УДК 004.8

ВПЛИВ РОЗМІРУ НАВЧАЛЬНОЇ ВИБІРКИ НА СТАБІЛЬНІСТЬ ТА УЗАГАЛЬНЮВАЛЬНУ ЗДАТНІСТЬ МОДЕЛЕЙ КЛАСИФІКАЦІЇ

Пилипенко В.І., аспірант
Київський національний університет технологій та дизайну

Ключові слова: машинне навчання, класифікація, стабільність моделі, розмір навчальної вибірки, ансамблеві методи, крос-валідація.

Сучасний розвиток методів машинного навчання супроводжується значним зростанням складності моделей та збільшенням вимог до якості та обсягу даних. Одним із ключових факторів, що визначає ефективність побудованих моделей, є розмір навчальної вибірки. Саме обсяг даних безпосередньо впливає на стабільність моделей, їхню узагальнювальну здатність та здатність коректно працювати на нових, раніше невідомих прикладах.

У задачах класифікації стабільність моделі зазвичай визначається через дисперсію її прогнозів при зміні навчальної вибірки, а також через чутливість до варіацій у даних. Теоретично в рамках статистичного навчання показано, що зі збільшенням розміру вибірки дисперсійна складова помилки зменшується, що підвищує загальну стабільність моделі. При цьому загальна помилка моделі може бути представлена у вигляді суми зміщення (bias), дисперсії (variance) та нездоланної випадкової похибки. Теоретично встановлено, що для лінійних моделей дисперсія оцінок параметрів обернено пропорційна розміру вибірки та може бути описана залежністю [1]:

$$Var(\theta) \approx \sigma^2 / (n \cdot I(\theta)), (1)$$

де n–розмір вибірки;
I(θ)–інформація Фішера.

Ця залежність демонструє фундаментальний принцип: зі збільшенням обсягу даних оцінки параметрів стають більш стабільними, а модель менш чутливою до випадкових флуктуацій у вибірці.

Для складних нелінійних моделей, таких як глибокі нейронні мережі, характер залежності є більш складним. У таких випадках спостерігається ефект подвійного спуску (doubledescent), коли помилка моделі спочатку зменшується, потім може зростати в області середніх розмірів вибірки, і знову зменшується при подальшому збільшенні обсягу даних. Це ускладнює визначення оптимального розміру вибірки для навчання моделей високої складності.

Експериментальні дослідження проведено на багатокласовій задачі класифікації академічної успішності студентів із варіацією розміру вибірки [2]. Оцінювання здійснювалося за допомогою 10-кратної стратифікованої крос-валідації, з урахуванням ефективності моделей прогнозування [3]. Побудований аналіз точок насичення представлено на рис. 1.

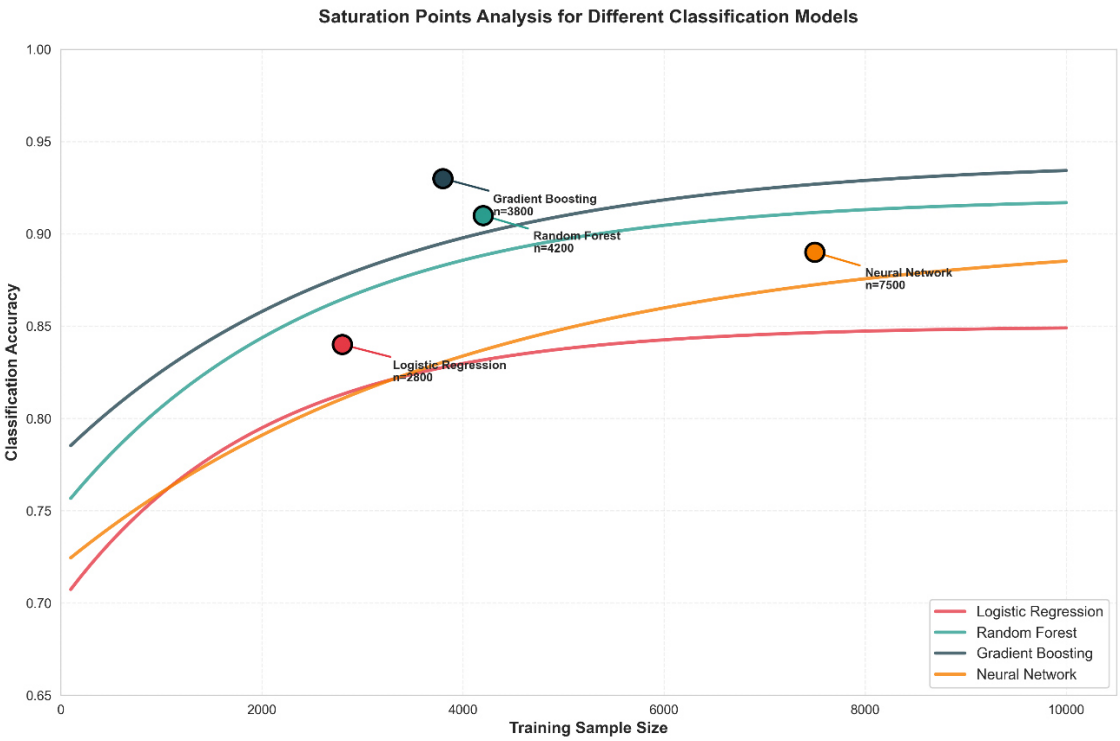

Рисунок 1 – Побудова отриманих точок насичення

Для логістичної регресії стабільна робота моделі, що характеризується низьким коефіцієнтом варіації результатів (CV < 0,10), досягається вже при обсязі вибірки близько 2500–3000 прикладів. Така поведінка пояснюється відносно невисокою складністю моделі та її лінійною природою. Логістична регресія належить до параметричних методів машинного навчання, у яких кількість параметрів є обмеженою та безпосередньо пов'язана з кількістю ознак. Завдяки цьому модель демонструє добру узагальнювальну здатність навіть на порівняно невеликих наборах даних. Точка насичення для даної моделі становить приблизно 2800 прикладів, після чого приріст якості прогнозування є мінімальним, а додавання нових даних практично не впливає на стабільність результатів. Це свідчить про досягнення оптимального співвідношення між зміщенням (bias) та дисперсією (variance) моделі.

Для моделей RandomForest та GradientBoosting стабільність прогнозування досягається при значно більшому обсязі вибірки–приблизно 3800–4200 прикладів. Це пояснюється особливостями ансамблевих методів навчання, які базуються на комбінуванні великої кількості базових моделей. RandomForest реалізує механізм bootstrap-агрегування (bagging), що дозволяє зменшити дисперсію шляхом усереднення прогнозів багатьох дерев рішень. У свою чергу GradientBoosting формує ансамбль послідовно, мінімізуючи помилки попередніх моделей через оптимізацію функції втрат. Незважаючи на високу здатність до виявлення складних нелінійних залежностей, такі моделі потребують більшого обсягу даних для уникнення перенавчання та забезпечення статистичної стійкості. Виявлена точка насичення свідчить про те, що після приблизно 4000 прикладів

зменшення варіативності результатів стає незначним, а learningcurves переходять у фазу плато.

Для нейронних мереж точка насичення спостерігається лише при значно більших обсягах даних 5000 –10000 прикладів. Це узгоджується з сучасними теоретичними уявленнями про глибоке навчання, згідно з якими моделі з великою кількістю параметрів мають високу апроксимаційну здатність, але одночасно характеризуються підвищеною дисперсією та чутливістю до розміру вибірки. Нейронні мережі здатні моделювати складні багатовимірні нелінійні залежності, однак ефективне налаштування вагових коефіцієнтів потребує значної кількості навчальних прикладів.

Отримані результати мають важливе теоретичне та практичне значення для розроблення систем машинного навчання у сфері освітньої аналітики, прогнозування успішності студентів та виявлення ризику академічної неуспішності [4]. Практична цінність дослідження також підтверджується результатами опитувань здобувачів освіти щодо доцільності впровадження програмного забезпечення для прогнозування успішності та персоналізації освітнього процесу [5]. Розуміння залежності між обсягом вибірки та стабільністю моделей дозволяє формувати ефективні стратегії збору й обробки даних, визначати мінімально необхідний обсяг навчальної інформації, зменшувати витрати на формування датасетів, оптимізувати вибір алгоритмів відповідно до доступних ресурсів, а також підвищувати надійність і точність прогнозних систем.

Список використаних джерел

1. Belkin, M., Hsu, D., Ma, S., &Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings of the National Academy of Sciences, 116(32), 15849-15854.

2. Pylypenko, V. (2025). The effect of training sample size on the stability of classification models. Technologies and Engineering, 26(6), 32-44. https://doi.org/10.30857/2786-5371.2025.6.3

3. S. Volodymyr, V. Pylypenko, V. Skidan and A. Volivach, "Investigation of the Accuracy of Machine Learning Methods in Prediction of Students Success," 2024 IEEE 5th KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2024, pp. 1-4, doi: 10.1109/KhPIWeek61434.2024.10877975

4. Пилипенко В. (2025). Прогнозування високого рівня академічної успішності студентів з використанням машинного навчання. *Наука і техніка сьогодні*, 8(45), 1634–1649. https://doi.org/10.52058/2786-6025-2025-8(49)-1634-1649

5. Пилипенко, В., Скідан, В., & Волівач, А. (2024). Аналіз опитування щодо впровадження програмного забезпечення прогнозування успішності здобувачів вищої освіти. *Herald of Khmelnytskyi National University. Technical sciences*, *345*(6 (2)), 108-112. https://doi.org/10.31891/2307-5732-2024-345-6-16