УДК 004.032.26

О. А. Козловський[1]
Я. В. Іванчук[1]
К. В. Добровольська[2]

МЕТОДИ СТРУКТУРНОЇ ОПТИМІЗАЦІЇ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

[1]Вінницький національний технічний університет
[2]Вінницький національний медичний університет ім. М. І. Пирогова

Анотація

У роботі розглянуто проблему зменшення обчислювальної та структурної складності штучних нейронних мереж. Проаналізовано традиційні підходи до спрощення архітектури, зокрема проріджування вагових коефіцієнтів і вилучення нейронів, та визначено їхні обмеження. Обґрунтовано доцільність застосування еволюційних методів для глобальної структурної оптимізації, що забезпечує формування компактних моделей без суттєвої втрати точності. Отримані результати підтверджують перспективність використання структурної оптимізації для підвищення ефективності нейромережевих систем.

Ключові слова: штучні нейронні мережі, структурна оптимізація, обчислювальна складність, проріджування ваг, вилучення нейронів, еволюційні методи.

Abstract

The paper addresses the problem of reducing the computational and structural complexity of artificial neural networks. Traditional approaches to architecture simplification, including weight pruning and neuron removal, are analyzed, and their limitations are identified. The feasibility of applying evolutionary methods for global structural optimization is substantiated, enabling the development of compact models without significant loss of accuracy. The obtained results confirm the prospects of structural optimization for improving the efficiency of neural network systems.

Keywords: artificial neural networks, structural optimization, computational complexity, weight pruning, neuron pruning, evolutionary methods.

Вступ

У сучасних умовах розвитку інформаційних технологій та зростання обсягів даних актуальним є підвищення ефективності штучних нейронних мереж (ШНМ). Висока точність таких моделей часто супроводжується значною обчислювальною складністю, що ускладнює їх практичне застосування [1]. Важливим напрямом є спрощення архітектури ШНМ без суттєвої втрати якості. Існуючі методи, зокрема проріджування ваг і вилучення нейронів, мають обмеження, пов’язані з локальним характером оптимізації та чутливістю до параметрів. Перспективним підходом є використання еволюційних методів, які забезпечують глобальну оптимізацію структури та параметрів моделей [2].

Метою роботи є аналіз і обґрунтування підходів до зменшення складності ШНМ на основі структурної оптимізації.

Результати досліджень

При практичному застосуванні ШНМ зовсім не останню роль відіграє їхня складність. Тому під час побудови мережі прагнуть досягти її максимально простої архітектури. Існує два підходи до вирішення даної проблеми: зменшення кількості елементів вагової матриці мережі [3] і зменшення кількості використовуваних нейронів [4]. Причому ці підходи використовуються як під час навчання ШНМ, так і після (рис. 1). Зокрема, у підході зменшення кількості елементів вагової матриці мережі ефективним є метод виключення вагових коефіцієнтів із малими значеннями [5]. Ідея методу заключається в тому, що після навчання багато вагових коефіцієнтів приймають достатньо малі значення по модулю (близькими до нуля) й відповідно, що внесок таких зв’язків у вихід нейромережі незначний, а тому їхні значення можуть бути зануленні. Даний метод може бути записаний в оптимізаційні формі із регуляризацією [2–4]:

$$\tilde{\theta} = \min_{\boldsymbol{W} \in \mathbb{R}^{m \times n}} \Im(\boldsymbol{W}) + \lambda \left\| \boldsymbol{W} \right\|_0, \quad (1)$$

де λ – параметр регулювання степені проріджування (кількість «штрафних» (занулених) вагових коефіцієнтів пропорційно значенню λ); $\Im(\boldsymbol{W})$ – функція втрат, $\left\| W \right\|_0 = \sum_{i=1}^{m} \sum_{j=1}^{n} 1\{W_{ij} \neq 0\}$ – «нульова норма» (кількість ненульових елементів в матриці вагових коефіцієнтів $\boldsymbol{W}$).

Рисунок 1 – Схема взаємозв'язку методів і підходів спрощення структури ШНМ

Метод виключення вагових коефіцієнтів із малими значеннями є достатньо простим, інтерпретованим й ефективним для стиснення самої моделі ШНМ, але має недолік неоптимального вибору порогу λ, що спричиняє втраті багато важливих зв'язків, і як наслідок спричиняє падіння точності нейромережі. Загалом у процесі застосування даного метода фактично не відбувається зміни архітектури ШНМ так як топологія нейромережі залишається незмінною, а це у свою чергу не призводить до реальної економії обчислень. Також застосування даного методу інколи зменшує «надлишковість» ШНМ, яка дозволяла боротися із шумом і перенавчанням, що у свою чергу зменшує узагальнювальну здатність нейромережі [2, 5].

У другому підході зменшення кількості використовуваних нейронів використовується метод вилучення нейронів з урахуванням їх важливості [1, 4]. Ідея методу заключається у вилученні нейронів як вхідного, так і прихованих шарів і заснований на використанні у функціоналі помилки додаткового члена – показника важливості нейрона $s^{(l)}$, призначеного як різниця між помилкою всієї мережі й помилкою мережі, з якої цей нейрон вилучений [3]. Загальне правило відбору нейронів із найменшою важливістю:

$$\boldsymbol{S}_p^{(l)} = \arg\min_{\boldsymbol{S} \subseteq \{1, \ldots, n_l\}, |\boldsymbol{S}| = k} \sum_{j \in \boldsymbol{S}} \boldsymbol{s}_j^{(l)}, \quad (2)$$

де k – кількість нейронів із найменшою важливістю; n_l – кількість нейронів в l-ому шарі нейромережі; $s_j^{(l)} = \sum_{i=1}^{n_l+1} \left(W_{i,j}^{(l+1)} \right)^2$ – важливість нейрона j; k – порогове значення кількості нейронів, які необхідно видалити.

Перевагою даного метода є те, що ефективно спрощує структуру ШНМ, але він не дозволяє виконувати глобальну оптимізацію із-за евристичного походження [4]. Також проблемою являється те, що параметр важливості нейрона $s^{(l)}$ у даному методі оцінюється окремо для кожного нейрона, а реальний внесок може залежати від комбінації нейронів (колективна взаємодія). Слід зауважити, що показник важливості нейрона $s^{(l)}$ оцінюється на основі конкретної навчальної вибірки, але якщо розподіл даних навчальної вибірки змінюється, тоді нейрони із малим значенням важливості можуть виявитися корисними.

Якщо необхідно оновити архітектуру нейромережі, тоді використовується підхід спрощення архітектури ШНМ під час навчання (див. рис. 1), а саме еволюційне структурування [5]:

$$\boldsymbol{A}^{(t+1)} = mut\left(\boldsymbol{A}^{(t)} \Pi_t(f^{(t)}) B_t^{arc} \right), \quad (3)$$

де $\boldsymbol{A}^{(t)} \in \{0, 1\}^{d \times N}$ – матриця архітектур ЕНМ (кожний стовпець – індивід, кожен елемент – наявність зв'язку/нейрона); $f^{(t)} \in \mathbb{R}^N$ – функція пристосованості по винагороді; $\Pi_t(f^{(t)}) \in \{0, 1\}^{N \times k}$ – оператор відбору k найкращих індивідів; $B_t^{arc} \in \{0, 1\}^{k \times N}$ – оператор рекомбінації (змішування батьківських індивідів в нащадки); mut(·) – оператор мутації (випадкове додавання/видалення зв'язків).

Еволюційні методи навчання ШНМ мають низку важливих переваг у порівнянні з іншими методами, а саме [1]: мають властивість глобальної багатокритеріальної оптимізації, що не дозволяє зупинятись у локальних мінімумах, на відміну від градієнтних методів; відсутність градієнтних обчислень, що дозволяє працювати із недиференційованими або зашумленими функціями втрат; дозволяє оптимізувати не тільки значення вагових коефіцієнтів, але й безспосередньо саму архітектуру ШНМ; мають стійкість до зашумлених даних, що дозволяє працювати нейромережі в умовах нестабільного навчання. Наявність вищевказаних переваг із властивістю природнього паралельного обчислення дозволяє вважати еволюційні методи навчання ШНМ як універсальними так і перспективним у використанні в комбінації з іншими методами навчання [4, 5] для вирішення різного роду складних задач.

Висновки

Розглянуто проблему зменшення обчислювальної та структурної складності штучних нейронних мереж. Показано, що надлишкова складність моделей ускладнює їх практичне застосування, особливо в умовах обмежених ресурсів. Проаналізовано традиційні підходи до спрощення архітектури, зокрема проріджування вагових коефіцієнтів і вилучення нейронів, та окреслено їхні основні обмеження, пов'язані з локальним характером оптимізації і залежністю від параметрів.

Обґрунтовано доцільність використання еволюційних методів як перспективного інструменту глобальної структурної оптимізації. Застосування таких підходів дозволяє одночасно оптимізувати архітектуру та параметри моделей, забезпечуючи баланс між компактністю та точністю. Отримані результати підтверджують перспективність подальших досліджень у напрямі розробки адаптивних методів структурної оптимізації ШНМ для підвищення ефективності інтелектуальних систем.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ротштейн О. Нейронні мережі. Генетичні алгоритми. Інтелектуальні технології ідентифікації. Вінниця: універсум. 1999.
2. Іванчук Я. В., Яковчук П. Л. Аналіз технологічних рішень моніторингу роботи мережевих високонавантажених систем / Я. В. Іванчук, П. Л. Яковчук // Наука і техніка сьогодні: Серія «Техніка»: – Київ, 2025. – №9(50). – С. 1213-1222. doi.org/10.52058/2786-6025-2025-9(50)-1213-1222.
3. Глухов, В. С. Дослідження і проектування комп'ютерних систем та мереж [Текст] : навч. посіб. / В. С. Глухов, А. Т. Костик ; НУ "Львівська політехніка". – Львів : Магнолія 2006, 2025. – 253 с.
4. Іванчук Я.В., Борисюк О.О. Синтез еволюційних механізмів в розробці адаптивного алгоритму оптимізації / Науковий журнал "Проблеми програмування" // Я. В. Іванчук, О. О. Борисюк. - № 3 (2025). – С. 53-65. http://doi.org/10.15407/pp2025.03.053.
5. Методи та алгоритми комп'ютерних обчислень. Теорія і практика: підручник / Р. Н. Квєтний, Я. В. Іванчук, І. В. Богач, О. Ю. Софина, М. В. Барабан. – Вінниця : ВНТУ, 2023. – 280 с. ISBN 978-966-641-952-4.

Козловський Олексій Андрійович – аспірант, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: ak@vin.ua.

Іванчук Ярослав Володимирович – д-р техн. наук, доцент, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, Вінниця, e-mail: ivanchuck@ukr.net.

Добровольська Катерина В'ячеславівна – к.п.н., доцент, доцент кафедри біологічної фізики, медичної апаратури та інформатики Вінницького національного медичного університету ім. М. І. Пирогова, м. Вінниця, e-mail: viekurs@ukr.net.

Kozlovskyi Oleksii A. – Faculty of Automation and Intelligent Information Technology, Vinnytsia National Technical University, Vinnytsia, e-mail: ak@vin.ua.

Ivanchuk Yaroslav V. – Dr. Sc. (Eng.), Professor of the Computer Science Department, Vinnytsia National Technical University, Vinnytsia, e-mail: ivanchuck@ukr.net.

Dobrovolska Kateryna Viacheslavivna – Cand. Sc. (Ped.), Assistant Professor, Ass. Prof. of Department of Biophysics, Medical Equipment and Informatics, National Pirogov Memorial Medical University, e-mail: viekurs@ukr.net.