огляд літератури

Іван Кущ

February 2025

1 Машине Навчання

Машинне навчання та штучний інтелект – це не просто інструменти, а нова філософія роботи з інформацією. Вони змінюють підхід до наукових досліджень, роблячи їх швидшими, точнішими та масштабнішими. У майбутньому ці технології можуть привести до відкриттів, які ми сьогодні навіть не можемо уявити. Машинне навчання (ML) ділиться на три основні категорії: навчання з учителем (Supervised Learning, SL), навчання без учителя (Unsupervised Learning, UL) і навчання з підкріпленням (Reinforcement Learning, RL).

У методі навчання з учителем(Supervised Learning, SL) , модель навчається на розмічених даних, тобто кожен вхідний об'єкт супроводжується відповідним вихідним значенням. Мета - знайти функцію f, яка апроксимує залежність між входами X і виходами Y:

$$Y = f(X) + \epsilon$$

де ϵ – випадкова помилка.

Навчання без вчителя (Unsupervised Learning) Цей метод використовується, коли дані не мають розмічених міток. Алгоритм шукає приховані структури в даних.Прикладом такого метроду , є метод кластеризації-об'єднує схожі об'єкти.

В навчання з підкріпленням (Reinforcement Learning) агент взаємодіє зі середовищем та отримує винагороду за дії. Мета - максимізувати функцію винагороди:

$$Q(s, a) = r + \gamma \max_{a'} Q(s', a')$$

де $Q(s, a)$ - цінність дії a в стані s, r - винагорода, γ - коефіцієнт дисконтування.

1.1 Основні Властивості Випадкового лісу (Random Forest, RF)

У сучасному світі величезний потік інформації вимагає ефективних методів її аналізу. Одним із найбільш популярних і потужних алгоритмів машинного навчання є випадковий ліс (Random Forest, RF). Він поєднує простоту дерева рішень з ансамблевим підходом, що дозволяє підвищити точність і зменшити ризик переобучення. Випадковий ліс є ансамблевим методом, який базується на методі бутстреп-агрегування (bagging). Основна ідея полягає в тому, щоб створити кілька незалежних дерев рішень і об'єднати їхні результати.У кожному вузлі дерева випадковим чином вибирається тільки частина ознак, що зменшує кореляцію між деревами.Дерево росте до максимальної глибини або досягнення порогу розбиття.Для класифікації використовується голосування більшості: клас, що зустрічається найчастіше серед дерев, обирається як остаточний. Нехай маємо датасет:

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

де x_i — вектор ознак, а y_i — мітка класу (у класифікації) або числове значення (у регресії).

Остаточний клас $\hat{y}$ визначається більшістю голосів:

$$\hat{y} = \arg\max_c \sum_{i=1}^{m} I(T_i(x) = c)$$

де:

- m — кількість дерев у лісі, - $T_i(x)$ — прогноз i-го дерева, - c — один із класів, - $I(\cdot)$ — індикаторна функція, що дорівнює 1, якщо прогноз дерева дорівнює класу c, і 0 в іншому випадку.

Ансамблевий підхід дозволяє випадковому лісу перевершувати класичні моделі, такі як лінійна регресія або поодинокі дерева рішень, особливо у випадку складних нелінійних залежностей.

Для кожного дерева використовується випадкова підмножина ознак, що зменшує кореляцію між деревами та підвищує стійкість алгоритму.

Формально, помилка узгодженого ансамблю (ensemble error) визначається як:

$$E = \rho E_T + \frac{1 - \rho}{m} E_I$$

де:

- E — загальна помилка ансамблю,
- ρ — середня кореляція між деревами,
- E_T — помилка окремого дерева,
- E_I — незалежна помилка.

Якщо $m \to \infty$ і ρ мале, тоді $E \to E_I$, тобто помилка зменшується!

Оскільки алгоритм використовує бутстреп-вибірки, окремі аномальні точки не впливають на всі дерева, що робить модель стійкою до викидів (outliers).

Формально, випадковий ліс може зменшувати вплив викидів на оцінку середнього значення в задачах регресії:

$$\hat{y}_{RF} = \frac{1}{m} \sum_{i=1}^{m} T_i(x) \approx \text{median}(T_i(x))$$

Медіана є більш стійкою до викидів, ніж середнє значення!

Випадковий ліс дозволяє визначити, які ознаки мають найбільший вплив на прогноз.

Метрика важливості ознаки (Gini Importance) обчислюється як:

$$I(X_j) = \sum_{t \in T} p(t)(1 - p(t)) - \sum_{k \in \{left, right\}} p(t_k)(1 - p(t_k))$$

де:

- $I(X_j)$ — важливість ознаки X_j,
- $p(t)$ — частка вибірки, яка потрапляє у вузол t,

- $p(t_k)$ — частки після розбиття вузла на дочірні вузли.

Це допомагає проводити відбір ознак, спрощуючи модель!

Головний недолік випадкового лісу — це час навчання та використання, оскільки він створює багато дерев і робить численні обчислення.

Складність алгоритму:

$$O(m \cdot n \log n)$$

де:

- m — кількість дерев,
- n — кількість зразків у навчальній вибірці.

Чим більше дерев, тим вища точність, але тим довше час навчання!

На відміну від окремого дерева рішень, де логіка ухвалення рішення є прозорою, випадковий ліс працює як чорний ящик.

Він не надає чітких правил ухвалення рішень, а лише усереднює прогнози дерев.

Приклад: В індивідуальному дереві ми можемо чітко сказати, що "Якщо дохід $>$ 50 тис., то людина бере кредит але у випадковому лісі кожне дерево може мати інший критерій.

Це ускладнює використання у чутливих сферах, таких як медицина чи право.

Якість випадкового лісу залежить від:

- Кількості дерев m (замало — низька точність, забагато — довге навчання),
- Максимальної глибини дерев (надмірно глибокі дерева можуть переобучатися),
- Кількості ознак для поділу (якщо вибирати забагато, дерева стають подібними).

Оптимальне налаштування:

$$m = 100 - 500 \quad \text{дерев}, \quad k = \frac{p}{\sqrt{m}}, \quad \text{де } p^{-}\text{загальна кількість ознак.}$$

1.2 Основні Властивості Символічної регресії (Symbolic Regression, SR)

Випадковий ліс, демонструючи вражаючу ефективність у задачах прогнозування, страждає від проблеми "чорної скриньки". Хоча модель здатна генерувати якісні результати, її внутрішня структура залишається непрозорою, що ускладнює інтерпретацію отриманих висновків та розуміння факторів, які впливають на результат.Для більш якіснох інтерпретації наших даних та їх передбаченню був сворений метод - Символічна регресія(Symbolic Regression, SR). Символічна регресія (Symbolic Regression, SR) — це потужний метод машинного навчання, який дозволяє автоматично знаходити математичні функції, що найкраще описують дані, без необхідності заздалегідь визначати функціональну форму. Цей метод є особливо корисним у випадках, коли немає можливості вивести функцію з перших принципів, або коли форма розподілу даних є довільною та складною для опису традиційними методами. Однією з ключових властивостей символічної регресії є її здатність автоматично шукати функції, які найкраще описують дані. На відміну від традиційних методів регресії, таких як поліноміальна регресія, які вимагають заздалегідь визначеної функціональної форми, SR дозволяє моделі самостійно знаходити оптимальну функцію. Це досягається за рахунок використання генетичного програмування, де функції представлені у вигляді дерев виразів, які еволюціонують протягом процесу навчання.

У генетичному програмуванні функції генеруються шляхом мутацій (зміни вузлів дерева) та кросоверу (обмін піддеревами між різними кандидатами). Цей процес дозволяє SR шукати функції в широкому просторі можливих математичних виразів, не обмежуючись попередньо заданими формами. Таким чином, SR усуває необхідність емпіричного підбору функцій, що є трудомістким і часто неточним процесом.

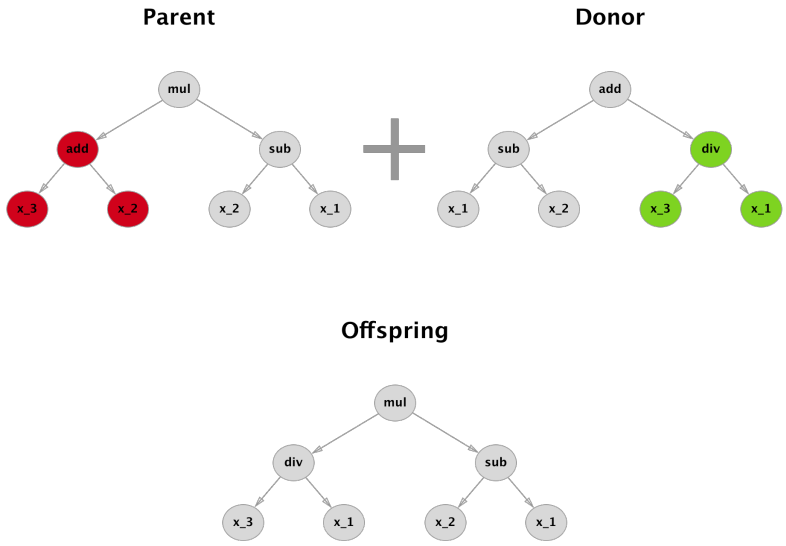

Рис. 1: Підхід генетичного програмування до символічної регресії. Функції представлені деревами виразів. Нові функції генеруються за допомогою мутації деревних вузлів.

SR дозволяє знаходити функції, які точно описують такі розподіли, навіть якщо вони не мають явного аналітичного виразу. Це досягається за рахунок того, що SR не обмежується попередньо заданими функціональними формами, а шукає їх у широкому просторі можливих математичних виразів. Сума квадратів похибок (SSE) є широко використовуваною метрикою оцінки якості моделі в алгоритмах символічної регресії. Вона базується на обчисленні різниці між прогнозованими та фактичними значеннями цільової змінної та визначається наступним чином

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{1}$$

де y_i - фактичне значення, $\hat{y}_i$ - прогнозоване значення, n - кількість спостережень.

Середньоквадратична похибка (MSE) є однією з найбільш поширених метрик для оцінки якості регресійних моделей. Вона відображає середнє значення квадрату різниці між прогнозованими та фактичними значеннями цільової змінної. MSE чутлива до викидів, оскільки вони зводяться в квадрат, що може значно збільшити значення метрики.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

де:

- n - кількість спостережень;

- y_i - фактичне значення i-го спостереження;
- $\hat{y}_i$ - прогнозоване значення i-го спостереження.

В процесі символічної регресії (SR) початкова популяція розв'язків формується випадковим чином з нескінченного пулу операторів та терміналів. Розмір популяції задається користувачем. Кожен окремий розв'язок може характеризуватися різною мірою придатності, впливаючи на загальну ефективність. Алгоритм SR постійно оцінює якість кожного розв'язку, відбираючи для подальшої еволюції найбільш перспективні варіанти, що демонструють мінімальну похибку, та відкидаючи неефективні.В алгоритмі символічної регресії (SR) еволюційний процес характеризується ітеративним зменшенням середньої помилки за рахунок елімінації розв'язків з низькою ефективністю. Завершення алгоритму SR відбувається при виконанні однієї з наступних умов: досягнення максимальної кількості ітерацій, заданої користувачем; досягнення оптимального розв'язку з нульовою помилкою; досягнення прийнятного рівня помилки, встановленого користувачем (наприклад, 0,01). Зазвичай використовується комбінація цих критеріїв. Важливо зазначити, що SR може генерувати не тільки окреме рівняння, але й набір рішень, що відрізняються за складністю. Символічна регресія є ефективним методом, який дозволяє значно скоротити час і зусилля, необхідні для моделювання даних. Завдяки автоматизації процесу пошуку функцій, SR усуває необхідність ручного підбору функціональних форм, що є трудомістким і часто неточним процесом. Це особливо важливо в таких галузях, як фізика високих енергій, де дані можуть мати складні форми, і традиційні методи моделювання можуть бути недостатньо ефективними.

2 Застосування символічної регресії у різних фізичних задачах

У цьому розділі більш детально розглянеться приклади використання SR та порівняння результатів з іншими методами ML та AI .

2.1 Computation of flow rates

Основна мета роботи полягала у зменшенні обчислювальних витрат та покращенні доступності розрахунків для мікро- та вакуумних систем, де традиційні методи кінетичної теорії є трудомісткими та непрактичними для оптимізації.У дослідженні використовувалися два підходи машинного навчання:Випадковий лісовий регресор (Random Forest Regression, RFR) , Символічна регресія (Symbolic Regression, SR).RFR показав високу точність прогнозування швидкості потоку. Максимальна абсолютна відносна похибка між прогнозами RFR та даними кінетичного моделювання становила менше 12.5%. Метод демонстрував швидку обчислювальну ефективність, що робить його корисним для уникнення трудомістких симуляцій. Однак RFR не надає аналітичних виразів, що обмежує його застосування в інженерних розрахунках. За допомогою методу символічної регресії (SR) було отримано аналітичні вирази для швидкості потоку W у вигляді функції відношення тисків p, відношення довжини трубки до радіусу l та параметра розрідження газу δ:

$$W = W(p, l, \delta)$$

Для всього діапазону параметрів було запропоновано загальне рівняння з максимальною абсолютною відносною похибкою менше 17%.

З метою підвищення точності, дані були розділені на три підмножини за значеннями параметра розрідження δ. Для кожної підмножини було отримано окреме рівняння з максимальною похибкою менше 9%.

Результати дослідження демонструють здатність SR генерувати прості та точні аналітичні вирази, які можуть бути легко інтегровані в інженерні розрахунки. Це робить метод символічної регресії особливо корисним для практичних застосувань. Символічна регресія (SR) виявилася ефективним методом для отримання аналітичних виразів, які можуть бути використані в інженерних розрахунках для визначення швидкості потоку розрідженого газу. Незважаючи на те, що RFR показав трохи кращу точність, SR має значну перевагу у вигляді інтерпретованих формул, що робить її більш придатною для практичного застосування. Обидва методи можуть бути корисними для подальших досліджень у галузі динаміки розріджених газів, але SR є більш перспективною для отримання фізично інтерпретованих результатів.

2.2 Analytical formulae for design of one-dimensional sonic crystals with smooth geometry based on symbolic regression

У дослідженні, представленому в статті, автори вивчали одномірні сонічні кристали, які є штучними періодичними структурами, здатними модифікувати передачу акустичних хвиль. Основна мета роботи полягала у знаходженні аналітичних формул, які б описували зв'язок між геометрією хвилеводу та акустичною передачею, зокрема шириною та центром заборонених зон (bandgaps). Для цього використовувалися методи машинного навчання (ML) та штучного інтелекту (AI), зокрема символічна регресія.Використовувалася бібліотека PySR, яка дозволяє будувати дерева виразів з аналітичних термів та операторів. Цей підхід дозволив знайти прості та інтерпретовані формули для центрів та ширини заборонених зон, а також для повного дисперсійного співвідношення. За допомогою символічної регресії були отримані аналітичні формули для центрів (m_1, m_2, m_3) та ширини (w_1, w_2, w_3) перших трьох заборонених зон. Наприклад, формула для центру першої забороненої зони має вигляд:

$$m_1 = \pi + 0.38 \max(0.17, x_{drd})$$

Результати, отримані за допомогою символічної регресії, були порівняні з чисельними розрахунками, виконаними за допомогою методу скінченних елементів у програмі COMSOL Multiphysics. Виявилося, що формули, знайдені за допомогою ML, досить точно відтворюють результати чисельних розрахунків, що підтверджує їхню практичну придатність.Символічна регресія дозволила знайти прості аналітичні формули, які легко інтерпретувати та використовувати для швидкого проектування хвилеводу. Це значно ефективніше, ніж проводити складні чисельні розрахунки для кожного нового варіанту геометрії.

2.3 Black HOLE

У дослідженні, основним фокусом було відкриття нової масштабної залежності між масою надмасивної чорної діри ($M_\bullet$) та параметрами спіральних галактик, зокрема кутом нахилу спіральних рукавів (ϕ) та максимальною швидкістю обертання галактичного диска ($v_{\max}$). Ця залежність, названа *тривимірною фундаментальною площиною* $M_\bullet$-ϕ-$v_{\max}$, була отримана за допомогою сучасних методів машинного навчання (ML) та штучного інтелекту (AI), зокрема *символічної регресії*.

Дослідження базувалося на вибірці з 41 спіральної галактики, для яких були доступні вимірювання маси чорної діри ($M_\bullet$), кута нахилу спіральних рукавів (ϕ) та макси-

мальної швидкості обертання ($v_{\max}$). Всі маси чорних дір були виміряні динамічними методами, що забезпечує високу точність.

Для знаходження оптимальної математичної залежності між $M_\bullet$, ϕ та $v_{\max}$ була використана *символічна регресія* (Symbolic Regression). У цьому випадку використовувався програмний пакет PySR.

Після того, як символічна регресія знайшла попередню залежність, було використано метод *Hyper-Fit* для уточнення параметрів моделі.

Підсумкова залежність має вигляд:

$$M_\bullet \sim \mathcal{N}\left[\mu = \alpha(\tan|\phi| - 0.24) + \beta \log\left(\frac{v_{\max}^2}{11\,\mathrm{km\,s^{-1}}}\right) + \gamma, \sigma = 0.22 \pm 0.06\right], \qquad (3)$$

де $M_\bullet = \log(M_\bullet/M_\odot)$ — логарифм маси чорної діри в сонячних масах, α, β та γ — коефіцієнти, а σ — внутрішній розкид моделі.