УДК 004.85:004.62.031.4:519.85

К. Д. Кириченко

МАСШТАБУВАННЯ АЛГОРИТМІВ МАШИННОГО НАВЧАННЯ В СЕРЕДОВИЩІ BIG DATA

Вінницький національний технічний університет

Анотація

Проаналізовано підходи до масштабування алгоритмів машинного навчання в середовищі Big Data, зокрема парадигму паралелізму даних і механізми розподіленого навчання. Розглянуто проблему синхронізації параметрів у процесі оптимізації на основі Stochastic Gradient Descent. Запропоновано метод адаптивної гібридної синхронізації, що передбачає динамічне регулювання обміну даними між вузлами кластера з урахуванням характеристик системи. Наведено результати експериментальної перевірки ефективності підходу.

Ключові слова: Big Data, масштабування, машинне навчання, розподілені обчислення, адаптивна синхронізація, Stochastic Gradient Descent.

SCALING OF MACHINE LEARNING ALGORITHMS IN BIG DATA ENVIRONMENTS

Abstract

This paper explores approaches to scaling machine learning algorithms within Big Data environments, specifically focusing on the data parallelism paradigm and distributed learning mechanisms. The study addresses the challenge of parameter synchronization during optimization based on Stochastic Gradient Descent (SGD). An adaptive hybrid synchronization method is proposed, featuring dynamic regulation of data exchange between cluster nodes contingent upon system performance characteristics. Experimental results are provided to validate the effectiveness and computational efficiency of the proposed approach.

Keywords: Big Data, scalability, machine learning, distributed computing, adaptive synchronization, Stochastic Gradient Descent.

Сучасний етап розвитку інформаційних технологій характеризується експоненціальним зростанням обсягів даних, що генеруються сенсорними мережами, соціальними медіа та транзакційними системами. У цих умовах класичні алгоритми машинного навчання (ML), орієнтовані на функціонування в межах локальної оперативної пам’яті одного обчислювального вузла, демонструють істотне зниження ефективності або стають непридатними для обробки даних терабайтного масштабу. Відтак науковий дискурс зміщується від екстенсивного нарощування обчислювальних ресурсів до оптимізації комунікаційних витрат, підвищення ефективності розподілених обчислень і вдосконалення механізмів синхронізації в кластерних середовищах [1–3].

Метою роботи є дослідження механізмів масштабування алгоритмів машинного навчання в середовищі Big Data та розробка методу адаптивної синхронізації для підвищення ефективності розподіленого навчання.

На відміну від існуючих підходів асинхронного навчання, запропонований метод передбачає динамічну адаптацію частоти синхронізації на основі комплексної оцінки стану обчислювального середовища, що дозволяє зменшити вплив гетерогенності ресурсів і нестабільності мережевих характеристик.

Результати дослідження

Ефективність масштабування в системах обробки великих даних визначається обраною парадигмою розподілу навантаження. Застосування концепції паралелізму даних (Data Parallelism) передбачає реплікацію моделі на множині обчислювальних вузлів (workers) із розподілом навчальної вибірки на окремі сегменти (shards). Такий підхід забезпечує підвищення продуктивності за рахунок паралельної обробки даних і є базовим у сучасних системах Big Data-аналітики [1; 2].

Основним об’єктом дослідження є процес розподіленого стохастичного градієнтного спуску (SGD), що широко застосовується для оптимізації параметрів моделей машинного навчання. У розподіленому середовищі процедура оновлення параметрів моделі формалізується як:

$$\theta_{t+1} = \theta_t - \eta \times \frac{1}{N} \sum_{i=1}^{N} \nabla L(x_i, y_i, \theta_t)$$

де N – кількість елементів навчальної вибірки;
η – коефіцієнт швидкості навчання;
∇L – градієнт функції втрат.

Установлено, що одним із основних обмежень масштабування є зростання накладних витрат на синхронізацію градієнтів між вузлами. Традиційна модель Bulk Synchronous Parallel (BSP) забезпечує узгодженість параметрів, однак спричиняє затримки, пов'язані з очікуванням повільніших обчислювальних вузлів (stragglers).

Запропонований метод гібридної асинхронної синхронізації передбачає адаптивне регулювання частоти обміну градієнтами залежно від:

- латентності мережі (вимірюється як середній час передачі повідомлень між вузлами);
- обчислювальної продуктивності вузлів (кількість операцій за одиницю часу);
- інтенсивності надходження даних.

Означене, дозволяє зменшити вплив «вузьких місць» і підвищити загальну ефективність системи без втрати збіжності алгоритму.

Експериментальна верифікація проведена в середовищі Apache Spark на кластері з 4 та 16 вузлів (CPU-based), із використанням набору структурованих даних обсягом 500 ГБ (задача класифікації, 120 ознак). Порівняння виконувалося для алгоритмів Random Forest та XGBoost.

Результати показали:

- скорочення часу навчання з 340 хв (локальна система) до 28 хв (16 вузлів);
- коефіцієнт прискорення – 12,1;
- підвищення пропускної здатності мережі (throughput) на 15–20% порівняно з BSP-підходом.

Водночас встановлено зниження ефективності масштабування при збільшенні кількості вузлів (з 92,5% до 75,6%), що обумовлено зростанням витрат на комунікацію та серіалізацію даних. Проте запропонований підхід демонструє вищу стійкість до цих факторів порівняно зі стандартними механізмами Spark MLlib.

Архітектура системи включає три рівні:

1) рівень зберігання даних (HDFS, S3);
2) рівень обробки даних (RDD, DataFrames);
3) рівень обчислювальної логіки, що реалізує адаптивний механізм синхронізації.

Отримані результати узгоджуються з сучасними підходами до використання Big Data як основи для побудови інтелектуальних систем підтримки прийняття рішень [3].

Висновки

У роботі доведено, що ефективне масштабування алгоритмів машинного навчання в середовищі Big Data можливе лише за умови оптимізації процесів міжвузлової взаємодії. Запропонований метод адаптивної гібридної синхронізації дозволяє зменшити вплив гетерогенності обчислювального середовища та мінімізувати комунікаційні витрати.

Експериментальні результати підтверджують, що застосування даного підходу забезпечує суттєве скорочення часу навчання моделей і підвищення ефективності використання ресурсів, що робить його доцільним для впровадження у високонавантажених аналітичних системах.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Gautam G. A Comprehensive Review of Machine Learning Algorithms for Big Data Analytics. *International Journal for Research in Applied Science and Engineering Technology*. 2025. Vol. 13. P. 3322–3328. https://doi.org/10.22214/ijraset.2025.76719
2. Che E., Dong J., Tong X. Stochastic Gradient Descent with Adaptive Data. *Operations Research*. 2026. March. https://doi.org/10.1287/opre.2024.1014
3. Міронова Ю. В., Юрчук Н. П. Формування адаптивних механізмів управління підприємством на засадах інтелектуального аналізу Big Data. *Інвестиції: практика та досвід*. 2026. № 6. С. 215–222. https://doi.org/10.32702/2306-6814.2026.6.215

Кириченко Катерина Дмитрівна – студентка групи 1ІСТ-24б, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця.

Науковий керівник: ***Юрчук Наталія Петрівна*** – канд. екон. наук, доцент, доцент кафедри менеджменту та безпеки інформаційних систем, Вінницький національний технічний університет, м. Вінниця

Kyrychenko Kateryna Dmytrivna – student of group 1IST-24b, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia.

Supervisor: ***Yurchuk Nataliia P.*** – PhD in Economics, Associate Professor, Associate Professor of the Department of Management and Security of Information Systems, Vinnytsia National Technical University, Vinnytsia.