УДК 004.8:004.6
DOI: https://doi.org/ 10.17721/3041-2323.2024.242-247

Наталія ОРЄХОВА, мол. наук. співроб.
ORCID ID: 0009-0000-5339-884X
e-mail: naorexova@gmail.com
Інститут кібернетики імені В. М. Глушкова
НАН України, Київ, Україна

Олександр ЛУК'ЯНОВ, канд. техн. наук
ORCID ID: 0009-0003-4704-8167
e-mail: ihorlukianov@gmail.com
Інститут кібернетики імені В. М. Глушкова
НАН України, Київ, Україна

АЛГОРИТМ ЗНАХОДЖЕННЯ ЗНАЧУЩИХ ФАКТОРІВ КОМП'ЮТЕРНОЇ МОДЕЛІ МЕТОДАМИ ГЛИБОКОГО НАВЧАННЯ

Запропоновано алгоритм на основі невеликої нейронної мережі для відбору ознак для наборів даних, у яких кількість навчальних вибірок трохи перевищує кількість атрибутів.

Ключові слова: ***моделювання, вибір ознак, нейронна мережа, щільний рівень, функція активації, кількість епох, вибірка даних, навчання, прогноз, значущість атрибутів.***

Вступ

Сучасні засоби комп'ютерної техніки надають можливість розробляти моделі високої складності. Це підштовхує розробників комп'ютерних моделей складних систем до більшої деталізації. Але сама по собі комп'ютерна модель має відносну цінність. Оскільки комп'ютерне моделювання як потужний інструмент системного аналізу завжди здійснюється з метою розв'язання конкретної проблеми, комп'ютерна модель набуває своєї повної цінності тільки разом із можливістю здійснювати на її основі комп'ютерні експерименти, а ефективність комп'ютерного моделювання залежить від ефективності планування експерименту.

Результати

Згідно з уніфікованою схемою комп'ютерного моделювання (Бігдан, Пепеляєв, & Чорний, 2006) можна виокремити такі три

життєві цикли розроблення і застосування моделі: Modeling, Simulation та Replication. На життєвому циклі Modeling здійснюється розроблення моделі, на життєвому циклі Simulation відбуваються основні комп'ютерні експерименти на основі розробленої моделі, а на життєвому циклі Replication у разі стохастичної моделі визначається найкраща альтернатива серед обраних на попередньому циклі претендентів.

Основною метою здійснення експериментів на життєвому циклі Modeling є обґрунтування адекватності розробленої моделі. Але не менш важливим для подальшого модельного дослідження, а особливо оптимізації складної системи, є визначення значущості факторів, що впливають на її функціонування. На початковому етапі розроблення моделі кількість малозначущих факторів може досягати 80 %. Таке збільшення розмірності значно ускладнює здійснення комп'ютерних експериментів, а також розуміння взаємодії важливих факторів, що визначають основу функціонування складної системи. Якщо в наявності є всього кілька зразків даних, інформаційна цінність їхніх ознак набуває вирішального значення. Тому бажано якомога раніше виявити значущі та малозначущі фактори.

Для розв'язання цієї проблеми запропоновано алгоритм на основі простої нейромережі, створеної за допомогою бібліотеки глибокого машинного навчання Keras (Chollet, 2021).Keras – відкрита нейромережева бібліотека, написана мовою Python, здатна працювати поверх TensorFlow, Microsoft Cognitive Toolkit, R, Theano. Вона призначена для проведення експериментів із мережами глибокого навчання, її основним автором і підтримувачем є Франсуа Шоллє (*François Chollet*).

Запропонована нейромережа реалізує регресію і складається з кількох послідовних повнозв'язних шарів із функцією активації **relu**, як це описано у (Keras, n. d.).Оптимальна кількість шарів та кількість нейронів у шарі, а також необхідна кількість епох навчання визначалися за допомогою експериментів. В результаті обрано архітектуру нейромережі, що складається з двох проміжних повнозв'язних шарів, кожен з яких містить 64 приховані нейрони, та закінчується одновимірним шаром, який не має функції активації.

Для експериментів використовували штучно створені набори даних. Зразок даних є масивом, що складається зі значень атрибутів X_i і відповідного їм значення цільової функції Y. Змінні X_i набувають випадкові значення, рівномірно розподілені в діапазоні [0; 1] із кроком 1/30. Значення цільової функції Y обчислюють як суму значень змінних (атрибутів) із певними коефіцієнтами.

Завдання полягало в тому, щоб за наявними даними за допомогою нейромережі визначити значущість змінних, тобто коефіцієнти, з якими вони входять до цільової функції.

Побудована нейромережа навчалася на наявному наборі даних, щоб можна було передбачати значення цільової функції для довільних значень змінних з указаного діапазону. Перебіг процесу навчання для визначення необхідної кількості епох відстежували порівнянням передбачених за допомогою частково навченої нейромережі значень цільової функції з реальними її значеннями на тестовому наборі даних перед початком кожної наступної епохи навчання.

Для візуалізації процесу навчання створювали діаграми залежності середньої абсолютної помилки (**mse**) від кількості епох. Зразок діаграми показано на рис. 1.

Навчання відбувалося тим швидше, чим більшою була кількість зразків у навчальному наборі. Визначено, що 15–20 епох – це цілком достатня кількість для отримання задовільних прогнозів.

Фіксувався також час навчання мережі. Він суттєво залежить від кількості атрибутів, кількості зразків даних у наборі, а також кількості епох навчання і варіювався від кількох секунд до двох десятків секунд.

Після закінчення навчання за допомогою навченої мережі оцінювали внесок до цільової функції кожної змінної двома способами, а саме: оцінювали, як зміниться значення цільової функції при додаванні певної змінної до порожнього набору, а також при виключенні цієї змінної з максимального набору. Отримані значення вважали оцінками значущості змінної.

Експерименти проводили для різної кількості атрибутів у зразку даних, різної кількості зразків даних у наборі, різних співвідношень значущих і малозначущих атрибутів. Результати експериментів подавали у вигляді діаграм.

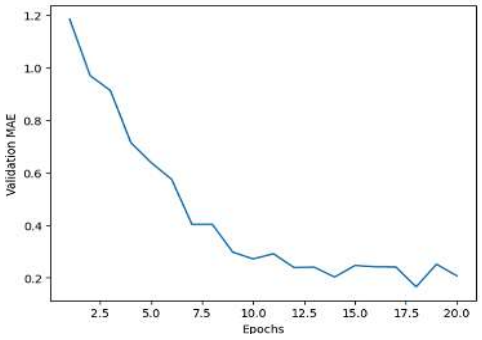
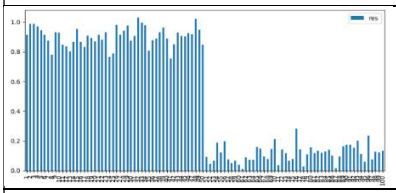
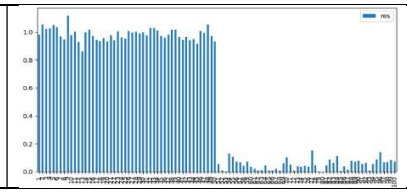
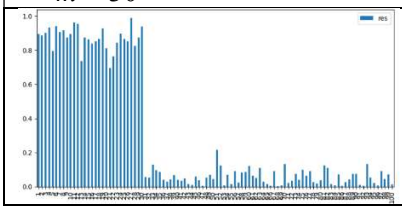
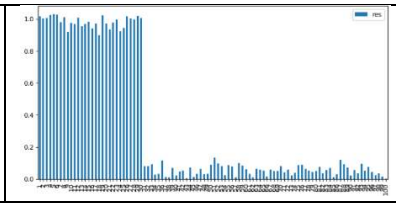


Рис. 1. *Приклад діаграми залежності середньої абсолютної помилки (mse) від кількості епох навчання*

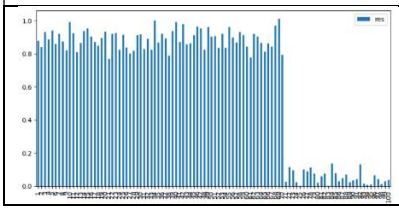
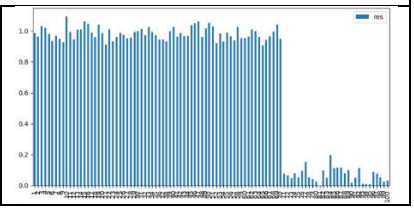
Приклади діаграм для різних співвідношень значущих і малозначущих змінних для наборів із 300 зразків, кожен з яких містить 100 атрибутів, причому перші *m* атрибутів входять до цільової функції з коефіцієнтами 1, інші – з коефіцієнтами 0.05, наведено в табл. 1. Алгоритм дозволяє впевнено виділити значущі фактори за наявності досить обмеженої кількості зразків даних у наборі.

Таблиця 1

Оцінки значущості для різних співвідношень значущих і малозначущих атрибутів

Оцінка значущості (додавання змінних)	**Оцінка значущості (виключення змінних)**
m = 50	
	
m = 30	
	

Закінчення табл. 1

Оцінка значущості (додавання змінних)	Оцінка значущості (виключення змінних)
$m = 70$	
	

Дискусія і висновки

Варто зазначити, що оцінки на різних прогонах для того самого набору дещо відрізнялися, оскільки початкові значення ваг нейромережі обирають випадковим чином. Тому в складніших випадках для більшої достовірності результатів рекомендується робити кілька прогонів.

Список використаних джерел

Бігдан, В. Б., Пепеляєв, В. А., & Чорний, Ю. М. (2006). Уніфікована схема реалізації оптимізаційно-імітаційних експериментів. *Проблеми програмування*, 2-3, 728–733.

Chollet, F. (2021). *Deep learning with Python*. Simon and Schuster.

Keras. (n. d.). *About Keras.* Retrieved May 6, 2025, from https://keras.io/about/

References

Bigdan, V. B., Pepelyaev, V. A., & Chornyi, Y. M. (2006). Unified scheme for optimization–simulation experiments. *Problems of Programming*, 2-3, 728–733 [in Ukrainian].

Chollet, F. (2021). *Deep learning with Python.* Simon and Schuster.

Keras. (n. d.). *About Keras.* Retrieved May 6, 2025, from https://keras.io/about/

Отримано редакцією журналу / Received: 17.09.24

Прорецензовано / Revised: 27.09.24

Схвалено до друку / Accepted: 01.10.24

Nataliia ORIEKHOVA, Junior Researcher
ORCID ID: 0009-0000-5339-884X
e-mail: naorexova@gmail.com
V. M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

Oleksandr LUKYANOV, PhD (Engin.)
ORCID ID: 0009-0003-4704-8167
e-mail: ihorlukianov@gmail.com
V. M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

ALGORITHM FOR FINDING SIGNIFICANT FACTORS OF A COMPUTER MODEL USING DEEP LEARNING METHODS

The article proposes an algorithm based on a small neural network for feature selection for data sets in which the number of training samples slightly exceeds the number of attributes.

Keywords: ***modeling, feature selection, neural network, dense level, activation function, number of epochs, data sample, training, prediction, attribute significance.***

Автори заявляють про відсутність конфлікту інтересів. Спонсори не брали участі в розробленні дослідження; у зборі, аналізі чи інтерпретації даних; у написанні рукопису; в рішенні про публікацію результатів.

The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.