УДК 004.8
DOI https://doi.org/10.15589/znp2025.2(500).34

ON THE PROBLEM OF UNBALANCED DATASETS IN TRAINING ARTIFICIAL NEURAL NETWORKS

ДО ПРОБЛЕМИ НЕЗБАЛАНСОВАНИХ НАБОРІВ ДАНИХ ПРИ НАВЧАННІ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

Anatoliy Yu. Gaida
cetus@ukr.net
ORCID: 0000-0002-8217-5044
Ihor L. Mykheliev
mihelevigor@gmail.com
ORCID: 0000-0001-9579-6547

А. Ю. Гайда,
канд. техн. наук
І. Л. Михелєв,
канд. техн. наук, доцент

Admiral Makarov National University of Shipbuilding, Mykolaiv
Національний університет кораблебудування імені адмірала Макарова, м. Миколаїв

Abstract. The paper considers the problem of training artificial neural networks on unbalanced data sets, which is manifested in the low quality of estimates of samples from minority classes during the operation of a neural network trained on such data. The purpose of the paper is to identify opportunities for improving the quality of training artificial neural networks on unbalanced data sets by generating synthetic samples to reduce the unevenness of class representation. To achieve the goal, the factors affecting the quality of training were identified, a comparative analysis of popular methods for equalizing data distribution when training artificial neural networks was performed, and the advantages and disadvantages of data balancing methods were determined. The study applied the results of experiments on training artificial neural networks of different architectures on academic data sets "Fisher Irises", "White Wine", "Red Wine", a comparison of training results for samples from majority and minority classes, mathematical modeling of training quality assessment, and synthesis of a model for selecting candidate samples for propagation. Based on the results of the analysis of existing methods, an original model for equalizing the distribution of data sets for training was proposed, which takes into account the distribution of data in the classes of the dataset and between classes. The equalization is carried out by synthesizing synthetic samples in critical areas of the data distribution. Unlike existing data balancing methods, the proposed model allows to increase the efficiency of the process of equalizing the data distribution in the dataset and to reduce the influence of the majority classes on the evaluation of samples belonging to the minority classes. The practical significance of the obtained results lies in the fact that the proposed model is relatively simple to implement and allows to significantly reduce the time for pre-processing of large data sets. The paper provides an example of practical testing of the developed model in solving the problem of minimizing the vibration of the ship's hull.
Key words: artificial neural network; artificial neural network training; data set for training; unbalanced data; "Red wine" dataset.

Анотація. В роботі розглянута проблема навчання штучних нейронних мереж на незбалансованих наборах даних, що виявляється у низькій якості оцінок зразків з міноритарних класів у процесі експлуатації навченої на таких даних нейронної мережі. Мета роботи полягає у визначенні можливостей підвищення якості навчання штучних нейронних мереж на незбалансованих наборах даних шляхом генерації синтетичних зразків для зменшення нерівномірності представництва класів. Для досягнення мети визначені чинники, що впливають на якість навчання, виконано порівняльний аналіз популярних методів вирівнювання розподілу даних при навчанні штучних нейронних мереж, визначено переваги та недоліки методів балансування даних. У дослідженні застосовані результати експериментів з навчання штучних нейронних мереж різної архітектури на академічних наборах даних «Іриси Фішера», «Біле вино», «Червоне вино», виконано порівняння результатів навчання для зразків з мажоритарних і міноритарних класів, математичне моделювання оцінки якості навчання, синтез моделі відбору зразків кандидатів для розмноження. За результатами аналізу наявних методів запропонована оригінальна модель вирівнювання розподілу наборів даних для навчання, що враховує розподіл даних у класах набору даних та між класами. Вирівнювання здійснюється шляхом синтезу синтетичних зразків у критичних

областях розподілу даних. На відміну від наявних методів балансування даних запропонована модель дозволяє підвищити ефективність процесу вирівнювання розподілу даних у наборі даних і зменшити вплив мажоритарних класів на оцінку зразків, що належать міноритарним класам. Практичне значення отриманих результатів полягає у тому, що запропонована модель є порівняно простою у реалізації і дозволяє значно скоротити час на попередню обробку великих наборів даних. У роботі наведений приклад практичної апробації розробленої моделі при вирішенні задачі мінімізації вібрації корпусу судна.
Ключові слова: штучна нейронна мережа; навчання штучної нейронної мережі; набір даних для навчання; незбалансовані дані; синтез даних; набір даних «Червоне вино».

ПОСТАНОВКА ЗАДАЧІ

Останнім часом все більшої актуальності набувають питання автоматизації процесів управління складними організаційними, економічними, технічними та іншими системами. Успішне вирішення пов'язаних з цими питаннями задач потребує наукового підходу до збору, аналізу і узагальнення значних обсягів даних про об'єкти і процеси управління в таких системах, пошуку і виявлення закономірностей в їх характеристиках і реакції на управління, математичного і комп'ютерного моделювання систем та їх складових. Одним із засобів, що дозволяють здійснювати моделювання складних систем і прогнозування реакції на управління, добуванні даних і виявлення закономірностей в цих даних є засоби машинного навчання, у тому числі – штучні нейронні мережі (ШНМ). Технології управління, що базуються на ШНМ, отримали назву нейроуправління і застосовуються на практиці для управління технічними (літаки, автомобілі, військові безпілотні системи), економічними, організаційними та іншими складними системами.

Основним обчислювальним блоком ШНМ є нейрон (зазвичай персептрон Ф. Розенблата), який «зважує» вхідні сигнали і виконує нелінійне перетворення і стиснення суми зважених сигналів. Таким чином, з математичної точки зору нейрон є композицією функцій, що разом визначають залежність вихідного сигналу нейрона у від функції перетворення і нормалізації s, вектора вхідних сигналів X і вектора вагових коефіцієнтів W.

$$y = s\left(w_0 + \sum\nolimits_{i=1}^{n}\left(w_i * x_i\right)\right) \qquad (1)$$

Нелінійна функція перетворення s (функція активації) та індивідуальні для кожного нейрона вагові коефіцієнти W дозволяють синтезувати більш складні передавальні функції шляхом укладання окремих нейронів у шари, а шарів у ШНМ. Процес вибору значень вагових коефіцієнтів W називається навчанням мережі. Теорема Д. Цибенка (універсальна теорема апроксимації) визначає можливість апроксимації з будь-якою точністю будь якої неперервної функції декількох аргументів у ШНМ прямого поширення з одним прихованим шаром, нейрони якого мають будь-яку неперервну сигмоїдальну функцією перетворення *s* [4]. Для успішної апроксимації необхідна лише достатня кількість нейронів прихованого шару.

Для моделювання складних залежностей у вхідних даних, що вимагають багаторівневих узагальнень, останнім часом широко застосовують ШНМ, які налічують декілька прихованих шарів – такі мережі отримали назву глибоких штучних нейронних мереж (ГШНМ). Істотною проблемою навчання повнозв'язних ГШНМ прямого поширення із сигмоїдальною функцією активації зі стисненням є проблема, відома як «зникнення градієнту», що виявляється у завмиранні процесу навчання (відомому як «параліч» мережі). Як наслідок, розробники ГШНМ змушені частково відмовлятись від використання сигмоїдальних функцій активації, які дозволяють синтезувати складні передавальні функції ШНМ, на користь ламано-лінійних функцій (ReLU, PReLU тощо). Для цього ж використовують мережі на основі нейронів з нелінійними функціями зважування (належності), прикладом яких є мережі на основі нео-нечітких (нео-фаззі) нейронів (NFN). Недоліком таких мереж є те, що ШНМ з кусково-лінійною апроксимацію не задовольняють умовам апроксимаційної теореми Д. Цибенка, що, у свою чергу, унеможливлює апроксимацію складних функцій з прийнятною точністю. Отже, заміна сигмоїдальних функцій стиснення на ламано-лінійні дозволяє створювати ГШНМ з високим рівнем узагальнень при відносно невисокій точності апроксимації [2, 8], що в окремих випадках є неприпустимим.

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ І ПУБЛІКАЦІЙ

Одним з рішень, що дозволяють на ГШНМ отримати одночасно прийнятний рівень апроксимації при прийнятному рівні узагальнення є істотне збільшення наборів даних навчання [11]. Але просте збільшення навчальних наборів даних не завжди забезпечує підвищення якості навчання, а в окремих випадках навіть призводить до її зменшення [1, 9, 13].

Першим негативним наслідком збільшення наборів даних для навчання ГШНМ є значне збільшення часу навчання глибокої мережі на великому наборі даних. В окремих випадках це становить серйозну перешкоду застосуванню ШНМ і для її подолання розроблені різні рішення: засоби паралельних обчислень і спеціалізовані процесори, орієнтовані на матричні обчислення [5]; пакетне навчання; зниження розмірності наборів даних; видалення з навчального набору близьких за характеристиками зразків тощо [11].

Другим негативним наслідком збільшення наборів даних для навчання є необхідність добування додаткових даних, що часто пов'язане зі складностями або взагалі неможливе [7]. У більшості випадків доповнення даних не змінює розподілу даних між класами, а в окремих випадках навіть призводить до погіршення рівномірності розподілу зразків даних за класами та нерівномірного розподілу зразків у окремих класах. Так, наприклад, популярні академічні набори даних для тестування засобів штучного інтелекту «Червоне вино» та «Біле вино» є значними за обсягом і одночасно вкрай незбалансованими [9, 16]. Загалом, незбалансованість наборів даних є типовою проблемою машинного навчання [1, 15] і є характерною ознакою таких задач, як виявлення випадків шахрайства, рідкісних захворювань тощо [17].

Таким чином, якість результатів, наданих ГШНМ при вирішенні складних задач визначається не тільки внутрішніми параметрами ГШНМ (кількість шарів, кількість нейронів у шарі, вид функції передачі тощо), але, значною мірою, також і якістю вхідних даних у частині їх розподілу за класами та розподілу за окремими характеристиками в кожному окремому класі.

Для подолання проблем, пов'язаних з нерівномірним розподілом значень окремих характеристик у класі застосовують методи перетворення (трансформації) даних для вирівнювання розподілу значень у діапазоні вимірювань і ці методи теоретично і практично доволі добре опрацьовані [5, 13]. Для вирівнювання представництва окремих класів застосовують методи генерації штучних зразків на основі наявних даних (random oversampling, SMOTE, ADASYN) [6], видалення близьких за значеннями зразків з мажоритарних класів тощо.

Random oversampling (випадкова передискретизація) дозволяє вирівняти розподіл зразків за класами шляхом дублювання наявних зразків. Основною перевагою метода є простота реалізації за відсутності попередніх знань про набір даних, основним недоліком – схильність до перенавчання мережі в областях значень, наближених до зразків, що дублюються. Різновидом цього методу є random oversampling with noise (випадкова передискретизація з шумом), що має меншу схильність до перенавчання за рахунок введення шуму, щоб запобігти точному дублюванню зразків.

SMOTE (Synthetic Minority Oversampling Techniques, техніка передискретизації синтетичної меншості) забезпечує створення нових зразків на основі наявних зразків та попередніх знань про їх розподіл у класі [6]. У базовому варіанті створення нового зразка здійснюється шляхом вибору випадкового зразка з міноритарного класу і одного з його найближчих сусідів з подальшим визначенням властивостей нового зразка як проміжного між двома обраними [3]. За рахунок створення синтетичних даних метод порівняно з методом Random oversampling є менш чутливим до перенавчання, але на зразках класу, що знаходяться між декількома щільними кластерами, дає результати, близькі до попереднього методу. Зазначені недоліки дещо послаблені у модифікаціях алгоритму, що акцентовано обирають зразки на межі класів (Borderline-SMOTE, SVM-SMOTE). Істотним недоліком SMOTE та модифікацій є необхідність багаторазового вирішення доволі ресурсомісткої задачі кластеризації даних [6].

ADASYN (Adaptive Synthetic, адаптивна синтетична вибірка) забезпечує створення нових зразків на основі наявних зразків, попередніх знань про їх розподіл у класі та наближення до зразків з інших класів. Створення нових зразків здійснюється шляхом вибору випадкового зразка з міноритарного класу з ймовірністю, пропорційній наближенню до скупчення зразків з інших класів. Чим більше у зразка сусідів з інших класів і менша відстань до них, тим більша ймовірність використання такого зразка як шаблону. Після вибору зразка генерація нового зразка виконується як у SMOTE [6]. Метод дозволяє отримати гарні результати на межі класів, але є найбільш ресурсомістким серед розглянутих.

ВІДОКРЕМЛЕННЯ НЕВИРІШЕНИХ РАНІШЕ ЧАСТИН ЗАГАЛЬНОЇ ПРОБЛЕМИ

З вищезазначеного видно, що забезпечення належної якості результатів роботи ГШНМ можливе за умови належного покриття набором даних простору рішень. Оскільки більш ефективні методи доповнення даних вимагають отримання інформації про розподіл даних у класі або всіх класах (зазвичай – виконанням кластеризації) – застосування таких методів вимагає значних обчислювальних ресурсів [12], а для великих наборів даних часто є надто складною і ресурсомісткою задачею, вирішення якої може становити серйозну проблему.

МЕТА, МЕТОДИ, ПРЕДМЕТ ТА ОБ'ЄКТ ДОСЛІДЖЕННЯ

Метою дослідження є підвищення ефективності процесу підготовки даних для навчання ШНМ у частині вирівнювання розподілу зразків між класами шляхом впровадження моделі генерації синтетичних зразків на основі особливостей розподілу даних у кожному окремому класі та навчальному наборі даних у цілому.

Методами дослідження є порівняльний аналіз наявних засобів вирівнювання розподілу зразків між класами; статистичний аналіз даних; кластерний аналіз даних. Об'єкт дослідження – штучні нейронні мережі, предмет дослідження – методи вирівнювання розподілу даних між класами при навчанні ШНМ.

ОСНОВНИЙ МАТЕРІАЛ

Найбільш простою оцінкою, що дозволяє визначити прогрес у навчанні ШНМ, є помилка навчання,

яка може бути визначена як відношення суми похибок на кожному зразку до кількості зразків для кожного кроку навчання:

$$e = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} < \zeta\,, \qquad (2)$$

де n – кількість зразків у наборі даних; e_i – помилка для зразка i.

Наслідком незбалансованого набору даних $S = \{s_1, s_2, ..., s_n\}$ є таке можливе спотворення поверхні рішень, коли мінімум функції втрат (2) визначається малою похибкою у великій групі з k близько розташованих зразків мажоритарних класів $S^g \in S$, яка компенсує відносно велику похибку в околицях декількох відокремлених від більшості m зразках з міноритарних класів $S^t \in S$ (Рис. 1).

З врахуванням (2) помилку навчання можна представити як суму відносних помилок на міноритарних та мажоритарних класах:

$$e = \sqrt{\frac{1}{m}\sum_{i=1}^{m} e_{t,i}^2} + \sqrt{\frac{1}{k}\sum_{j=1}^{k} e_{g,j}^2} < \zeta \Leftrightarrow \forall e_{t,i} \gg \gg e_{g,j} \wedge m \ll k \wedge m + k = n\,, \qquad (3)$$

де m – кількість зразків з міноритарних класів; k – кількість зразків з мажоритарних класів; $e_{t,i}$ – помилка у точці i класу t з групи міноритарних класів; $e_{g,j}$ – помилка у точці j класу g з групи близько розташованих точок мажоритарних класів. У свою чергу, кожен з класів x представлений набором зразків $S^x = \{s_{x,1}, s_{x,2}, ..., s_{x,h}\}$, розподіл яких, як правило, є подібним до нормального розподілу, що призводить до недостатнього покриття граничних областей класу. Так, з Рис. 1 видно, що результати роботи мережі на штучно розбалансованому наборі даних «Іриси Фішера» можуть бути прийнятними при обчисленні значень цільової функції у точках, найближчих до точок з навчального набору даних, але в області відносно відокремлених точок (права частина графіка) відхилення змодельованої ШНМ поверхні рішень від істинної вже є значними.

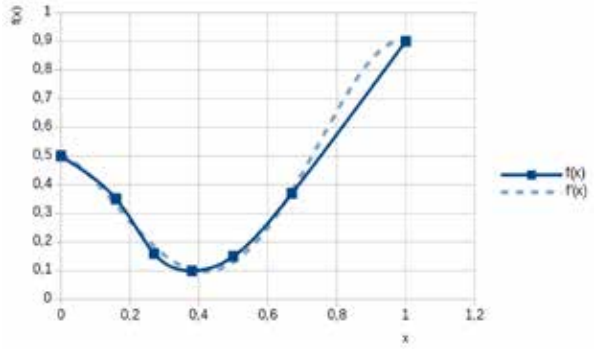

Рис. 1. Приклад перетину поверхні рішень на штучно розбалансованому наборі даних «Іриси Фішера» (*f(x)* – дійсна поверхня рішень; *f'(x)* – змодельована ШНМ)

Отже, оскільки навчання ШНМ реалізоване як процес мінімізації значення функції помилок (3), то якби ймовірність покриття навчальним набором даних певної області поверхні рішень могла б відповідати ймовірності звернення до цієї області при використанні мережі, то такий навчальний набір можна було б вважати найбільш прийнятним. Але при вирішенні задач, подібних до виявлення шахрайства та рідкісних захворювань [17], на мережі, навченій на незбалансованому наборі даних, результати виявляться зміщеними у бік мажоритарних класів і тому неприйнятними.

Таким чином, із зазначеного можна зробити висновок, що при створенні синтетичних зразків для доповнення міноритарних класів слід надавати перевагу їх належності області, що лежить на межі класів.

Для уникнення зазначеної проблеми, як було показано вище, застосовують методи балансування даних, що при їх належному застосуванні дозволяє доповнити набір даних для навчання штучно згенерованими зразками і, як наслідок, значно зменшити відхилення змодельованої ШНМ поверхні рішень від істинної (Рис. 2).

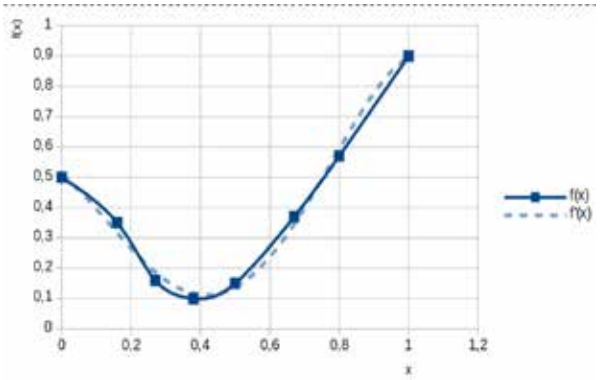

Рис. 2. Приклад перетину поверхні рішень на доповненому наборі даних «Іриси Фішера» (*f(x)* – істинна поверхня рішень; *f'(x)* – змодельована ШНМ)

З результатів аналізу популярних методів доповнення наборів даних видно, що створення штучних зразків на основі випадково обраних наявних зразків, які з відносно більшою ймовірністю будуть належати області з більшою щільністю розподілу зразків, часто не дозволяє покращити якість навчання (Рис. 3), а в окремих випадках навіть призводить до погіршення якості. При цьому з точки зору зменшення похибки мережі у всьому діапазоні значень не всі методи є однаково ефективними – більш ефективними виявляються методи, що враховують розподіл зразків у класі або також і в інших класах з метою доповнення вибірки зразками, що знаходяться на межі розділу окремих класів (Рис. 2) [6]. Але такі методи є також і найбільш витратними за часом виконання, споживаними обчислювальними ресурсами та необхідним інструментальним забезпеченням.

У відносно простому випадку (Рис. 4) зразки, що знаходяться на межі розділу класів, повинні задовольняти наступним умовам:

- знаходитись на значній відстані від центру розподілу зразків власного класу;
- знаходитись на відносно невеликій відстані від центру розподілу зразків з іншого класу або центрів розподілу з інших класів.

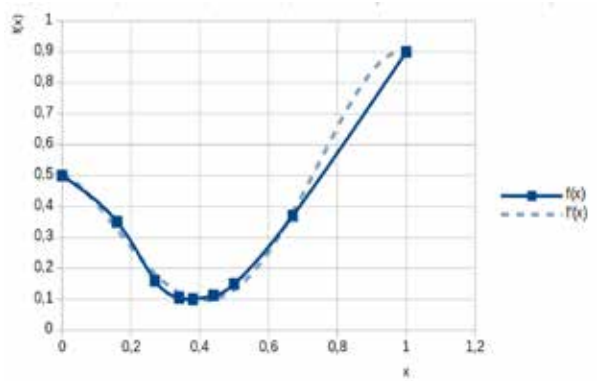

Рис. 3. Приклад перетину поверхні рішень на неефективно доповненому наборі даних «Ірриси Фішера» (*f(x)* – істинна поверхня рішень; *f'(x)* – змодельована ШНМ)

Таким чином, задачу відбору зразків $p_{i,c}$ на межі розділу класів S^c та S^x можна визначити як:

$$p_{i,c,x} = \begin{vmatrix} 1, if r_{i,c,c} > \alpha \wedge r_{i,c,x} < \beta \\ 0, otherwise \end{vmatrix}, \quad (4)$$

де $r_{i,c,c}$ – відстань зразка *i* з класу *c* до центру власного класу; $r_{i,c,x}$ – відстань зразка *i* з класу *c* до центру класу *x*; α і β – гіперпараметри.

Відібрані зразки можна розглядати як множини $S^{c,x} \in S^c$:

$$S^{c,x} = \left\{ s_{i,c} \vee \forall x \in X, p_{i,c,x} = true \right\}. \quad (5)$$

За міру відстані $r_{i,c,x}$ зручно прийняти нормалізовану евклідову відстань (відстань Махаланобіса) від точки до центру множини точок (центру кластера) у багатовимірному просторі [14]:

$$r_{i,c,x} = \sqrt{\frac{\sum_{j=1}^{n} \left(v_{i,c,j} - y_{x,j} \right)^2}{\sigma_{x,j}^2}}, \quad (6)$$

де *n* – розмірність простору, у якому задані точки; $v_{i,c,j}$ – значення властивості *j* точки і з класу *c*; $y_{x,j}$ – математичне сподівання властивості *j* класу *x*; $\sigma_{x,j}$ – середнє відхилення властивості j з класу *x*. Аналогічно можна обчислити і відстань $r_{i,c,c}$.

Оскільки результат відбору зразків залежить від значень гіперпараметрів α і β, важливо оцінити принцип, на основі якого мають встановлюватись ці значення. Якщо обирати зразки випадково, тоді ці гіперпараметри мають визначатись через середню відстань зразка від центру кластера, до якого він належить.

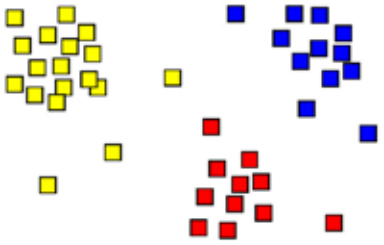

Рис. 4. Приклад відносно простого розподілу зразків за класами

Цю відстань легко отримати, оскільки вона визначається як середнє відхилення σ випадкового значення від математичного сподівання. Оскільки відстані між точками обчислені як нормалізовані значення, тоді за початкові значення цих параметрів можна прийняти такі множники до σ, що дозволить розпочати відбір найбільш віддалених від центрів множин класів кандидатів:

$$\alpha_0 = 2, \quad (7а)$$

$$\beta_0 = 2. \quad (7б)$$

Значення гіперпараметрів α і β не тільки визначають область, з якої обираються зразки, але, фактично, і кількість відібраних з цієї області зразків. І ця кількість має бути достатньою. У найбільш простому випадку множина зразків кожного класу може бути поділена за ознакою наближення до множин з інших класів на *n* кластерів (Рис. 5), тоді:

$$\left| S^{c,x} \right| \approx \frac{\left| S^c \right|}{n}, \quad (8)$$

де *n* – кількість класів у навчальній вибірці.

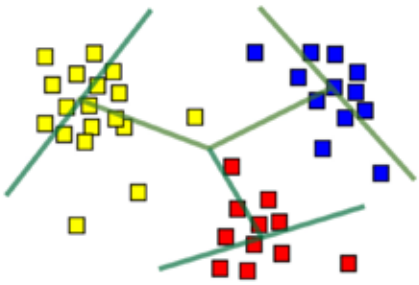

Рис. 5. Спрощений приклад поділу множин зразків у класах за наближенням до множин з інших класів

Надалі значення α і β можуть бути уточнені на кожному наступному кроці *t* відбору зразків залежно від відносної кількості обраних граничних елементів класу до кількості елементів у цьому класі:

$$\alpha = k_t \alpha_0 \quad (9а)$$

$$\beta = k_t \beta_0, \quad (9б)$$

де k_t – коефіцієнт, що визначає ширину області відбору зразків. Для його розрахунку зручно застосу-

вати експоненційне згладжування, тоді, з врахуванням (8):

$$k_t = \left| \begin{array}{c} 1.0, if\, t = 0 \\ k_{t-1} + a\left(\frac{n\left|S_t^{c,x}\right|}{\left|S^c\right|} - \frac{n\left|S_{t-1}^{c,x}\right|}{\left|S^c\right|} \right), if\, t > 1 \end{array} \right|. \quad (10)$$

Після вибору зразків на межі класів необхідно обрати зразки, на основі яких будуть згенеровані нові зразки. Зазвичай для цього використовують два зразки з одного класу: один в якості базового (*b*), інший (*a*) – для визначення напрямку зміни координат базового зразка у просторі рішень. Зважаючи на те, що кандидати для генерації синтетичного зразка мають належати одному класу *c*, вони мають належати множині кандидатів $S^{c,x}$:

$$s_a^c \in S^{c,x} \wedge s_b^c \in S^{c,x}. \quad (11)$$

Для унеможливлення генерації синтетичних зразків, близьких за властивостями базовому зразку або надто віддалених від нього бажано обирати для нього пару за умови:

$$\gamma_1 < r_{a,b,c} < \gamma_2, \quad (12)$$

де $r_{a,b,c}$ – відстань між зразками *a* і *b* класу *c*; γ_1, γ_2 – значення гіперпараметрів.

Подібно до (7) за їх початкові значення можна прийняти:

$$\gamma_1 > 0.1 \wedge \gamma_2 < 1.0. \quad (13)$$

Запропонована модель відбору може застосовуватись і для більш складних випадків розподілу зразків у класах (Рис. 6), оскільки і тут зразки, що знаходяться на межі розділу класів, повинні задовольняти тим самим умовам:

- знаходитись на значній відстані від центру розподілу зразків власного класу;
- знаходитись на відносно невеликій відстані від центру розподілу зразків з будь якого іншого класу.

РЕЗУЛЬТАТИ

Отримані результати дозволяють ефективно здійснювати вирівнювання розподілу класів у наборі даних при навчанні штучних нейронних мереж. Наукова новизна запропонованого методу порівняно з поширеними методами балансування даних полягає у можливості вирішення задачі балансування даних з істотно меншими витратами ресурсів на підготовку набору даних, що дозволяє застосовувати запропонований метод для обробки великих наборів даних, де час попередньої обробки даних може значно збільшувати час навчання мережі.

Представлені в роботі результати були застосовані в практичній задачі мінімізації вібрації корпусу судна в межах міжнародного науково-дослідного проєкту і були представлені на міжнародній конференції Товариства підводних технологій у Китаї [10]. Вирішення задачі мінімізації вібрації корпусу судна здійснювалось на нейромережевій моделі корпусу судна з встановленим головним валопроводом.

ОБГОВОРЕННЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Для апробації запропонованого методу на академічних наборах даних були використані попередньо збалансовані запропонованим методом набори даних «Біле вино», «Червоне вино». Отримані результати виявились близькими до результатів роботи з даними, доповнених методом ADASYN [3] при значно менших витратах часу на підготовку даних і кращими за метод випадкової передискретизації [16].

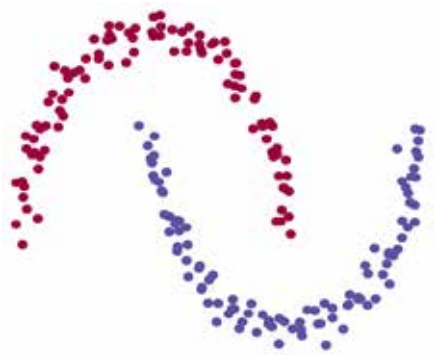

Рис. 6. Приклад відносно складного розподілу зразків за класами

Серед недоліків запропонованого оригінального методу (які також притаманні і наявним методам [3, 6]) є недостатнє врахування інформації про закон розподілу даних у кожному окремому класі. Для усунення цього недоліку можна запропонувати здійснення попередньої кластеризації всього набору даних, а не лише окремих класів, з виділенням окремих кластерів і визначенням відстані центрів отриманих кластерів до центрів кластерів, що визначаються кожним окремим класом для визначенням закону розподілу зразків на межі класів. Вирішення цього питання потребує додаткових досліджень.

ВИСНОВКИ

З наведеного вище можна зробити висновок, що є наступні основні причини невідповідності оцінок якості навчання і якості застосування навченої ШНМ:

- спотворення поверхні рішень, що виникає через складності створення мережі з «достатньою кількістю нейронів» для досягнення скільки завгодно високої точності апроксимації;
- спотворення поверхні рішень, що виникає через істотно нерівномірну щільність покриття поверхні рішень навчальною вибіркою і невідповідність розподілу зразків у навчальній вибірці розподілу запитів при застосуванні навченої ШНМ.

Для успішного подолання наслідків зазначених причин необхідно забезпечити відносно рівномірне покриття навчальною вибіркою поверхні рішень з перевагою областей на межі різних класів.

Вирішення цієї задачі можливо отримати без виконання ресурсомісткої кластеризації з результатами, близькими до кращих методів генерації синтетичних зразків.

Розроблене рішення можна рекомендувати для навчання ГШНМ, що вимагають відносно високої точності оцінок і, відповідно, великих за обсягом зразків наборів даних.

REFERENCES

[1] Ben, G, Stan, M., Aflalo, E., Madasu, A. (2025) Learning from Reasoning Failures via Synthetic Data Generation. Gabriela Ben, Melech Stan, Estelle Aflalo, Avinash Madasu and others. Retrieved from: https://arxiv.org/pdf/2504.14523 / Cornel University.

[2] Brownlee, J. (2001) Two-Dimensional (2D) Test Functions for Function Optimization. Retrieved from: https://machinelearningmastery.com/2d-test-functions-for-function-optimization/

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002) SMOTE: Synthetic Minority Over-sampling Technique / *Journal Of Artificial Intelligence Research*, Volume 16, pp. 321–357, Retrieved from: https://arxiv.org/abs/1106.1813.

[4] Cybenko, G.V. (1989) Approximation by Superpositions of a Sigmoidal function. *Mathematics of Control, Signals and Systems*, 1989, vol. 2, № 4 pp. 303–314. Retrieved from: https://web.njit.edu/~usman/courses/cs677/10.1.1.441.7873.pdf

[5] Dean, J. (2004) MapReduce: Simplified Data Processing on Large Clusters / Jeffrey Dean and Sanjay Ghemawat / OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA.

[6] Elreedy, D., Atiya, A. F., Kamalov, F. (2023) A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning / Springer, Retrieved from: https://link.springer.com/article/10.1007/s10994-022-06296-4.

[7] Fissler, T., Lorentzen, C., Mayer M. (2022) Model Comparison and Calibration Assessment: User Guide for Consistent Scoring Functions in Machine Learning and Actuarial Practice / Tobias Fissler, Christian Lorentzen, Michael Mayer / USA: Cornel University. Retrieved from: https://arxiv.org/abs/2202.12780

[8] Gaida, A. Yu., Morozova, H. S. (2024) Mekhanizm adaptyvnoho navchannia hlybokykh shtuchnykh neironnykh merezh [Adaptive learning mechanism of deep artificial neural networks]. Vcheni zapysky Tavriiskoho natsionalnoho universytetu imeni V. I. Vernadskoho. Seriia: Tekhnichni nauky. Tom 35 (74) No 3 2024 – Scientific Notes of the V. I. Vernadsky Tavrichesky National University. Series: Technical Sciences. Volume 35 (74). № 3 2024 – Kherson: S. 48–53.

[9] Gaida, A. Yu., Mykheliev, I. L., Morozova, H. S. (2023) Do problemy znykaiuchoho hradiientu pry navchanni hlybokykh shtuchnykh neironnykh merezh [Towards the vanishing gradient problem in training deep artificial neural networks]. Innovatsii v sudnobuduvanni ta okeanotekhnitsi: XIV Mizhnarodna naukovo-tekhnichna konferentsiia: materialy – Innovations in Shipbuilding and Ocean Engineering: XIV International Scientific and Technical Conference: Proceedings. – Mykolaiv: NUK.

[10] Gaida, A., Luchenkov, Ye., Orischuck, Yu. (2019) Application of Artificial Intelligence and Finite Element Method for Design of Marine Propulsion Units and Reduction of Noise and Vibration / During 8-th International SUT (China) Technical Conference – Zhejiang Ocean University / Zhoushan of Zhejiang, Chine.

[11] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J. (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies / Kremer, S. C.; Kolen, J. F. A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press. ISBN 0-7803-5369-2.

[12] Jain, A., Murty, M., Flynn, P. (1999) Data Clustering: A Review / ACM Computing Surveys. Vol. 31, № 3.

[13] Leskovec, J. (2010) Mining of Massive Datasets / Jure Leskovec Anand Rajaraman, Jeffrey David Ullman / Stanford Univ.

[14] McLachlan, G. (2004) Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons Inc. USA.

[15] Nabil, T., Agoua, G., Cauchois, P., De Moliner, A, Grossin, B. (2025) A synthetic dataset of French electric load curves with temperature conditioning / USA: Cornel University. Retrieved from: https://arxiv.org/abs/2504.14046

[16] Ramya, B., Shankar, R., Kruthika, R. (2023) A Study on Red Wine Quality Detection Using Machine Learning / International Journal of Research Publication and Reviews, Vol 4, no 12, pp 1579-1587 December 2023. Retrieved from: https://ijrpr.com/uploads/V4ISSUE12/IJRPR20259.pdf

[17] Suhas, B. N., Mattioli, D., Abdullah, S. (2025) Thousand Voices of Trauma: A Large-Scale Synthetic Dataset for Modeling Prolonged Exposure Therapy Conversations. Suhas BN, Mattioli D. Abdullah S. and others / USA: Cornel University. Retrieved from: https://arxiv.org/pdf/2504.13955

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Ben, G, Stan, M., Aflalo, E., Madasu, A. (2025) Learning from Reasoning Failures via Synthetic Data Generation. Gabriela Ben, Melech Stan, Estelle Aflalo, Avinash Madasu and others. Retrieved from: https://arxiv.org/pdf/2504.14523 / Cornel University.

[2] Brownlee, J. (2001) Two-Dimensional (2D) Test Functions for Function Optimization. Retrieved from: https://machinelearningmastery.com/2d-test-functions-for-function-optimization/

[3] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002) SMOTE: Synthetic Minority Over-sampling Technique / *Journal Of Artificial Intelligence Research*, Volume 16, pp. 321–357, Retrieved from: https://arxiv.org/abs/1106.1813.

[4] Cybenko, G.V. (1989) Approximation by Superpositions of a Sigmoidal function. Mathematics of Control, Signals and Systems, 1989, vol. 2, № 4, pp. 303–314. Retrieved from: https://web.njit.edu/~usman/courses/cs677/10.1.1.441.7873.pdf

[5] Dean, J. (2004) MapReduce: Simplified Data Processing on Large Clusters / Jeffrey Dean and Sanjay Ghemawat / OSDI'04: Sixth Symposium on Operating System Design and Implementation. San Francisco, CA.

[6] Elreedy, D., Atiya, A. F., Kamalov, F. (2023) A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning / Springer, Retrieved from: https://link.springer.com/article/10.1007/s10994-022-06296-4.

[7] Fissler, T., Lorentzen, C., Mayer M. (2022) Model Comparison and Calibration Assessment: User Guide for Consistent Scoring Functions in Machine Learning and Actuarial Practice / Tobias Fissler, Christian Lorentzen, Michael Mayer / USA: Cornel University. Retrieved from: https://arxiv.org/abs/2202.12780

[8] Гайда, А. Ю., Морозова, Г. С. (2024) Механізм адаптивного навчання глибоких штучних нейронних мереж. *Вчені записки Таврійського національного університету імені В. І. Вернадського.* Серія: Технічні науки. 2024. Том 35 (74). № 3. Херсон : С. 48–53.

[9] Гайда, А. Ю., Михелєв, І. Л., Морозова, Г. С. (2023) До проблеми зникаючого градієнту при навчанні глибоких штучних нейронних мереж. *Інновації в суднобудуванні та океанотехніці : XIV Міжнародна науково-технічна конференція: матеріали.* Миколаїв : НУК.

[10] Gaida, A., Luchenkov, Ye., Orischuck, Yu. (2019) Application of Artificial Intelligence and Finite Element Method for Design of Marine Propulsion Units and Reduction of Noise and Vibration / During 8-th International SUT (China) Technical Conference – Zhejiang Ocean University / Zhoushan of Zhejiang, Chine.

[11] Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J. (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies / Kremer, S. C.; Kolen, J. F. A Field Guide to Dynamical Recurrent Neural Networks. IEEE Press. ISBN 0-7803-5369-2.

[12] Jain, A., Murty, M., Flynn, P. (1999) Data Clustering: A Review / ACM Computing Surveys. Vol. 31, № 3.

[13] Leskovec, J. (2010) Mining of Massive Datasets / Jure Leskovec Anand Rajaraman, Jeffrey David Ullman / Stanford Univ.

[14] McLachlan, G. (2004) Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons Inc. USA.

[15] Nabil, T., Agoua, G., Cauchois, P., De Moliner, A, Grossin, B. (2025) A synthetic dataset of French electric load curves with temperature conditioning / USA: Cornel University. Retrieved from: https://arxiv.org/abs/2504.14046

[16] Ramya, B., Shankar, R., Kruthika, R. (2023) A Study on Red Wine Quality Detection Using Machine Learning / International Journal of Research Publication and Reviews, Vol 4, no 12, pp 1579-1587 December 2023. Retrieved from: https://ijrpr.com/uploads/V4ISSUE12/IJRPR20259.pdf

[17] Suhas, B. N., Mattioli, D., Abdullah, S. (2025) Thousand Voices of Trauma: A Large-Scale Synthetic Dataset for Modeling Prolonged Exposure Therapy Conversations. Suhas BN, Mattioli D. Abdullah S. and others / USA: Cornel University. Retrieved from: https://arxiv.org/pdf/2504.13955

Дата надходження статті до редакції: 23.05.2025
Дата затвердження статті до друку: 11.06.2025