

Д.Г. Зеленцов, Т.М. Шаптала

МОДЕЛІ ТА МЕТОДИ НАВЧАННЯ НЕЙРОННИХ МЕРЕЖ З ДИФЕРЕНЦІЙОВАНИМИ ФУНКЦІЯМИ АКТИВАЦІЇ

Анотація. Пропонується нова постановка задачі навчання нейронних мереж з використанням розширеного вектору варійованих параметрів. Під розширеним вектором варійованих параметрів мається на увазі те, що він включає до себе окрім вагових коефіцієнтів ще і деякі параметри функцій активації. Введення додаткових варійованих параметрів не передбачає зміни архітектури нейронної мережі, але унеможливорює застосування методу зворотного поширення похибки. Задача навчання формулюється як задача неперервної багатовимірної безумовної оптимізації. Для розв'язання оптимізаційних задач були використано низку градієнтних методів. Проведено дослідження різних постановок оптимізаційних задач та методів їх розв'язування за критеріями точності та ефективності.

Ключові слова: нейронні мережі, задача навчання нейронних мереж, багатовимірна оптимізація, вектор варійованих параметрів, градієнтні методи.

Постановка проблеми. Оскільки головною ідеєю штучної нейронної мережі (ШНМ) є відтворення принципу функціонування біологічного нейрону мозку, то заздалегідь ШНМ не знає як розв'язувати поставлені перед собою задачі. Таким чином, одним із ключових етапів роботи ШНМ – це процес її навчання. Алгоритм навчання – це метод або математична модель, яка підвищує продуктивність ШНМ. Як правило, цей алгоритм застосовується багаторазово до мережі [1]. Алгоритм навчання є одним із факторів, що визначають, наскільки швидко та наскільки точно може працювати ШНМ. Залежно від процесу розробки мережі є три основні моделі машинного навчання.

Навчання з вчителем – це парадигма машинного навчання для завдань, у якій доступні дані складаються з помічених прикладів, що означає, що кожна точка даних містить функції та пов'язану мітку. Метою алгоритмів навчання з учителем є вивчення функції, яка відображає вектори ознак (вхідні дані) до міток (вихідних даних) на основі прикладів пар вхід-вихід [2].

Навчання без вчителя – на відміну від навчання з вчителем цей метод демонструє самоорганізацію, яка фіксує закономірності у вигляді щільності ймо-

вірності або комбінації переваг нейронних функцій, закодованих у вагу зв'язку між нейронами [3].

Навчання з підкріпленням – таке навчання є окремим випадком навчання з вчителем, але не вимагає подання зазначених пар вхідних/вихідних даних і не вимагає явної корекції не оптимальних дій. У цієї моделі вчителем є середовище або його модель [4].

Задача навчання нейронної мережі формулюється як задача багатовимірної безумовної оптимізації, де варійованими параметрами є вагові коефіцієнти: $\overline{W} = [w_1, w_2, \dots, w_n]^T$. Функція похибки ШНМ має вигляд [1]:

$$E(\overline{W}) = \frac{1}{2} \sum_{s=1}^N \sum_{j=1}^k (t_j^s - y_j^s(\overline{W}))^2 \rightarrow \min; \quad \overline{W} \in E^n. \quad (1)$$

Тут N – кількість навчальних зразків; k – розмірність вихідного вектору; $\overline{y}^s(\overline{W})$ – вихідний вектор для s -го навчального зразку, якій залежить від усіх вагових коефіцієнтів; $\overline{t}^s$ – еталонний вихідний вектор для s -го навчального зразку.

Для переважної більшості прикладних задач, методи розв'язування яких передбачають використання ШНМ, саме якість навчання має визначальне значення.

В задачах розрахунку довговічності та оптимального проектування конструкцій, що підлягають впливу агресивних середовищ ШНМ використовуються для управління точністю чисельного розв'язування систем диференціальних рівнянь (СДР) [5] або для обчислення поправочних функцій [6]. Низька якість навчання ШНМ ставить під сумнів доцільність розв'язання задачі в цілому. В роботі пропонується підхід до підвищення точності розв'язання ШНМ, оснований на збільшенні розмірності оптимізаційної задачі з збереженням первісної архітектури мережі.

Аналіз останніх досліджень та публікацій. Алгоритм управління точністю чисельного розв'язання СДР, які описують процес накопичення геометричних пошкоджень в конструкційних елементах, був запропонований, очевидно уперше, в [5]. ШНМ виступала як контролер, що дозволяв визначити параметр чисельного розв'язування при конкретному значенні припустимої похибки для даного вектора варійованих параметрів. Такий алгоритм використовував одразу декілька ШНМ. В подальших його модифікаціях використовувалась лише од-

на ШНМ, для якої одним з вхідних параметрів було значення припустимої похибки.

В методі поправочних функцій [6] на відміну від попереднього підходу враховувалась зміна за часом внутрішніх зусиль в конструкційних елементах, а ШНМ апроксимувала похибку наближеного розв'язку СДУ. Використання сигмоїдальних функцій активації обумовлювало використання для розв'язання задачі (1) методу зворотного поширення похибки (МЗПП).

При розв'язанні СДР чисельно-аналітичних методів для управління точністю була потрібна не відстань між вузлами часової сітки, а кількість вузлів на вісі напружень [7]. Отже, в ШНМ використовувалась дискретна функція активації вихідного нейрону, що виключало можливість застосування для її навчання градієнтних методів математичного програмування. Для розв'язання задачі у цьому випадку був задіяний генетичний метод.

Незалежно від обраного методу забезпечити потрібну якість навчання ШНМ не завжди вдавалося через проблему локальних мінімумів.

Мета дослідження полягає в пошуку нових моделей оптимізаційних задач щодо сформульованої проблеми та ефективних методів їх реалізації, які дозволили би підвищити якість навчання ШНМ, зокрема через подолання проблеми локальних мінімумів.

Викладання основного матеріалу дослідження. Метод зворотного поширення похибки передбачає корегування вагових коефіцієнтів за формулою:

$$w_{ij}^s = w_{ij}^{s-1} + \lambda \cdot \delta_j \cdot x_i, \quad (2)$$

де λ – крок навчання; δ_j – похибка нейрону.

Припустимо, що в процесі навчання нейронної мережі на $(s-1)$ -й ітерації досягнута точка локального мінімуму $\bar{w}_{lok}^*$, тобто значення функції похибки (1) у цієї точці менш, ніж у будь-якій іншій точці, що належить її околу (рис. 1):

$$\forall \bar{W} : \left| \bar{w}_{lok}^* - \bar{W} \right| \leq \varepsilon \Rightarrow E(\bar{w}_{lok}^*) < E(\bar{W}). \quad (3)$$

У цьому випадку зміна значень вагових коефіцієнтів згідно з формулою (2) приведе не до покращення функції похибки, а навпаки – до її погіршення, а необхідна точність навчання досягнута не буде.

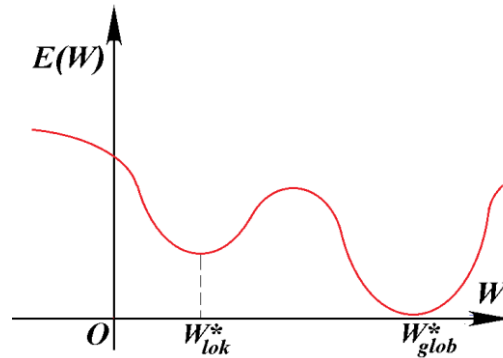


Рисунок 1 –Локальний та глобальний мінімуми функції

Для подолання таких ситуацій доцільно ввести в формулу (2) інерційний коефіцієнт η [8]:

$$w_{ij}^s = w_{ij}^{s-1} + \lambda \cdot \delta_j \cdot x_i + \eta \cdot \Delta w_{ij}^{s-1}. \quad (4)$$

Таким чином, при корегуванні вагових коефіцієнтів на s -й ітерації будуть враховані їх значення не тільки на $(s-1)$ -й, а і на $(s-2)$ -й ітераціях. Ефективність такого підходу виявлятиметься лише в околі локального екстремуму.

Авторами пропонується альтернативний підхід, який базується на збільшенні розмірності оптимізаційної задачі шляхом додання до вектору варійованих параметрів, які не впливатимуть на архітектуру ШНМ.

Розширення вектору варійованих параметрів пропонується за рахунок m параметрів функцій активації прихованих та вихідного шарів ШНМ: $\overline{W}_{new} = [w_1, w_2, \dots, w_n, \alpha_1, \alpha_2, \dots, \alpha_m]^T$. Постановка задачі навчання ШНМ з розширеним вектором варійованих параметрів як задачі багатовимірної оптимізації має вид:

$$E(\overline{W}_{new}) = \frac{1}{2} \sum_{s=1}^N \sum_{j=1}^k (t_j^s - y_j^s(\overline{W}_{new}))^2 \rightarrow \min; \quad \overline{W}_{new} \in E^{n+m}. \quad (5)$$

У цьому випадку архітектура НМ залишається незмінною, але використання МЗПП в класичному виді для розв'язання задачі навчання стає проблематичним через ускладнення аналітичного диференціювання функції похибки по параметрам функцій активації.

В даній роботі авторами досліджуються деякі градієнтні методи: метод градієнтного спуску, метод найшвидшого спуску та методи спряжених градієнтів (алгоритми Флетчера-Рівза та Полака-Ріб'єра). Навчання ШНМ відбувалось в пакетному режимі, тобто корегування варійованих параметрів здійснювалось після обробки усієї множини навчальних зразків. На відміну від МЗПП градієн-

ти функції похибки обчислювались приблизно з використанням скінченно-різницевої апроксимації диференціальних операторів. В методах найшвидшого спуску та спряжених градієнтів для визначення точки мінімуму на обраному напрямку використовувався метод парабол.

Для ілюстрації запропонованого підходу розглядається задача апроксимації функцій двох аргументів на обмеженому просторі. Дана постановка обумовлює вибір архітектури ШНМ (рис. 2). На підставі теореми Хехт-Нільсена кількість нейронів прихованого шару дорівнювала п'яти. Кількість варійованих параметрів (вагових коефіцієнтів) у постановці (1) з урахуванням нейронів зсуву дорівнювала 21.

Як функції активації прихованого та вихідного шарів використовувались логістичні функції:

$$f_i(z) = \frac{1}{1 + \exp(-\alpha_i \cdot z)}; \quad i = 1, 2. \quad (6)$$

Для постановки (1) значення параметрів функцій активації обирались $\alpha_1 = \alpha_2 = 1$. Логістична функція є асимптотичною, тобто область її значень обмежена інтервалом $f(x) \in (0; 1)$. Таким чином, параметри α_i впливають лише на довжину інтервалу ефективної зміни значень цих функцій (рис. 3, 4).

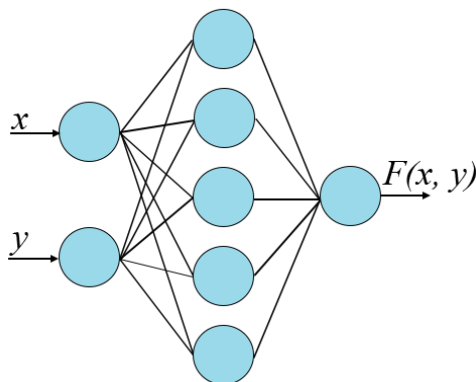


Рисунок 2 – Архітектура нейронної мережі

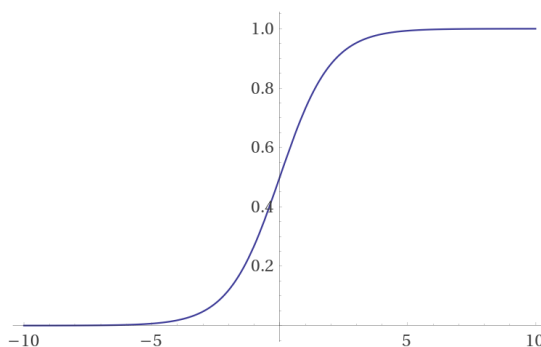


Рисунок 3 – Графік функції (6) при $\alpha = 1$

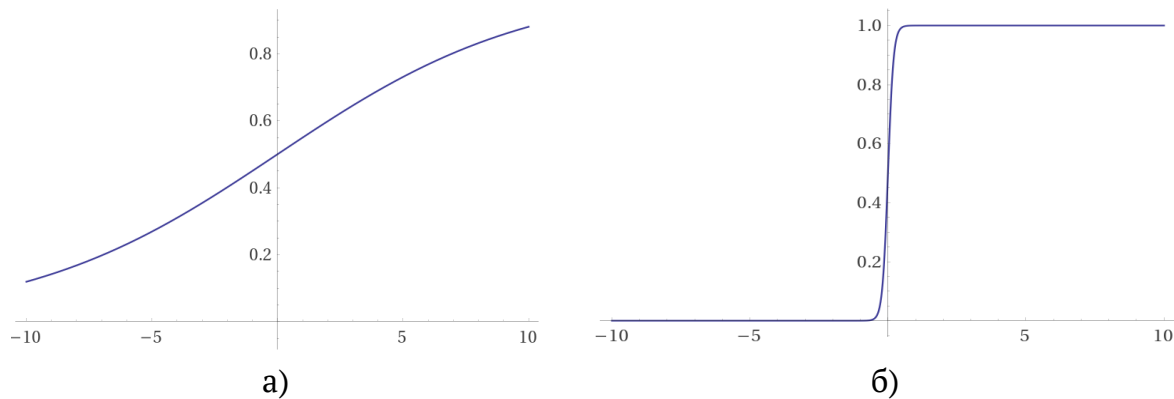


Рисунок 4 – Графік функції (6) при $\alpha = 0,2$ (а) та $\alpha = 10$ (б)

При збільшенні параметра α довжина інтервалу прагне до нуля і функція починає вести себе більш схоже на дискретну функцію Гевісайда (порогову функцію), яка може приймати лише два значення: 0 або 1 (рис. 4 а). При зменшенні цього параметру довжина інтервалу зростає, а на деякому підінтервалі в околі точки $(0; 0,5)$ функція поводить себе схоже на звичайну лінійну функцію (рис. 4 б). При зміні знака параметра графік функції стає симетричним відносно вісі ординат.

Для постановки (5) значення додаткових варійованих параметрів α_1 та α_2 обмежувалися інтервалом $\alpha_i \in [-10; 10]$. Загальна кількість варійованих параметрів досягала 23.

Першою функцією, для якої розв'язувалась задача апроксимації, була обрана функція Розенброка – неопукла функція, що використовується для тестування продуктивності алгоритмів оптимізації:

$$F_1(x, y) = (a - x)^2 + b(y - x^2)^2 \quad (7)$$

На рис. 5 зображена поверхня функції (7) при $a = b = 1$ в області $x \in [0; 2]$, $y \in [0; 3]$. Область значень функції при обраних параметрах $F_1(x, y) \in [0; 17]$.

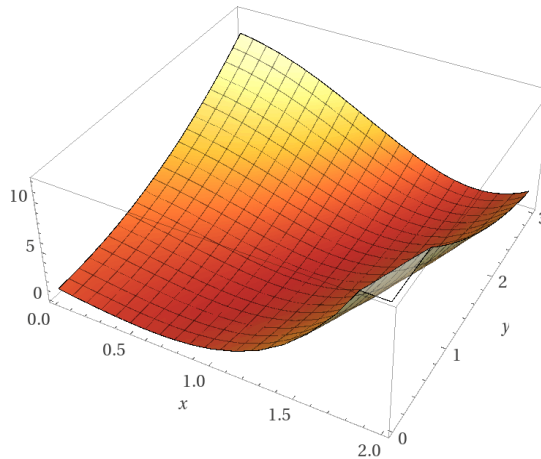


Рисунок 5 – Поверхня функції Розенброка

Останні дві функції мають вигляд:

$$F_2(x, y) = (x - 4)^2 + 2(y - 3)^2; x \in [2; 8], y \in [1; 5] \quad (8)$$

та

$$F_3(x, y) = 3xy - \sqrt{x^2 + y^2}; x \in [0; 3], y \in [1; 4]; \quad (9)$$

їх поверхні зображені на рисунках 6 та 7.

Область значень функцій: $F_2(x, y) \in [0; 24]$, $F_3(x, y) \in [-4; 31]$.

Виходячи з рекомендацій, наведених в [1, 6-8], кількість навчальних зразків – 25000; контрольних зразків – 5000; тестових зразків – 10000. Результати навчання ШНМ в постановці (1) наведені в таблицях: 1 для функції (7), 2 для функції (8) та 3 для функції (9). При проведенні обчислювальних експериментів значення припустимої похибки приймалось $\varepsilon^* = 0,0001$, а максимальна кількість епох – 3000.

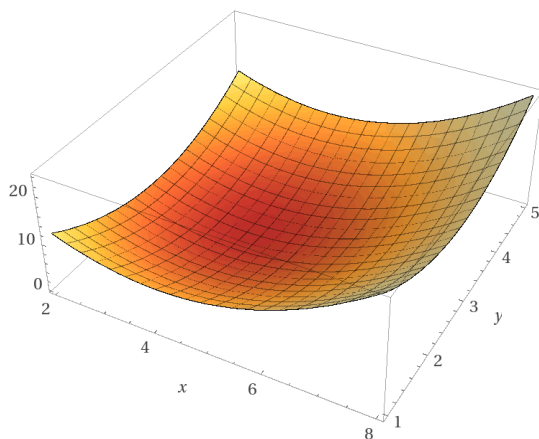


Рисунок 6 – Поверхня функції (8)

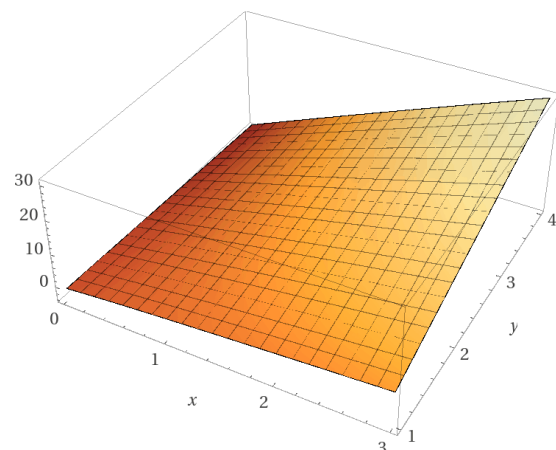


Рисунок 7 – Поверхня функції (9)

Для всіх використаних методів градієнт функції похибки в точці простору розв'язків E^n визначався чисельно, тобто функція похибки обчислювалась $(n+1)$ раз. Отже, найбільш значні обчислювальні витрати потребувала саме процедура визначення градієнту, тому як критерій ефективності приймалась загальна кількість його визначень.

Таблиця 1

Метод	Кількість обчислень градієнту функції похибки	Функція похибки
Метод градієнтного спуску	3000	0,000298
Метод найшвидшого спуску	535	0,000273
Метод Флетчера-Рівза	529	0,000124
Метод Полака-Ріб'єра	494	0,000153

На підставі даних таблиці 1 зроблено висновок про доцільність подальшого використання методів спряжених градієнтів для розв'язання задачі навчання ШНМ в різних постановках.

Для порівняння ефективності моделей (1) та (5) використовувались результати навчання ШНМ для задачі апроксимації функцій (7-9). Значення припустимої похибки приймалось $\varepsilon^* = 0,0001$; максимальна кількість епох – 10000. Для коректного зіставлення результатів розглядалися шість стартових точок, перші 21 координати яких співпадали для обох постановок. Як метод навчання використовувався метод Флетчера-Рівза так як він показав більш точний результат виходячи з отриманих результатів у таблиці 1.

Для даних задач обчислювальні витрати визначалися не тільки кількістю обчислень градієнту функції похибки, а і розмірністю оптимізаційної задачі: для одного обчислення градієнту функція похибки визначалась 22 рази для постановки (1) та 24 рази для постановки (5).

Результати обчислювального експерименту наведені таблицях: для функції (7) Розенброка з параметрами $a = b = 1$ у таблиці 2, для функції (8) у таблиці 3 та для функції (9) у таблиці 4. Кількісні показники розв'язків задачі (5) указані в дужках.

Таблиця 2

№ стартової точки	Кількість обчислень градієнту функції похибки	Кількість обчислень функції похибки
1	624 (340)	275184 (179860)
2	1571 (732)	692811 (387228)
3	839 (911)	401751 (443831)
4	1324 (488)	583884 (258152)
5	501 (565)	249165 (265029)
6	1357 (852)	598437 (450708)

Таблиця 3

№ стартової точки	Кількість обчислень градієнту функції похибки	Кількість обчислень функції похибки
1	487 (269)	214644 (142089)
2	696 (665)	333453 (352818)
3	1152 (420)	507979 (222011)
4	401 (408)	199332 (214673)
5	1099 (682)	484734 (360566)
6	1194 (549)	526536 (290421)

Таблиця 4

№ стартової точки	Кількість обчислень градієнту функції похибки	Кількість обчислень функції похибки
1	1351 (602)	595817 318144
2	684 (368)	306367 194826
3	451 (504)	224249 241176
4	835 (367)	368578 194137
5	780 (756)	373628 399924
6	549 (303)	242162 160075

Висновки. Аналіз результатів чисельних експериментів дозволив зробити висновок про доцільність розширення вектору варійованих параметрів в задачах навчання ШНМ з неперервними та диференційованими функціями активації. Незважаючи на збільшення розмірності задачі оптимізації, в цілому ефективність нової постановки значно більш висока, ніж традиційної. Пояснюється це, на думку авторів, тим, що значна частка обчислювальних витрат в традиційній постановці припадає на спроби покинути околиці локальних мінімумів,

в той час, коли збільшення розмірності простору розв'язків дозволяє здійснити це зі значно меншими витратами.

ЛІТЕРАТУРА

1. Haykin S. Neural Networks: A comprehensive foundation. Prentice Hall, 1999. – 842 p.
2. Russell S., Norvig P. Artificial Intelligence: A Modern Approach, fourth edition. London: Pearson, 2020. – 1136 p.
3. Freund Y., Hausler D. Unsupervised learning of distributions on binary vectors using two-layer networks // In Advances in Neural Information Processing Systems 4: Conference and Workshop on Neural Information Processing Systems. – 1992. – P. 912 – 919.
4. Kelbling, Leslie P., Littman, Michael L., Moore, Andrew W. Reinforcement Learning: A Survey // Journal of Artificial Intelligence Research. - Archived from the original on November 20, 2001. – P. 237 – 285.
5. Зеленцов Д.Г., Короткая Л.И. Использование нейронных сетей при решении задач долговечности корродирующих конструкций. // Вісник Кременчуцького національного університету ім. М. Остроградського. – Кременчук: КрНУ, 2011. – Вип. 3 (68), част. 1. – С. 24 – 27.
6. Zelentsov D.G., Denysiuk O.R., Korotka L.I. The Method of Correction Functions in Problems of Optimization of Corroding Structures. // Advances in Computer Science for Engineering and Education III (ICCSEEA 2020), 2020. – pp. 132 – 142.
7. Зеленцов Д.Г., Новикова Л.В., Науменко Н.Ю. Алгоритм управления точностью численного решения некоторых классов систем дифференциальных уравнений. // Системні технології. Регіональний міжвузівський збірник наукових праць. – Вип. 5 (82). – Дніпропетровськ, 2012. – С. 71 – 79.
8. Каллан Р. Основные концепции нейронных сетей. – М.: Издательский дом «Вильямс», 2001. – 287 с.

REFERENCE

1. Haykin S. Neural Networks: A comprehensive foundation. Prentice Hall, 1999. – 842 p.
2. Russell S., Norvig P. Artificial Intelligence: A Modern Approach, fourth edition. London: Pearson, 2020. – 1136 p.
3. Freund Y., Hausler D. Unsupervised learning of distributions on binary vectors using two-layer networks // In Advances in Neural Information Processing Systems 4: Conference and Workshop on Neural Information Processing Systems. – 1992. – P. 912 – 919.

4. Kelbling, Leslie P., Littman, Michael L., Moore, Andrew W. Reinforcement Learning: A Survey // Journal of Artificial Intelligence Research. - Archived from the original on November 20, 2001. – P. 237 – 285.
5. Zelentsov D.G., Korotka L.I. The use of neural networks in solving problems of the durability of corroding structures. // Bulletin of the Kremenchug National University named after M. Ostrogradskyi. – Kremenchug: KrNU, 2011. – Vol. 3 (68), part. 1. – P. 24 – 27.
6. Zelentsov D.G., Denysiuk O.R., Korotka L.I. The Method of Correction Functions in Problems of Optimization of Corroding Structures. // Advances in Computer Science for Engineering and Education III (ICCSEE 2020), 2020. – pp. 132 – 142.
7. Zelentsov D.G., Korotka L.I., Naumenko N.U. Accuracy Control Algorithm for Numerical Solution of Certain Classes of Systems of Differential Equations // System technologies. Regional interuniversity collection of scientific works. – Issue 5 (82). – Dnipropetrovsk, 2012. – P. 71 – 79.
8. Callan R. Basic concepts of neural networks. – M.: Williams Publishing House, 2001. – 287 p.

Received 01.11.2022.
Accepted 07.11.2022.

Models and methods of learning neural networks with differentiated activation functions

Analysis of the literature made it clear that the problem associated with improving the performance and acceleration of ANN learning is quite actual, as ANNs are used every day in more and more industries. The concepts of finding more profitable activation functions have been outlined a lot, but changing their behavior as a result of learning is a fresh look at the problem.

The aim of the study is to find new models of optimization tasks for the formulated problem and effective methods for their implementation, which would improve the quality of ANN training, in particular by overcoming the problem of local minima.

A studied of models and methods for training neural networks using an extended vector of varying parameters is conducted. The training problem is formulated as a continuous multidimensional unconditional optimization problem. The extended vector of varying parameters implies that it includes some parameters of activation functions in addition to weight coefficients. The introduction of additional varying parameters does not change the architecture of a neural network, but makes it impossible to use the back propagation method. A number of gradient methods have been used to solve optimization problems. Different formulations of optimization problems and methods for their solution have been investigated according to accuracy and efficiency criteria.

The analysis of the results of numerical experiments allowed us to conclude that it is expedient to expand the vector of varying parameters in the tasks of training ANNs with continuous and differentiated activation functions. Despite the increase in the dimensionality of the optimization problem, the efficiency of the new formulation is higher than the generalized one. According to the authors, this is due to the fact that a significant share of computational costs in the generalized formulation falls on attempts to leave the neighborhood of local minima, while increasing the dimensionality of the solution space allows this to be done with much lower costs.

Dmytro Zelentsov – Doctor of Technical Sciences, Professor, Head of Department of Information Systems, Ukrainian State University of Chemical Technologies

Shaptala Taras – Postgraduate Student, Department of Information Systems, Ukrainian State University of Chemical Technologies.

Зеленцов Дмитро Гегемонович – доктор технічних наук, професор, завідувач кафедри інформаційних систем, Український державний хіміко-технологічний університет.

Шаптала Тарас Максимович – аспірант кафедри інформаційних систем, Український державний хіміко-технологічний університет.