ОПТИМІЗАЦІЯ КОНФІГУРУВАННЯ РОЗПОДІЛЕНИХ СИСТЕМ МАШИННОГО НАВЧАННЯ З УРАХУВАННЯМ ВИТРАТ РЕСУРСІВ

Колтун Ю.М., Нго За Фат
Харківський національний університет радіоелектроніки, Харків, Україна

У сучасних інфраструктурах машинного навчання (ML) дедалі ширше використовуються розподілені обчислювальні середовища, зокрема Apache Spark, Dask та Kubernetes-орієнтовані фреймворки. Продуктивність фреймворків критично залежить від конфігураційних параметрів: як гіперпараметрів моделей, так і параметрів обчислювального середовища. Неефективне налаштування може призвести до надмірних витрат ресурсів без приросту якості моделі [1].

Більшість сучасних систем машинного навчання працюють у розподіленому або хмарному середовищі, де витрати на обчислювальні ресурси є критичним фактором. Існуючі підходи до конфігурування таких систем зазвичай фокусуються або на налаштуванні гіперпараметрів моделей, або на параметрах обчислювальної платформи, не враховуючи їхню взаємодію.

Крім того, вартість запуску рідко розглядається як повноцінний критерій оптимізації. Це створює потребу в комплексному підході, що одночасно оптимізує якість моделей, продуктивність і вартість обчислень.

Метою доповіді є розробка та експериментальна перевірка підходу до багаторівневого конфігурування розподілених систем машинного навчання, що дозволяє одночасно оптимізувати точність моделей, час навчання та вартість обчислювальних ресурсів.

Запропоноване рішення полягає в розробці програмного інструменту, що виконує автоматизоване багаторівневе конфігурування розподіленої системи машинного навчання, орієнтоване на одночасну оптимізацію якості моделі, часу її навчання та вартості обчислювальних ресурсів.

Інструмент реалізовано на мові Python із використанням Apache Spark як базової обчислювальної платформи. Він дозволяє створювати та запускати експерименти з різними конфігураціями – як на рівні гіперпараметрів алгоритмів (наприклад, розмір батчу, кількість ітерацій, глибина дерева), так і на рівні системних параметрів Spark (кількість виконавців, обсяг пам’яті, число ядер тощо).

У процесі кожного запуску автоматично вимірюються основні метрики: точність моделі, час навчання та обчислювана вартість (розрахована як добуток тривалості використання ресурсів на їх вартість за одиницю часу).

Теоретично задача налаштування параметрів системи формулюється як мультиоб’єктивна оптимізаційна задача.

Нехай $h \in H$ – гіперпараметри моделі, $s \in S$ – параметри обчислювального середовища, тоді оптимізація формулюється як пошук конфігурацій, що мінімізують векторну цільову функцію.

Для цього використовується метод генерації конфігурацій із подальшим відбором оптимальних комбінацій за принципом Pareto. Кожна конфігурація оцінюється за трьома критеріями, що дозволяє виявити набір рішень, які представляють найкращі компроміси між точністю, продуктивністю та вартістю.

Таким чином, користувач або система може обрати найкращу конфігурацію відповідно до поточних обмежень або пріоритетів. Такий підхід забезпечує формальну, масштабовану й практичну основу для конфігурації складних DML-систем, особливо в умовах хмарних або ресурсно обмежених середовищ, де кожна неефективна конфігурація прямо впливає на вартість розгортання та навчання моделі.

У результаті експериментів було встановлено, що запропонований підхід до спільного конфігурування гіперпараметрів та параметрів розподіленого середовища дозволяє досягти кращого балансу між точністю моделей, часом навчання та вартістю обчислень, порівняно з однорівневими стратегіями. Зокрема, для моделі Gradient Boosted Trees точність зросла на 1-2%, час виконання зменшився на 20–30%, а вартість ресурсів – до 40% у порівнянні з початковою конфігурацією.

Аналіз Pareto-фронту показав, що спільне налаштування значно збільшує кількість ефективних рішень у просторі параметрів.

Також було виявлено, що деякі системні параметри (наприклад, рівень паралелізму) суттєво впливають не лише на швидкодію, а й на якість моделі, що рідко враховується в традиційних AutoML-інструментах.

Отримані результати підтверджують доцільність багаторівневого конфігурування з урахуванням вартості як окремого оптимізаційного критерію.

Список літератури

1. Ляшенко, О. і Михайліченко, І. (2025) «Модель автономної системи моніторингу та оптимізації IT-інфраструктури з використанням трансформерів», Сучасний стан наукових досліджень та технологій в промисловості, (1(31), с. 73–82. doi: 10.30837/2522-9818.2025.1.073.

2. Коваленко А. А., Кучук Г. А. Методи синтезу інформаційної та технічної структур системи управління об’єктом критичного застосування. *Сучасні інформаційні системи*. 2018. Т. 2, № 1. С. 22–27. DOI: https://doi.org/10.20998/2522-9052.2018.1.04

3. Yaloveha V., Hlavcheva D., Podorozhniak A. Usage of convolutional neural network for multispectral image processing applied to the problem of detecting fire hazardous forest areas. *Сучасні інформаційні системи*. 2019. Т. 3, № 1. С. 116–120. DOI: https://doi.org/10.20998/2522-9052.2019.1.19.