ІНФОКОМУНІКАЦІЙНІ ТЕХНОЛОГІЇ ТА ЕЛЕКТРОННА ІНЖЕНЕРІЯ
INFORMATION AND COMMUNICATION TECHNOLOGIES, ELECTRONIC ENGINEERING

№ 5 (2), 2025

https://doi.org/10.23939/ictee2025.02.

ОПТИМІЗАЦІЯ НАВЧАЛЬНОЇ ВИБІРКИ ЗА ДОПОМОГОЮ ВИПАДКОВИХ ТОЧКОВИХ ПРОЦЕСІВ

О. Луцик [ORCID: 0000-0003-1707-3532], **Б. Русин** [ORCID: 0000-0001-8654-2270], **Р. Косаревич** [ORCID: 0000-0001-9108-0365]

Фізико-механічний інститут ім. Г. В. Карпенка Національної академії наук України, вул. Наукова, 5, Львів, 79013, Україна

Відповідальний за рукопис: О. Луцик (e-mail: olutsyk@yahoo.com)

(Подано 1 серпня 2025)

У роботі розглянуто методи оптимізації навчальних вибірок для алгоритмів глибокого навчання із застосуванням випадкових точкових процесів, таких як Матерна першого і другого типу, гіббсівський, гауссівський та пуассонівський процеси. Запропоновано підхід до зменшення навчальних даних без втрати їхньої інформативності, що дає змогу зменшити обчислювальні витрати та проблеми із перенавчання. Запропоновано новий підхід подання зображень у вигляді випадкового точкового процесу, який дає можливість перейти від піксельного вигляду зображення до моделі, яка підходить для аналізу методами випадкових точкових процесів. Такий перехід до компактнішої емпіричної форми полегшує подальші аналіз та моделювання. Завдяки цьому уможливлюється використання інструментів статистики для описання закономірностей, прихованих у зображеннях. Продемонстровано ефективність випадкових точкових процесів для формування простору ознак, аналізу покриття та структурованості навчальної вибірки. Розглянуто вплив різних типів точкових процесів на збалансованість даних та їх здатність зменшувати надмірність у вибірці. Увагу звернено зокрема на репрезентативність даних, оскільки вона визначає стійкість і узагальнювальну здатність моделей глибокого навчання. Наведено алгоритми перетворення зображень на точкові процеси та їх використання для балансування класів, розрідження даних та підвищення репрезентативності вибірки. Подано результати оцінювання ймовірності правильної класифікації на прикладі моделей ResNet, які демонструють переваги застосування точкових процесів порівняно із випадковим проріджанням даних. Підтверджено ефективність точкових процесів для оптимізації великих навчальних наборів та підвищення точності глибокого навчання. Отримані результати свідчать, що застосування таких методів може стати ключовим для створення ефективніших моделей нейронних мереж. Окреслено перспективи подальших досліджень у напрямі адаптивної оптимізації навчальних вибірок, під час яких випадкові точкові процеси можуть стати основою для нових методів підготовки.

Ключові слова: *навчальна вибірка, нейронні мережі, випадкові точкові процеси, оптимізація, обчислювальна складність.*
УДК: 004.8, 681.3, 621.391, 004.9

Вступ

У сучасних технологіях машинне навчання займає центральне місце, забезпечуючи автоматизацію складних процесів, допомогу для прийняття рішень, контроль якості та інтелектуальну обробку інформації. Сьогодні значна частка інтелектуальних систем спирається на методи глибокого навчання, зокрема на нейронні мережі, здатні ефективно працювати з величезними обсягами

даних. На відміну від традиційних алгоритмів машинного навчання, які часто супроводжуються обмеженнями під час оброблення складних або великих наборів даних, нейромережеві моделі дають змогу виявляти складні закономірності, розпізнавати приховані структури та будувати точні прогнози на основі великої кількості інформації [1–3].

Проте під час упровадження таких систем постають вимоги щодо масштабованості та ефективності обчислень. Глибокі нейронні мережі потребують значних обчислювальних ресурсів та великих обсягів тренувальних даних. У багатьох випадках цілісне використання всіх доступних даних стає надмірно витратним як за часом, так і за обчислювальними ресурсами. Це зумовлює важливість розроблення підходу оптимізації навчальної вибірки, які дають змогу зменшити її обсяг, одночасно зберігаючи варіативність та репрезентативність даних. Ефективне зменшення навчальної вибірки допомагає пришвидшити навчання, зменшити обчислювальні витрати та знизити ризик перенавчання моделей, підтримуючи високу точність класифікації. Важливо, де замість повної множини даних використовується обмежена кількість узагальнених представлень класів. За структурою це схоже на прототипне навчання. Наприклад, у роботі [4] запропоновані прототипні мережі, які використовують середні вектори у гіперпросторі як прототипи класів, що показало високу ефективність під час навчання. Аналогічні ідеї застосовано у методах класифікації з малим обсягом даних у роботах [5, 6].

Паралельно розвивається напрям пониження навчальної вибірки, метою якого є усунення надлишкових або неінформативних прикладів без погіршення якості навчання. Класичні методи [6–8] заклали основи для сучасних алгоритмів пониження навчальної вибірки. Сучасніші підходи [10, 11] пропонують формувати підвибірку, що охоплює структуру повної множини, з мінімальними втратами інформації. Зокрема у роботі [10] запропоновано стратегію активного вибору найінформативніших елементів для глибоких мереж за допомогою крос-аналізу, що істотно зменшує потребу у великих обсягах даних.

З погляду подання даних важливе місце займають автоенкодери, які дають змогу будувати стиснені векторні образи даних у прихованому просторі. Робота [13] є фундаментальною і показала, що нейронні мережі можуть ефективно зменшувати розмірність даних з мінімальними втратами. Приховані вектори часто використовують як основу для побудови точкових образів.

У випадках роботи з 3D-даними та хмар точок особливу увагу варто звернути на архітектури, описані в роботі [14], які оперують із необробленими точковими даними, демонструючи можливість ефективного навчання із мінімальними структурами. Цей підхід є основним для ієрархічного подання точок, де кожен фрагмент простору представлено зведеною групою точок, що може інтерпретуватися як точковий образ.

Окремим напрямом є активне навчання та семплювання, де дані вибирають не випадково, а з урахуванням їхнього внеску в покращення моделі. Робота [9] описує методи активного навчання, зокрема семплювання та крос-підходи. У роботі [15] наведено практичні алгоритми побудови ефективних навчальних множин на основі вибору найважливіших елементів вибірки. Точкові образи в роботі [17] також можна використовувати для адаптації моделей до нових доменів з обмеженим набором даних.

У багатьох сучасних завданнях комп’ютерного зору дані природно виникають у вигляді випадкових точкових образів (*random point patterns*). Це множини точок, розміщених у просторі (2D, 3D або більше); як кількість точок, так і їх положення можуть бути випадковими та змінюватися від одного зразка до іншого. Такі образи трапляються, наприклад, у даних із сенсорів LiDAR, зображеннях дистанційного зондування Землі або у даних щодо розсіяння зірок у астрономії.

Класичні методи аналізу зображень або векторних даних часто виявляються неефективними, оскільки не враховують інваріантність до перестановок точок, змін у кількості точок, і специфіку просторової структури. Тому виникає потреба у створенні або адаптації моделей, які можуть опрацьовувати цілі образи точок як вхідні об’єкти, а не як фіксовані вектори ознак. З цією метою необхідно розвивати методи аналізу зі статистичної теорії випадкових точкових процесів, а також із глибоким навчанням на множинах, які дають змогу моделювати та класифікувати образи з ура-

хуванням інваріантності до перестановок, масштабів та зсувів. Одним з важливих завдань є побудова компактного, узагальнювального подання, для кожного образу, що зберігає важливу інформацію про його структуру, щільність та форму.

У роботі показано можливості використання випадкових точкових образів для оптимізації навчальної вибірки під час тренування нейронних мереж чи загалом у задачах машинного навчання. Зокрема, розроблено підходи до генерації навчальних вибірок на основі точкових образів з контрольованими просторовими властивостями, такими як щільність, кластеризація та рівномірність, які є ключовими для створення навчальних вибірок. Проаналізовано вплив різних типів точкових образів (наприклад, пуассонівських, регулярних, Матерна, кластеризаційних) на якість узагальнення моделей глибокого навчання. Показано переваги вибірок, створених за допомогою точкових образів, порівняно з традиційними методами вибору навчальних баз даних, такими як випадкова вибірка із усієї множини або стратифікація. Реалізовано застосування точкових образів у задачах з обмеженою кількістю анотованих даних, коли важливе максимальне охоплення простору ознак.

Розроблено практичні рекомендації щодо використання випадкових точкових образів як евристики для активного навчання моделей. Дослідження поєднує методи стохастичної теорії, статистики даних і глибокого навчання для створення ефективніших, компактніших та інформативніших навчальних наборів.

2. Подання зображень у вигляді випадкового точкового процесу

Зображення подаються у вигляді регулярних матриць інтенсивностей пікселів. Для статистичного моделювання та аналізу просторових структур зображення потрібно перетворити у форму для інтерпретації зображень як точкових випадкових процесів. Це дасть змогу розглядати окремі об’єкти, деталі або ознаки зображення як випадкові точки у деякому просторі, що уможливлює використання теорії точкових процесів.

Наприклад, маємо зображення $I(x,y)$, де (x,y) є індексами окремих пікселів, а матриця $I(x,y)$ – це інтенсивність, яка у випадку кольорового зображення складається із трьох матриць. Важливо привести до єдиного стандарту порівнюваність зображень, для цього можна здійснити нормалізацію відповідно до такого виразу:

$$\overset{N}{I}(x,y)=\frac{I(x,y)-\min(I(x,y))}{\max(I(x,y))-\min(I(x,y))}. \tag{1}$$

За допомогою порогування можна ефективно зменшити розмірність зображення. Найпростішим варіантом є бінарне представлення:

$$B(x,y)=\begin{cases}1, \overset{N}{I}(x,y)>h\\ 0, \overset{N}{I}(x,y)<h\end{cases}, \tag{2}$$

де h – порогове значення, яке вибирають адаптивним способом залежно від інтенсивності зображення.

Після цього для кожного пікселя (x,y) зображення обчислюють ймовірність потрапляння у точковий процес як функцію від інтенсивності:

$$p(x,y)=f(B(x,y)). \tag{3}$$

У результаті формується реалізація точкового випадкового процесу. Коли $Z(x,y)=1$, то конкретний піксель (x,y) залучається у процес:

$$\Phi=\{(x,y)\in W \mid Z(x,y)=1\}. \tag{4}$$

Цей підхід перетворення зображення на точковий випадковий процес дає змогу перейти від піксельного представлення зображення до моделі, яка підходить для аналізу методами випадкових точкових процесів у компактній емпіричній формі, що полегшує надалі аналіз та моделювання. Завдяки цьому можна використовувати інструменти статистики для описання закономірностей, прихованих у зображеннях. Також цей підхід не має обмежень за розміром і дає змогу застосовувати його як для простих, так і для високорозмірних зображень, незалежно від їхнього призначення для медицини, супутникові зображення дистанційного зондування Землі, технічні зображення. Опрацювання даних у цьому випадку здійснюється не із суцільними масивами пікселів, а з набором точок у просторі, що знижує обчислювальні затрати.

Недоліком такого подання, як виявлено під час роботи, є втрата частини інформації у результаті дискретизації у вигляді точок, бо не завжди можливо точно передати градієнти інтенсивності, текстурні особливості чи дрібні деталі, іноді критично важливі для аналізу зображень. Але вплив втрат на загальну ефективність перетворення зображень у вигляді випадкових точкових процесів не є істотним.

3. Застосування випадкових точкових процесів

Одним із найпростіших точкових випадкових процесів є пуассонівський процес. Він уможливлює моделювання кластеризаційних процесів. У цьому випадку дані не розподіляються рівномірно у просторі ознак, а формують групи, які відображають локальні структури даних.

Алгоритм формування процесу складається із двох етапів. Спочатку генерується множина початкових точок, які є центрами майбутніх кластерів. Вони з'являються випадково відповідно до однорідного пуассонівського процесу з інтенсивністю λ_p. Після цього навколо кожного такого центра формується певна кількість вторинних точок, які зміщуються відносно початкових за нормальним законом розподілу із нульовим середнім і дисперсією. В результаті отримуємо вибірку із об'єктами, зібраними в групи.

Позначимо інтенсивність пуассонівського процесу, який породжує вторинні точки, через λ_p. Тоді кількість таких точок у фіксованій області визначається розподілом Пуассона:

$$N_p(A) = P(\lambda_p \mid A). \tag{5}$$

Для кожної первинної точки генерується випадкова кількість вторинних точок C_i, кожна з яких також може мати пуассонівський розподіл з параметром μ:

$$C_i = P(\mu). \tag{6}$$

Координати кожної вторинної точки утворюються як випадкове зміщення відносно положення початкової точки:

$$y_{ij} = x_i + e_{ij}. \tag{7}$$

У результаті повна множина даних має вигляд:

$$Y = \bigcup_{i=1}^{N_p} \bigcup_{j=1}^{C_i} y_{ij} \tag{8}$$

Це означає, що вибірка складається із набору кластерів, кожен з яких локалізований навколо випадково вибраного центра, а структура кластера описується нормальним розподілом.

Процеси Матерна першого і другого типу складніші. Процес Матерна першого типу ґрунтується на процесі Пуассона із густиною $\lambda > 0$, що генерує множину точок:

$$\Phi = \{x_i\} \subset R^d. \tag{9}$$

Відтак необхідно виконати відбір, за якого кожна точка $x_i \hat{I}\ F$ залишається у фінальному процесі F_M за умови, що в радіусі $r > 0$ навколо неї не міститься жодної іншої точки:

$$F_M = \{x_i \hat{I}\ F : B(x_i, r) \c{C} (F\{x_i\}) = 0\}, \tag{10}$$

де $B(x_i, r)$ – простір, обмежений кулею із центром x_i і радіусом r.

Відмінністю процесу Матерна першого типу є те, що залишаються лише точки, які не мають сусідів на близькій відстані. Це створює вибірку з мінімальною відстанню між точками, усуваючи надмірні скупчення.

Якщо кожен елемент у навчальній вибірці відповідає окремій точці у багатовимірному просторі ознак, то густина цих точок може істотно впливати на якість побудованої моделі. В реальних умовах навчальна вибірка часто містить дублікати, які є дуже близькими або повністю повторюються, розташовані надзвичайно близько один до одного у цьому просторі. Такі надмірності не лише збільшують обчислювальну складність процесу навчання, але й можуть призводити до погіршення здатності нейронної мережі узагальнювати, оскільки вона навчається інформацією, яка не додає нової інформації. Точкові випадкові процеси Матерна першого типу дають змогу уникнути цієї проблеми завдяки параметру мінімальної відстані, який визначає, наскільки близько дві точки можуть розташовуватися одна від одної, щоб вважатися незалежними. Якщо у навчальній вибірці з'являються два вектори з відстанню менше ніж r, один із них відкидається. У результаті залишається лише та частина навчальної вибірки, у якій кожна точка ізольована від інших на визначеному рівні.

Такий підхід уможливлює проріджування даних. Наприклад, якщо вихідна вибірка містить багато повторюваних або майже однакових зображень, застосування фільтра Матерна першого типу приводить до істотного зменшення їх кількості без втрати різноманітності інформації, яку містять зображення. За допомогою цього методу зберігаються лише ті елементи навчальної вибірки, які вносять нову інформацію щодо структури даних.

З практичного погляду це означає, що нейронна мережа отримує навчальну вибірку без надлишкових дублікатів. В цих умовах оптимізація ваг відбувається ефективніше, бо нейромережева модель більше уваги приділяє суттєвим відмінностям між елементами навчальної вибірки, а не повторює багаторазове запам'ятовування того самого образу. Також зменшуються ризики перенавчання, тривалість навчання, оскільки кількість унікальних елементі вибірки менша, ніж у початковій навчальній вибірці.

Обчислювано складнішим є процес Матерна другого типу. Він також є процесом Пуассона, але використовує впорядкування точок.

Кожній точці $x_i \hat{I}\ F$ довільно присвоюється деяка мітка t_i. Точка потрапляє під процес F_M, якщо в її околі $B(x_i, r)$ не існує іншої точки x_j із меншою міткою $t_j < t_i$, а саме:

$$F_M = \{x_i \hat{I}\ F : " x_j \hat{I}\ B(x_i, r), t_j < t_i\}. \tag{11}$$

Цей підхід гарантує, що в околі кожного радіуса r залишиться лише одна точка, яка за рахунок найменшої мітки набула найбільшої ваги.

Процес Матерна другого типу є модифікацією стохастичних методів проріджування точок у просторі та має істотні переваги порівняно з базовим підходом Матерна. Якщо у першому випадку після генерації випадкової множини точок деякі з них відкидаються за певним критерієм відстані, що може призвести до нерівномірності покриття простору оснак, то в другому підході застосовується збалансованіша процедура відбору. Тут кожна точка має рівні умови для відбору, і кінцева структура не містить надлишкових скупчень або частин простору, де немає точок. У результаті виходить набір точок, які розташовані регулярніше, з однаковими інтервалами між сусідніми

елементами. В задачі оптимізації навчальної вибірки цей підхід забезпечує краще представлення всього простору ознак. Якщо у вибірці залишати лише ті дані, які отримали найвищу вагу за відстань до сусідів, то гарантовано, що навчальний набіп даних охоплюватиме різні ділянки простору рівномірніше. Це дуже важливо, бо алгоритми навчання чутливі до дисбалансу вхідних даних: надмірна концентрація елементів у певній області може призвести до того, що модель переорієнтується саме на цю локальну частину, оминаючи увагою інші ділянки простору. А цей підхід мінімізує такий ризик, створюючи краще збалансоване відображення структури даних.

Експерименти із реальними базами зображень показують, що вибірка, сформована на основі Матерна другого типу, має властивість кращої репрезентативності даних. Класифікація даних у багатовимірних просторах забезпечує, що жоден клас або підклас не буде представлений локалізовано, а розташування елементів сприяє ефективнішому навчанню моделі, яка узагальнює закономірності, а не підлаштовує свої параметри під локальні особливості. Отже, цей підхід дає змогує зменшити кількість даних без втрати нформативності а також істотно поліпшити якість навчання.

4. Задача балансування класів

Окремо можна виділити задачу балансування класів за допомогою точкових випадкових процесів. У навчальних вибірках часто є проблема нерівномірного розподілу даних між класами: один клас може бути представлений тисячами зразків, тоді як інший лише кількома десятками. Це приводить до перекосу під час навчання моделі, що проявляється у поганій класифікації менш представленого класу. В класах, де елементів замало, за допомогою точкових випадкових процесів штучно відтворюються нові карти точок, зберігаючи закономірності просторового розташування, які схожі зі справжніми даними. У результаті набір даних доповнюється, і співвідношення між класами вирівнюється.

Це реалізується послідовністю кроків. Спочатку відбувається аналіз даних: для кожного класу визначають, наскільки він збалансований стосовно інших. Якщо клас має значно менше елементів, він стає кандидатом на доповнення. Далі зображення з цього класу перетворюють на точкові процеси, щоб отримати узагальнену модель просторового розташування точок. Ця модель є статистичною основою для генерації нових елементів. Під час генерації точковий випадковий процес доповнює нові конфігурації точок, щільність, розподіл та характерні локальні закономірності яких такі самі, як і вихідних даних. Після цього точкові карти можна використовувати безпосередньо як вхідні ознаки для навчання нейронних мереж.

Балансування за допомогою точкових випадкових процесів також дає змогу легко контролювати варіативність навчальної вибірки. Можна задавати параметри щільності, рівень шуму чи масштаб відхилень, досягаючи потрібного рівня узагальнення моделі на цих даних. Якщо якийсь клас навчальної вибірки складний, можна локально зберегти більшу подібність до оригінальних елементів або, навпаки, вносити більшу варіативність, якщо модель потребує ширшого покриття класів.

На практиці ефект від застосування цього методу для покращення якості навчання особливо відчутний у складних задачах – навчальних вибірках із нерівномірними наборами даних.

5. Обчислювальна складність під час оптимізації навчальної вибірки

Питання обчислювальної складності алгоритмів дуже важливе, особливо під час роботи із навчальними вибірками, оскільки вони можуть бути дуже великими. За великої навчальної вибірки висока обчислювальна складність призводить до істотних часових затрат та неефективного використання обчислювальних потужностей. У випадку із випадковими точковими процесами кожний із підходів має певні переваги, які, однак, супроводжуються додатковою обчислювальною складністю.

Простий пуассонівський точковий процес має низьку обчислювальну складність близько $O(N)$ і не потребує додаткових обчислень для взаємодії між точками. Його доцільно використовувати, коли потрібне максимально швидке оброблення навчальної вибірки. Підходить для задач балансування класів, де головне – отримати розподіл даних, близький до рівномірного, без складних залежностей.

Для гауссівського точкового процесу оптимізація вибірки потребує попереднього обчислення щільності через гауссівський процес. Обчислювальна складність для побудови ядра гауссівського процесу приблизно $O(N^3)$. Для великих даних алгоритм стає дуже повільним. З погляду обчислювальної складності його доцільно використовувати, коли у даних складна просторова структура та чітко виражені кластери, а також для задач, у яких важливо враховувати кореляцію між зразками. Через значну обчислювальну складність їх доцільно використовувати, якщо кількість даних порівняно невелика і потрібна висока точність врахування залежностей.

Складність точкових процесів Гіббса – $O(N^2)$. Вони ефективні, коли потрібно уникнути скупчення точок у навчальній вибірці. Добре проявляють себе в задачах, де важлива варіативність вибраних елементів вибірки і якщо поставлено завдання відбору репрезентативних даних із великих наборів.

Для процесу Матерна першого типу основна складність полягає у перевірці відстаней між точками. Якщо наївно порівнювати кожну точку з усіма іншими, це потребує $O(N^2)$ операцій, N – кількість початково згенерованих точок. Але на практиці застосовують спрощення, що дає змогу зменшити складність до приблизно $O(N \log N)$.

Процес Матерна другого типу з погляду обчислювальної складності близький до Матерна першого типу, загальна обчислювальна складність у цьому випадку залишається у межах $O(N \log N)$. Його використання виправдане, коли необхідно зберегти контроль за розташуванням точок, залишивши найпріоритетніші. Добре підходить для оптимізації навчальної вибірки, коли треба уникати як надто щільних скупчень, так і проріджения. Використовується у випадках, коли важливо, щоб розподіл вибірки був стабільним і відтворюваним, а не як за процесу Матерна першого типу.

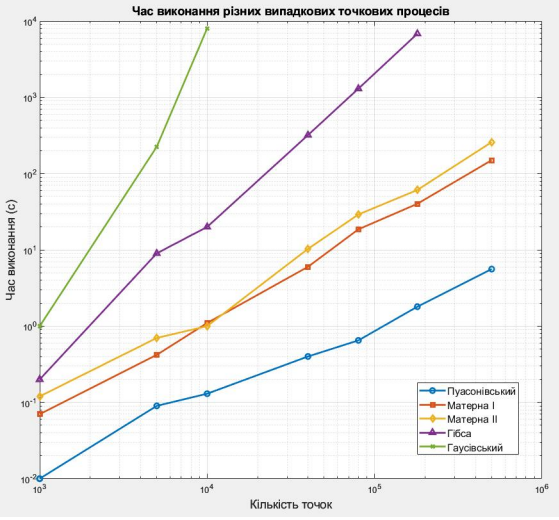

Рис. 1. Залежність часу обчислень від кількості точок

На рис. 1 показано залежність часу обчислень (у секундах, по вертикальній осі) від кількості точок (по горизонтальній осі у логарифмічному масштабі) для п’яти основних класів випадкових точкових процесів. Графіки отримано для процесора Intel Core i5-8300H з 32 гігабайтами оперативної пам’яті.

6. Застосування до реальних даних

Використання випадкових точкових процесів відкриває нові можливості для оптимізації навчальних вібірок. Як тестову вибрано базу даних ImageNet, яка складається із величезної кількості зображень, розділених на окремі класи. На рис. 2 показано залежність правильної імовірності класифікації моделі ResNet від розміру навчальної вибірки після застосування різних методів оптимізації за допомогою точкових випадкових процесів. Можна оцінити ефективність цих методів порівняно зі звичайним прорідженням даних. Вісь X, побудована в логарифмічному масштабі та у послідовності зниження, відображає зменшення розміру навчальної вибірки від максимальної кількості елементів до мінімальної. Вісь Y відображає ймовірність правильної класифікації, що дає змогу оцінити, як зменшення об'єму даних впливає на точність моделі.

На графіку можна спостерігати, що звичайне прорідження спричиняє істотне зниження точності моделі у разі зменшення вибірки. Ймовірність правильних передбачень знижується різкіше порівняно з усіма кривими, які відповідають точковим процесам, що свідчить про втрату репрезентативності даних. Це пояснюється тим, що звичайне випадкове зменшення не контролює покриття простору ознак, і деякі важливі області даних можуть залишатися недостатньо представленими. Криві, отримані за допомогою пуассонівського процесу, демонструють плавніше зниження точності. Навіть за істотного зменшення вибірки ймовірність правильної класифікації на порівняно високому рівні, оскільки пуассонівський процес забезпечує рівномірне розподілення точок, що допомагає уникнути пропуску критичних ділянок простору ознак.

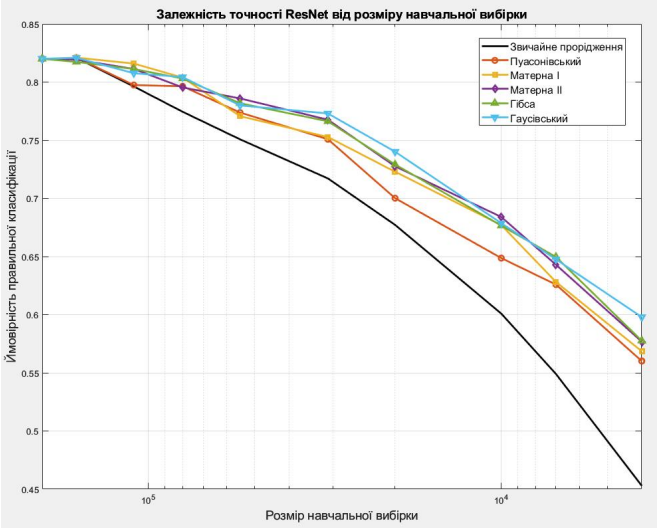

Рис. 2. Залежність точності правильної класифікації від розміру навчальної вибірки

Процеси Матерна першого та другого типів показують кращу стабільність точності із зменшенням навчальної вибірки. Процес Матерна другого типу забезпечує рівномірніше та регулярніше розташування точок у вибірці, це відображається як найбільш плавна та стабільна крива, не втрачає точності у разі істотного зменшення розміру навчальної вибірки.

Гауссівський точковий процес дає найкращі результати і дає змогу зберігати якість класифікації у разі зменшення вибірки, але висока обчислювальна складність робить цей метод недоцільним.

Загалом графік демонструє перевагу використання точкових випадкових процесів для оптимізації навчальної вибірки. Вони дають змогу зменшувати обсяг даних без суттєвої втрати точності, забезпечуючи ефективніше навчання моделей і зменшення ризику перенавчання. Використання структурованих методів зменшення навчальної вибірки за допомогою випадкових точкових про-

цесів забезпечує хороший баланс між розміром вибірки і збереженням високої якості класифікації, і це підтверджує ефективність підходів Матерна і Гібса. Така оптимізація особливо корисна у разі роботи з великими наборами даних, такими як ImageNet, коли пряме навчання на всій вибірці потребує значних обчислювальних ресурсів.

Висновок

У результаті аналізування різних типів випадкових точкових процесів показано, що їх застосування до оптимізації навчальної вибірки нейронних мереж дає змогу досягти істотного підвищення ефективності навчання. Процеси Матерна другого типу забезпечують контрольоване розрідження даних, що дає змогу уникнути надлишкових дублювань та підвищити варіативність елементів у просторі ознак. Гауссовий випадковий точковий процес корисний для моделювання природних кластерів даних і може бути застосований у задачах кластеризації та виявлення структур у вибірці. Інші процеси, такі як пуассонівський чи гіббсовий, дають змогу варіювати рівень випадковості та регулярності розподілу точок, що робить їх гнучкими інструментами для підготовки репрезентативних навчальних наборів. Використання цих підходів дає змогу істотно зменшити обсяг даних без втрати інформативності, зменшити тривалість навчання нейронних мереж та підвищити їх здатність до узагальнення.

Отже, застосування випадкових точкових процесів відкриває нові можливості для оптимізації навчальних вибірок у сучасних методах глибокого навчання. Воно дає змогу ефективно використовувати обмежені обчислювальні ресурси, гарантує рівномірніше покриття простору ознак, що критично для побудови точних моделей. Використання таких методів особливо актуальне під час роботи з великими масивами даних, коли ручна оптимізація вибірки або класичні методи підбору прикладів малоефективні. Отримані результати підкреслюють значення інтеграції методів теорії точкових процесів у практику побудови навчальних наборів для нейронних мереж, відкриваючи перспективи для подальших досліджень у цій сфері.

Список використаних літературних джерел

[1] Alzubaidi, L., Zhang, J., Humaidi, A., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. & Al-Amidie, M. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data, 53. DOI: 10.1186/s40537-021-00444-8

[2] Pichler, M. & Hartig, F. (2023). Machine learning and deep learning – A review for ecologists. Methods in Ecology and Evolution. DOI: 10.1111/2041-210X.14061

[3] Awais, M., Chiari, L., Ihlen, E., Helbostad, J. & Palmerini, L. (2021). Classical Machine Learning versus Deep Learning for the Older Adults Free-Living Activity Classification. Sensors, 21. DOI: 10.3390/s21144669

[4] Snell, J., Swersky, K. & Zemel, R. (2017). Prototypical Networks for Few-shot Learning. Advances in Neural Information Processing Systems (NeurIPS).

[5] Mensink, T., Verbeek, J., Perronnin, F. & Csurka, G. (2013). Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[6] Hart, P. E. (1968). The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory, 14(3), pp. 515–516.

[7] Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. IEEE Transactions on Systems, Man, and Cybernetics, 2(3), pp. 408–421.

[8] Angelova, A., Zhu, S. & Yan, S. (2005). Pruning Training Sets for Learning of Object Categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[9] Settles, B. (2009). Active Learning Literature Survey. University of Wisconsin-Madison.

[10] Sener, O. & Savarese, S. (2018). Active Learning for Convolutional Neural Networks: A Core-Set Approach. International Conference on Learning Representations (ICLR).

[11] Bachem, O., Lucic, M. & Krause, A. (2017). Practical Coreset Constructions for Machine Learning. arXiv preprint.

[12] Mirzasoleiman, B., Bilmes, J. & Leskovec, J. (2020). Coresets for Data-efficient Training of Machine Learning Models. arXiv preprint.
[13] Hinton, G. E. & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. Science, 313(5786), pp. 504–507. DOI:10.1126/science.1127647
[14] Qi, C. R., Su, H., Mo, K. & Guibas, L. J. (2017). PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. CVPR.
[15] Huang, S., Zhang, Y., Deng, J. & Yu, K. (2020). Deep Prototype Learning for Classification. In: International Conference on Learning Representations (ICLR).
[16] Bengio, Y., Louradour, J., Collobert, R. & Weston, J. (2009). Curriculum Learning. Advances in Neural Information Processing Systems (NeurIPS).
[17] Tran, K., Nguyen, M., Phung, D. & Venkatesh, S. (2021). Transferable Prototype Learning across Domains. arXiv preprint

OPTIMIZATION OF TRAINING SAMPLE USING RANDOM POINT PROCESSES

Oleksiy Lutsyk, Bohdan Rusyn, Rostyslav Kosarevych

Karpenko Institute of Physics and Mechanics, NAS of Ukraine, 5, Naukova str., Lviv, 79000, Ukraine

The paper considers methods for optimizing training samples for deep learning algorithms through the use of random point processes, such as Matern of the first and second types, Gibbs, Gaussian, and Poisson processes. An approach to reducing training data without sacrificing informativeness is proposed, enabling a decrease in computational costs and mitigating the risk of overfitting. A novel method of representing images as random point processes is introduced, allowing a transition from the pixel based representation of an image to a model more suitable for analysis with point process techniques. This transition to a more compact empirical form facilitates subsequent analysis and modeling. In addition, it enables the use of statistical tools to uncover patterns hidden within images. The effectiveness of random point processes in shaping the feature space, analyzing coverage, and structuring the training dataset is demonstrated. The study also considers the impact of different types of point processes on data balance and their ability to reduce redundancy within the sample. Particular attention is given to the issue of data representativeness, as it directly affects the stability and generalization capability of deep learning models. Algorithms for converting images into point processes and their application for class balancing, data thinning, and enhancing the representativeness of samples are presented. The evaluation of classification accuracy, conducted using ResNet models, highlights the advantages of applying point processes over random data thinning. The results confirm the effectiveness of point processes for optimizing large-scale training datasets and improving the accuracy of deep learning. Furthermore, the findings indicate that these methods may play a key role in developing more lightweight and efficient neural network models. The study outlines promising directions for future research in adaptive optimization of training samples, where random point processes may serve as the foundation for new approaches to data preparation in neural network training.

Keywords: *training sample, neural networks, random point processes, optimization, computational complexity.*