DOI: 10.34185/1991-7848.itmm.2026.01.079

ПИТАННЯ ВИЗНАЧЕННЯ МІНІМАЛЬНО ДОСТАТНЬОГО ОБСЯГУ НАВЧАЛЬНОЇ ВИБІРКИ ДЛЯ МОДЕЛЕЙ МАШИННОГО НАВЧАННЯ

Кісельов Б.Г.[1] [ORCID], Сенько А.О.[2] [ORCID],

Купін А.І.[3] [ORCID], Балик Д.К.[4] [ORCID]

[1]Криворізький національний університет, Кривий Ріг, аспірант, Україна

[2]Криворізький національний університет, Кривий Ріг, к.т.н., Україна

[3]Криворізький національний університет, Кривий Ріг, д.т.н., професор, Україна

[4]ТОВ «НВП Гамаюн», Кривий Ріг, інженер-програміст, Україна

Анотація. *Тези доповіді присвячено задачі емпіричного визначення мінімально достатнього обсягу навчальної вибірки для регресійних моделей машинного навчання у системах сенсорного сортування руд. Запропоновано методику на основі ієрархії підходів: крива навчання – параметрична степенева екстраполяція – GP-based learning-type curve. Дослідження виконано на реальному наборі сенсорних даних (699 спостережень). Як основну модель обрано HistGradientBoostingRegressor (R^2 = 0,93, 10-fold GroupKFold крос-валідація). Параметрична екстраполяція дозволила отримати точкові оцінки мінімального обсягу вибірки для заданого порогу похибки. GP-based підхід забезпечив ймовірнісні оцінки з урахуванням невизначеності. Сформульовано практичні рекомендації щодо достатнього обсягу вибірки для досягнення цільового рівня точності.*

Ключові слова: *машинне навчання; крива навчання; мінімальний обсяг вибірки; степенева апроксимація; гауссівський процес; сортування руд; крос-валідація; HistGradientBoosting; Neural Scaling Laws; екстраполяція.*

Вступ

У задачах машинного навчання не існує універсальної відповіді на питання про те, яким має бути достатній обсяг навчальної вибірки для досягнення належної якості моделі. Для задачі сенсорного сортування руд ця проблема є особливо актуальною: формування репрезентативної вибірки потребує значних витрат на отримання зразків і виконання вимірювань для різних типів матеріалу. При цьому недостатній обсяг навчальної вибірки може призводити до нестійкого навчання і погіршення узагальнювальної здатності моделі [1].

Виникає потреба в методиці, яка б дозволяла на основі відносно невеликої вибірки емпірично оцінити, як змінюється похибка моделі зі зростанням кількості навчальних прикладів, та спрогнозувати обсяг, необхідний для досягнення цільового рівня точності. Мета роботи – розробка та практична перевірка такої методики для регресійних моделей машинного навчання.

Основний матеріал

Набір даних та модель. Дослідження базується на реальних даних системи сенсорного сортування руд, що містить 699 спостережень без пропущених значень. Цільова змінна – KT10valueMax (максимальне значення вихідного сигналу котушки), яка корелює з вмістом металу у зразку. Ознаковий простір охоплює 8 первинних вимірювань (об'єм, площа, висота зразка, температура сенсора та ін.) і 3 синтезованих похідних ознаки. Найбільш інформативною є ознака SensorValueRAW (r = 0,985). Як основну модель обрано HistGradientBoostingRegressor [2]. Підбір гіперпараметрів виконано за допомогою бібліотеки Optuna (байєсівська оптимізація, 100 спроб). Для оцінювання застосовано 10-fold GroupKFold крос-валідацію (групи – мітки типу руди), що виключає витік інформації між фолдами. Середнє R^2 = 0,93 підтверджує високу якість моделі.

Аналіз кривої навчання. Криву навчання побудовано для 12 розмірів навчальної підвибірки від 20 % до 100 % загального обсягу. Для кожного розміру виконано 10-fold GroupKFold крос-валідацію з трьома повтореннями та випадковим перемішуванням. Крива демонструє монотонне зниження RMSE зі збільшенням N із тенденцією до стабілізації у правій частині – ознаку входження в зону насичення [3].

Параметрична степенева екстраполяція. Для апроксимації кривої навчання обрано степеневу функцію [1]:

$$E(N) = a + b \cdot \left(\frac{N}{Nref}\right)^{-\gamma} \quad (1)$$

де E(N) – значення RMSE при обсязі вибірки N; a – асимптотична RMSE; b – масштаб початкового відхилення від асимптоти; γ – швидкість спадання похибки; Nref – медіана обсягів навчальних підвибірок. Параметри підібрано методом нелінійної регресії (scipy.optimize.curve_fit) з ваговим урахуванням

стандартного відхилення RMSE по фолдах. Прогнозована асимптотична RMSE a = 10,795 задає нижню межу досяжної якості за наявного рівня шуму та неповноти ознак. На основі підігнаної функції отримано точкові оцінки: для досягнення RMSE ⩽ 12 необхідно 559 спостережень; для малопомітного приросту якості (виграш від подвоєння < 0,2 RMSE) – 1004 спостереження. Монте-Карло симуляція (3000 зразків) дала ймовірнісні оцінки P_{50}: 778 та 1197 спостережень відповідно.

GP-based learning-type curve. Як більш просунутий статистичний підхід застосовано GP-based learning-type curve [4]: залишки NLS-підгонки апроксимовано гауссівським процесом із композитним ядром RBF + WhiteKernel (вхід: x = log(N)). Мале значення рівня шуму WhiteKernel підтвердило, що детермінований скелет формули (1) вловлює основну тенденцію, а GP-компонента фіксує систематичні залишки. Порівняльні результати оцінювання мінімально достатнього обсягу вибірки наведено у Таблиці 1.

Таблиця 1

Порівняння оцінок мінімально достатнього обсягу навчальної вибірки

Критерій	NLS (точкова)	NLS + GP (P95)
RMSE ⩽ 12	559 спостережень	810 спостережень
Виграш від подвоєння < 0,2 RMSE	1004 спостереження	1004 спостереження
P_{50} (Монте-Карло)	778 спостережень	~ 810 спостережень

GP-based підхід дає більш консервативну, але статистично обґрунтовану оцінку: для досягнення RMSE ⩽ 12 з надійністю 95 % необхідно 810 спостережень. Оцінки порогу малопомітного приросту якості між двома методами збігаються (≈ 1000 спостережень). Таким чином, для досягнення цільового рівня похибки з достатньою статистичною надійністю рекомендується розширити датасет до 780–810 зразків.

Висновки

Проведена робота підтвердила, що задача визначення мінімально достатнього обсягу навчальної вибірки для моделей машинного навчання є розв'язуваною емпіричними методами на основі аналізу кривих навчання та їх екстраполяції. Модель HistGradientBoostingRegressor забезпечила R^2 = 0,93 на реальних сенсорних даних. Степенева екстраполяція дозволила отримати

кількісні точкові оцінки мінімально достатнього обсягу вибірки; GP-based підхід доповнив їх ймовірнісними оцінками з урахуванням невизначеності параметрів. Для досягнення RMSE ⩽ 12 рекомендується 780–810 спостережень. Запропонована методика може бути застосована до інших задач промислового машинного навчання з обмеженими даними в рамках інтелектуальних інформаційно-управляючих систем.

ЛІТЕРАТУРА

1. Кісельов Б. Г., Сенько А. О. Вплив адитивних стохастичних збурень на нижню межу узагальнювальної похибки моделей регресії в сенсорних системах. Комп'ютерні інтелектуальні системи та мережі : матеріали XIX Всеукраїнської науково-практичної WEB-конференції. Кривий Ріг, 2026. С. 156–159.
2. Ke G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 3146–3154.
3. Viering T., Loog M. The Shape of Learning Curves: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2023. Vol. 45, no. 12. P. 15050–15067. DOI: 10.1109/TPAMI.2021.3085003.
4. Snell K. I. E. et al. Sample size requirements for training clinical prediction models using participant-level meta-analysis. Statistics in Medicine. 2024. Vol. 43, no. 15. P. 2945–2975. DOI: 10.1002/sim.10121.
5. Figueroa R. L. et al. Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making. 2012. Vol. 12. Article 8. DOI: 10.1186/1472-6947-12-8.
6. Kaplan J. et al. Scaling Laws for Neural Language Models. arXiv. 2020. arXiv:2001.08361. DOI: 10.48550/arXiv.2001.08361.
7. Domingos P. A few useful things to know about machine learning. Communications of the ACM. 2012. Vol. 55, no. 10. P. 78–87. DOI: 10.1145/2347736.2347755.

QUESTIONS OF DETERMINING THE MINIMUM SUFFICIENT TRAINING SAMPLE SIZE FOR MACHINE LEARNING MODELS

B. Kiselov, A. Senko, A. Kupin, D. Balyk

Abstract. *This paper addresses the problem of empirically determining the minimum sufficient training sample size for machine learning regression models in ore sensor sorting systems. A methodology based on a hierarchy of approaches is proposed: learning curve – parametric power-law extrapolation – GP-based learning-type curve. The study is conducted on a real sensor dataset (699 observations). HistGradientBoostingRegressor was selected as the primary model ($R^2 = 0.93$, 10-fold GroupKFold cross-validation). Power-law extrapolation provided point estimates of the minimum sample size for a given RMSE threshold. The GP-based approach yielded probabilistic estimates*

accounting for parameter uncertainty. For RMSE ⩽ 12 with 95 % confidence, 810 observations are required. Practical recommendations for datasets of similar type are formulated.

Keywords: *machine learning; learning curve; minimum sample size; power-law approximation; Gaussian process; ore sorting; cross-validation; HistGradientBoosting; Neural Scaling Laws; extrapolation.*

REFERENCE

1. Kiselov, B. H., & Senko, A. O. (2026). Vplyv adytyvnykh stokhastychnykh zburen′ na nyzhniu mezhu uzahalniuval′noi pokhybky modelei rehresii v sensornykh systemakh. Komp'iuterni intelektualni systemy ta merezhi: materialy XIX Vseukrainskoi naukovo-praktychnoi WEB-konferentsii. Kryvyi Rih. P. 96–101. [in Ukrainian].
2. Ke, G. et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.
3. Viering, T., & Loog, M. (2023). The shape of learning curves: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12), 15050–15067. https://doi.org/10.1109/TPAMI.2021.3085003
4. Snell, K. I. E. et al. (2024). Sample size requirements for training clinical prediction models using participant-level meta-analysis. Statistics in Medicine, 43(15), 2945–2975. https://doi.org/10.1002/sim.10121
5. Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. BMC Medical Informatics and Decision Making, 12, Article 8. https://doi.org/10.1186/1472-6947-12-8
6. Kaplan, J. et al. (2020). Scaling laws for neural language models. arXiv. https://doi.org/10.48550/arXiv.2001.08361
7. Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78–87. https://doi.org/10.1145/2347736.2347755