм, який закладений в основу, ваги, додаткові параметри, основні параметри: винагорода, жаль, оцінка);
- Характеристики результатів та вимоги їхнього подання (графік, таблиця, числові дані);

Отримані інформаційні об'єкти за характеристиками поділено на узагальнені, конкретні та агрегатні:

- *Узагальнені*: Біржа, Агент – ці об'єкти є узагальненими, оскільки вони представляються цілий клас об'єктів даного проблемного середовища;
- *Конкретні*: акції IBM, акції Amazon, акції LG, жадібний агент, епсилон-жадібний агент, LSTM агент та ін. – ці об'єкти є конкретними в межах класу, оскільки вони конкретизують їх характеристики;

- *Агрегатні*: Експериментальне середовище, багаторукі агенти та агенти на основі рекурентної нейронної мережі, Компанія, експериментальне середовище – ці об'єкти я виділив як агрегатні, так як до їх складу входять інші об'єкти з заданої предметної області.

Система буде генерувати такі вихідні дані:

- Кумулятивну винагороду, яка представляє собою суму винагород, отриманих від відповідного агента під час кожної ітерації експерименту;
- Графік накопиченої винагороди;
- Діаграму, що відображає обрані компанії та винагороди, які їм належать.

Ці вихідні дані будуть використані для проведення аналізу ефективності використання різних алгоритмів у реальних умовах і для подальшого порівняння їх продуктивності.

Перший крок у дослідженні полягає у вивченні вхідних даних. Для цього виконується завантаження інформації в середовище з відповідних файлів та побудова графіків, що відображають зміну цін компаній відповідно до цих даних (рис. 2)

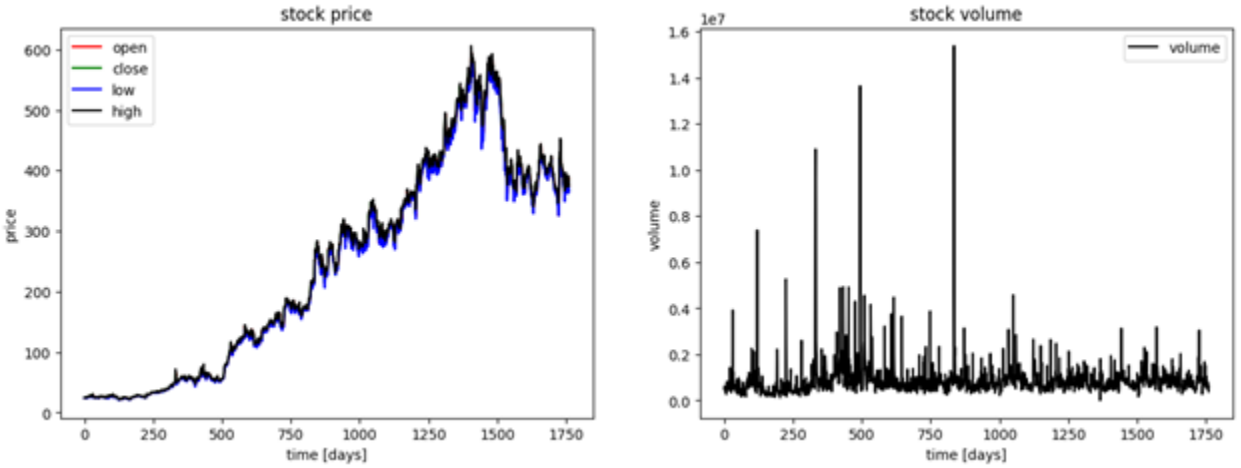

Рис. 2. Ціни та попит на акції компанії REGN (2010-2017 рр.)

Наступним завданням в рамках цього дослідження є опрацювання та підготовка даних для наповнення середовища. Воно включає в себе процес формування винагород, отримуваних агентом впродовж проведення експериментів. Спосіб визначення винагороди буде наступним: для розрахунку винагороди ми візьмемо ціни на акції на початку та в кінці дня та обчислимо різницю між ними. Цей підхід дозволить там обчислити прибуток, або збитки, отримані внаслідок перепродажу цінних паперів відповідної компанії в межах однієї ітерації. В результаті використання цього підходу, агент буде мати змогу вкладати у відповідні компанії протягом проведення експериментів та формувати оцінки. Крім того, цей підхід дозволяє аналізувати чистий прибуток та збитки, адже обчислюючи винагороду саме таким чином маємо змогу об'єктивно оцінити ефективність дій нашого агенту в досягненні основної мети – примноження статків. Після формування винагород, графічно відобразимо зміни цін на акції (рис. 3).

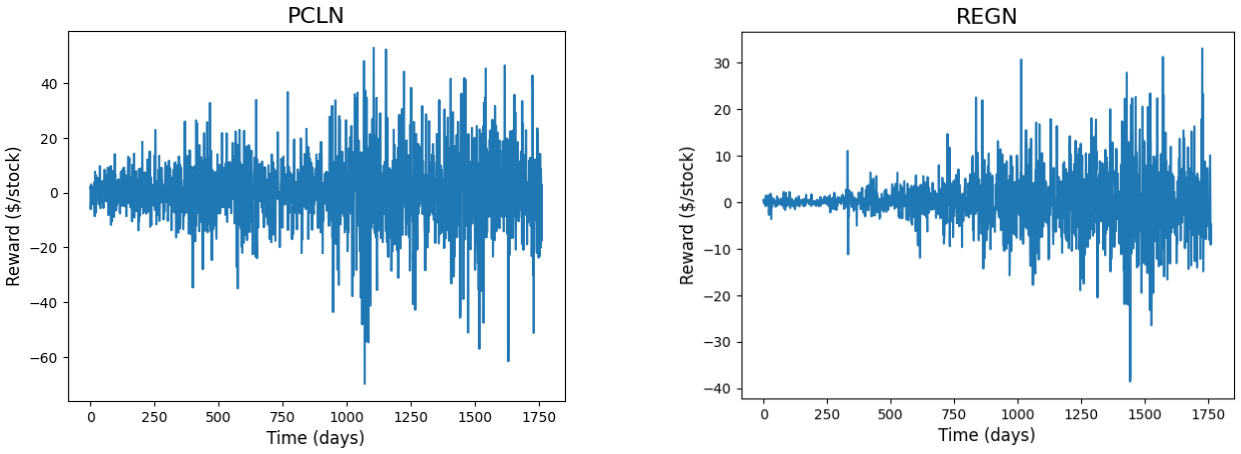

Рис. 3. Зміни цін на акції компаній PCLN та REGN (2010-2017 рр.)

Для оцінювання ефективності роботи кожного з алгоритмів будемо аналізувати їхні основні параметри: кількість неприбуткових вкладень, кількість прибуткових вкладень, максимальну сума винагород, мінімальну сума винагород. Кожен з алгоритмів буде запускатись в середовищі по одній тисячі разів. Така велика кількість запусків дозволить добре оцінити їхню справжню ефективність, мінімізуючи вплив стохастичних факторів на результат роботи кожного з них. Результати роботи найефективніших комбінацій

представлені в таблиці 1 для подальшого аналізу.

Таблиця 1

Результати роботи алгоритмів

Назва алгоритму	К-ть неприбут-ков их вкладень	К-ть прибут-ков их вкладень	Максимальна сума винагород	Мінімальна сума винагород
Випадковий алгоритм	430	570	188.72	-759.53
Жадібний алгоритм	286	714	280.54	-654.36
Епсилон-жадібний алгоритм	162	838	385.88	-557.01
Епсилон-жадібний алгоритм із послабленням	72	928	427.53	-362.36
Оптимістичний алгоритм	216	784	247.68	-284.62
UCB	0	1000	897.702	95.47

В ході експериментів було визначено, що найкращі результати, при використанні алгоритму верхньої межі довіри можна отримати використавши параметр довіри рівний 17,16. Неприбуткові вкладення відсутні, що є вельми хорошим знаком. Також, можемо побачити вражаючий максимальний прибуток. Мінімальний прибуток складає невисоке значення, але невеликий прибуток завжди краще збитків. Отже, при правильно підібраних параметрах даний алгоритм в результаті своєї роботи надає нам достойний прибуток. Аналіз поведінки кожного з алгоритмів наведено на рис.4 та рис.5.

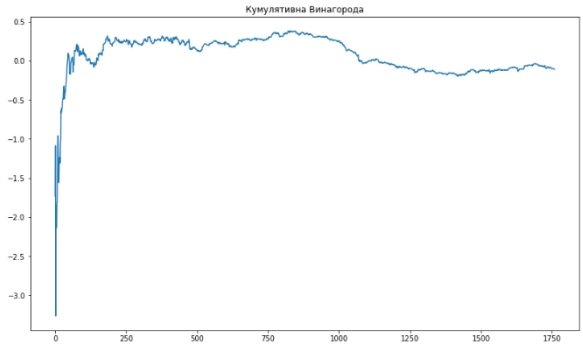

Рис.4. Кумулятивна винагорода агента

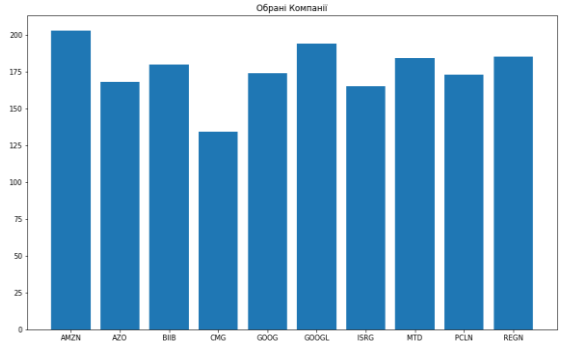

Рис. 5. Частота вибору агентом компаній

Використання рекурентних нейронних мереж для торгівлі акціями на біржі передбачає наступні етапи:

- Створення і тренування нейронних мереж;
- Отримання прогнозів;
- Модифікація жадібного багаторукого бандита за принципом заміни його класичних оцінок на передбачення, отримані з нейронних мереж.

Спочатку потрібно нормалізувати дані для тренування. Для цього використовуємо мінімаксне масштабування, щоб забезпечити однаковий діапазон значень для всіх атрибутів. Потім розбиваємо наші дані на тренувальні (80%), тестові (10%) та валідаційні (10%).

Для оцінювання ефективності роботи кожного з алгоритмів будемо аналізувати їхню кінцеву кумулятивну винагороду. Для того, щоб отримати кінцеву кумулятивну винагороду, потрібно отримати прогнозовані ціни на акції відповідних компаній. Після отримання прогнозів потенційного прибутку з відповідних компаній, жадібний агент обирає потенційно найбільш прибуткове вкладення та вкладає у відповідну компанію. Як результат, він отримує реальну винагороду за перепродаж відповідної компанії. Результати роботи наведені в таблиці 2.

Таблиця 2

Результати роботи рекурентних моделей

Назва моделі	Кумулятивна винагорода
RNN	362.12
LSTM	345.32
GRU	362.16

З наведених результатів видно, що найкраще себе проявили моделі на основі RNN та GRU. Вони хоча й з невеликим відривом відійшли від LSTM (16.8 ум.од.). Такий результат був досить передбачуваний, адже LSTM краща для запам'ятовування довготривалих залежностей, що не є завжди найкращою ідеєю, коли мова йде про інвестування та коливання ціни на акції компаній. Поведінка моделі RNN наведена на рис. 6 та рис.7.

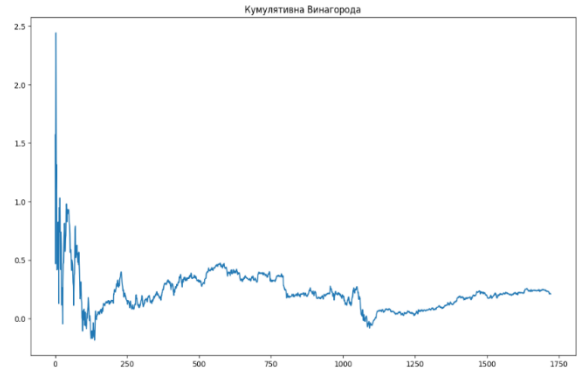

Рис.6. Кумулятивна винагорода моделі на основі RNN

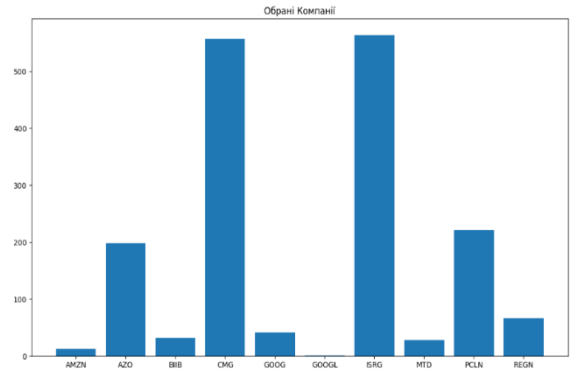

Рис.7. Частота вибору моделі на основі RNN

Висновки

Для реалізації поставленого завдання створено експериментальне середовище, яке дозволило спостерігати та аналізувати поведінку кожного з алгоритмів за однакових умов впродовж семи років. Дані для наповнення середовища були оброблені та проаналізовані. Також було програмно реалізовано кожен алгоритм. Впродовж досліджень основний фокус був спрямований на аналіз восьми базових алгоритмів. Обґрунтовано вибір двох найкращих алгоритмів з кожної предметної області: UCB та жадібного агенту, оцінки якого сформовані рекурентною нейронною мережею на основі GRU. Також, проаналізовані результати використання інших алгоритмів, які не потребують попередніх знань про середовище, водночас надаючи непоганий прибуток.

Найкращі результати були отримані при використанні UCB та жадібного агенту, оцінки якого сформовані рекурентною нейронною мережею на основі GRU. Хоча прибуток, отриманий при використанні UCB й був втричі більшим за прибуток, отриманий GRU агентом, варто зауважити, що ймовірність правильного підбору параметру довіри в UCB є вельми малим. Отже, в залежності від потреб потенційного користувача можна обирати один з цих підходів, не забуваючи про ризик використання UCB. Однак, використання результатів обох підходів є найкращим варіантом для аналізу інвестиційної привабливості, адже поєднання потенціалу великого прибутку з стабільною стратегією надає найкращі оцінки компаній для інвестування.

Література

1. Zhao C., Yang B., and Hirate Y., A reward optimization model for decision-making under budget constraint, Journal of Information Processing, vol. 27, 2019, pp. 190–200, doi: 10.2197/ipsjjip.27.190.
2. Chang H.S., An asymptotically optimal strategy for constrained multi-armed bandit problems, Mathematical Methods of Operations Research, vol. 91, no. 3, 2020, pp. 545–557, doi: 10.1007/s00186-019-00697-3.
3. Jiang Y., Application and Comparison of Multiple Machine Learning Models in Finance, Scientific Programming, vol. 2022, 2022, doi: 10.1155/2022/9613554.
4. Oliveira A.V.D., Dazzi M.C.S., Fernandes A.M.D.R., Dazzi R.L.S., Ferreira P., and Leithardt V.R.Q., Decision Support Using Machine Learning Indication for Financial Investment, Future Internet, vol. 14, no. 11, 2022, doi: 10.3390/fi14110304.
5. Olorunnimbe K. and Viktor H., Deep learning in the stock market—a systematic survey of practice, backtesting, and applications, Artificial Intelligence Review, vol. 56, no. 3, 2023, pp. 2057–2109, doi: 10.1007/s10462-022-10226-0.
6. Cayci S., Eryilmaz A., and Srikant R., Learning to Control Renewal Processes with Bandit Feedback, Performance Evaluation Review, vol. 47, no. 1, 2019, pp. 41–42, doi: 10.1145/3309697.3331515.
7. Sankararaman A., Ganesh A., and Shakkottai S., Social Learning in Multi Agent Multi Armed Bandits, Performance Evaluation Review, vol. 48, no. 1, 2020, pp. 29–30, doi: 10.1145/3393691.3394217.

References

1. Zhao C., Yang B., and Hirate Y., A reward optimization model for decision-making under budget constraint, Journal of Information Processing, vol. 27, 2019, pp. 190–200, doi: 10.2197/ipsjjip.27.190.
2. Chang H.S., An asymptotically optimal strategy for constrained multi-armed bandit problems, Mathematical Methods of Operations Research, vol. 91, no. 3, 2020, pp. 545–557, doi: 10.1007/s00186-019-00697-3.
3. Jiang Y., Application and Comparison of Multiple Machine Learning Models in Finance, Scientific Programming, vol. 2022, 2022, doi: 10.1155/2022/9613554.
4. Oliveira A.V.D., Dazzi M.C.S., Fernandes A.M.D.R., Dazzi R.L.S., Ferreira P., and Leithardt V.R.Q., Decision Support Using Machine Learning Indication for Financial Investment, Future Internet, vol. 14, no. 11, 2022, doi: 10.3390/fi14110304.
5. Olorunnimbe K. and Viktor H., Deep learning in the stock market—a systematic survey of practice, backtesting, and applications, Artificial Intelligence Review, vol. 56, № 3, 2023, pp. 2057–2109, doi: 10.1007/s10462-022-10226-0.
6. Cayci S., Eryilmaz A., and Srikant R., Learning to Control Renewal Processes with Bandit Feedback, Performance Evaluation Review, vol. 47, no. 1, 2019, pp. 41–42, doi: 10.1145/3309697.3331515.

7. Sankararaman A., Ganesh A., and Shakkottai S., Social Learning in Multi Agent Multi Armed Bandits, Performance Evaluation Review, vol. 48, №. 1, 2020, pp. 29–30, doi: 10.1145/3393691.3394217.