ЗАСТОСУВАННЯ СИМВОЛЬНОЇ РЕГРЕСІЇ ДЛЯ АНАЛІЗУ ДАНИХ У ФІЗИЦІ ВИСОКИХ ЕНЕРГІЙ

Д. М. Клекоць[1], О. А. Безшийко[1], Л. О. Голінка-Безшийко[1]

[1]Київський Національний Університет Імені Тараса Шевченка, Київ, Україна

Дослідження в області експериментальної фізики високих енергій завжди вимагали проведення значного аналізу отриманих експериментальних даних. Починаючи з аналізу фотографій треків нових частинок з бульбашкових камер до теперішніх експериментів на колайдерах з частотою зіткнення пучків порядку десятків наносекунд. Генерування великої кількості даних у експериментах вимагає алгоритмів їх надійного аналізу, зокрема відділення сигнальних подій від комбінаторного фону, що є особливо важливим для рідкісних події, таких як наприклад утворення бозону Хігса [1-2], що вимагають максимальної ефективності відбору сигналу та придушення фону.

Найпоширеніші на сьогодні методи аналізу даних використовують машинне навчання, що також було підтверджено врученням нобелівської премії з фізики 2024 [3-4] за піонерські дослідження в даному напрямку.

Історично застосування машинного навчання для аналізу даних у фізиці високих енергії розпочалося з застосування нейронних мереж [5], однак згодом посиленні дерева рішень [6] стали більш популярними та стали переважаючим інструментом аналізу на сьогодні. Незважаючи на велику ефективність класичних моделей машинного навчання, вони мають обмежену інтерпритовність, та працюють за принципом «темної скрині». Хоча моделі машинного навчання проходять валідацію, їх розділення подій на сигнальні та фонові, залишається дещо поза аналізом науковців, що потенційно може призвести до упередженості моделі.

Одним з можливих альтернатив є символьні моделі, основним результатом тренування яких є формула скалярної функції декількох аргументів, на основі значень якої приймається рішення про те чи належить подія до фону чи сигналу. Принцип роботи символьних моделей такий же як і класичних посилених дерев рішень чи нейронних мереж. В якості вхідних параметрів вони отримують параметри реконструйованих треків у колайдерах, та їх основі розраховують вихідну величину, яка використовується як достовірність того що подія є сигналом. Особливістю символьних методів є те, що вихідна модель є формулою, яка може бути проаналізована аналітично.

Принцип роботи посилених дерев рішень, можна порівняти з розбиванням фазового простору вхідних параметрів на області сигнальних та фонових подій, в той час як символьні методи спрямовані на пошук закономірностей за якими розподілені сигнальні та фонові події.

Розпізнавання закономірностей символьними моделями може бути продемонстроване на прикладі даних зображених на Рис. 1. Де кожна подія характеризується двома параметрами, які позначенні координатами відповідних точок на графіку. Приведенні дані були використані для навчання моделі посилених дерев рішень, з використанням фреймворку XGBoost [7], та символьної моделі з використанням фреймворку Symbolic Regression [8]. З Рис. 1. візуально видно закономірність за якою розташовані сигнальні події та фонові, однак слід зазначити що символьні моделі можуть розпізнавати закономірності які не завжди вдається встановити візуально. Фоновим кольором на Рис. 1 позначено вихідні значення натренованої символьної моделі. Як видно з графіка, передбачення натренованої символьної моделі відповідає закономірності за якою розподілений сигнальні події.

Формула даної символьної моделі має наступний аналітичний вигляд,

$$F(x_1, x_2) = \left|1 - 0.84\frac{x_1}{x_2}\right|$$

Ефективність символьної моделі, та моделі посилених дерев рішення порівнювалася за допомогою величини площі під ROC кривою (див Рис. 2), яка становить 89,54 для символьної моделі, та 81,91 для моделі посилених дерев рішень. Що свідчить про кращу ефективність символьної моделі.

Варто відмітити, що хоча символьна регресія в даному випадку показує кращі характеристики порівняно з просиленими деревами рішень, це не гарантовано для будь якого набору даних. Також тренування символьних моделей зазвичай займає значно більше часу, порівняно з оптимізованими алгоритмами тренування посилених дерев рішень.

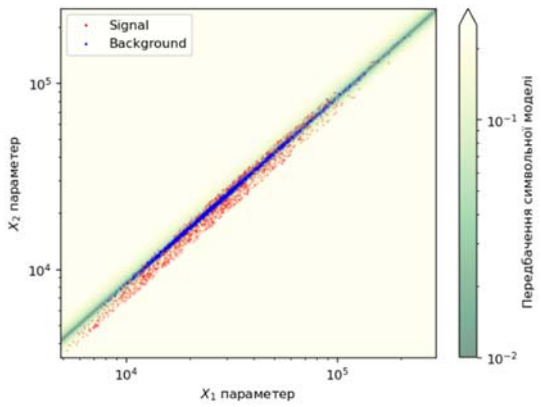

Рис. 1: Візуалізація розподілу параметрів тренувального набору даних. Сигнальні події позначені червоним кольором, фонові події - синім. Фоновим кольором позначено передбачення символьної моделі.

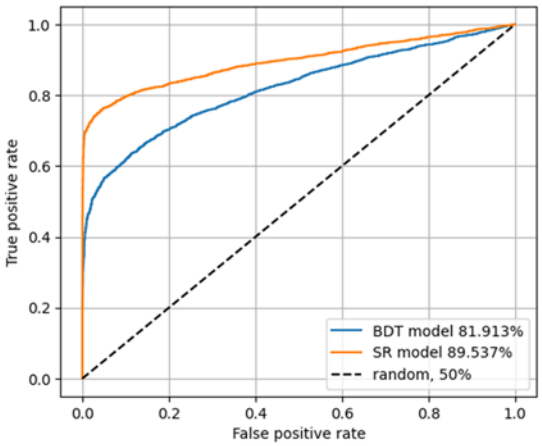

Рис. 1: ROC криві символьної моделі (помаранчева крива) та моделі посилених дерев рішень (блакитна крива), побудовані на валідаційному наборі даних. Площа під кривою зазначена у відсотках від максимальної.

Підсумовуючи, варто зазначити що символьні моделі машинного навчання мають перспективу застосування у фізиці високих енергії, де прозорість моделі має одне з пріоритетних значень. Також символьні моделі демонструють можливість розпізнавання закономірностей у даних, що класичні моделі машинного навчання не завжди можуть розпізнати. Проте символьні моделі не гарантують кращу ефективність порівняно з посиленими деревами рішення та вимагають значно більше обчислювальних ресурсів на етапі тренування, однак через їх відносну простоту, натреновані символьні моделі можуть швидше робити передбачення що має потенціал у застосуванні де швидкість має вирішальне значення, наприклад у тригерних системах колайдерних експериментів.

Дослідження частково підтримано фінансуванням НФДУ в рамках проєкту «Вирішення сучасних проблем хімії, біомедицини, фізики та матеріалознавства з використанням центру високопродуктивних обчислень і машинного навчання», реєстраційний номер заявки 2023.05/0024, (Конкурс «Дослідницькі інфраструктури для проведення передових наукових досліджень»)

1. ATLAS Collaboration, et al., Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC, Physics Letters B, 2012, (716) 1, pp. 1-29. https://doi.org/10.1016/j.physletb.2012.08.020
2. CMS Collaboration, et al., Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC, Physics Letters B, 2012, (716) 1, pp. 30-61. https://doi.org/10.1016/j.physletb.2012.08.021
3. Hopfield J. J, Neural networks and physical systems with emergent collective computational abilities, Proceedings of the National Academy of Sciences. 1982. (79) 8. pp. 2554–2558.
4. Ackley D. H., Hinton G. E., Sejnowski T. J., A learning algorithm for Boltzmann machines, Cognitive Science. 1985. (9) 1. pp. 147–169.
5. L. Teodorescu, Artificial neural networks in high-energy physics, Inverted CERN School of Computing, 2008, 2005 and 2006 edition, pp.13-22.
6. Byron P. Roe, Hai-Jun Yang, Ji Zhu, et al., Boosted decision trees as an alternative to artificial neural networks for particle identification, NIM-A, 2005, (543) 2-3, pp. 577-584. https://doi.org/10.1016/j.nima.2004.12.018
7. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
8. M. Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, DOI: https://doi.org/10.48550/arXiv.2305.01582