**УДК 007.5**
**DOI: 10.15587/1729-4061.2019.164591**

# Розробка методу структурної оптимізації нейронної мережі за критерієм ефективності використання ресурсів

**І. А. Луценко, О. Ю. Михайленко, О. І. Дмитрієва, О. В. Рудковський, Д. В. Мосьпан, Д. В. Кухаренко, Г. В. Коломіц, А. С. Кузьменко**

*Для вирішення задач апроксимації широко використовуються математичні моделі у вигляді штучних нейронних мереж (ШНМ). Використання цієї технології передбачає двох етапний підхід. На першому етапі визначається структура моделі ШНМ, а на другому етапі здійснюється навчання для отримання максимального наближення до еталонної моделі. Максимальне значення наближення до еталону визначається складністю архітектури ШНМ. Тобто, підвищення складності моделі ШНМ дозволяє підвищувати точність апроксимації, а, відповідно, і результату навчання. При цьому визначення структури моделі ШНМ, що здійснює апроксимацію із заданою точністю, визначається як процес оптимізації.*

*Однак підвищення складності ШНМ призводить не тільки до підвищення точності, а і до підвищення часу обчислювального процесу.*

*Таким чином, показник «задана точність» не може використовуватися в задачах визначення оптимальної архітектури нейронної мережі. Це пов'язано з тим, що результат вибору структури моделі і процесу її навчання, котрий базується на забезпеченні необхідної точності апроксимації, може зайняти неприйнятний для користувача часовий проміжок.*

*Для вирішення завдання структурної ідентифікації нейронної мережі використовується підхід, у рамках якого здійснюється визначення конфігурації моделі за критерієм ефективності. У процесі реалізації розробленого методу узгоджується часовий чинник вирішення завдання і точністю апроксимації.*

*Запропонований підхід дозволяє обґрунтувати принцип вибору структури і параметрів нейронної мережі, спираючись на максимальне значення показника ефективності використання ресурсів*

*Ключові слова: штучна нейронна мережа, оптимізація структури, апроксимація функцій, критерій ефективності*

## 1. Вступ

Розробка та використання математичних моделей, робота яких подібна принципам функціонування нервової системи людини, а саме біологічної нейронної мережі, є одним з напрямків наукових досліджень, що останнім часом найбільш динамічно розвиваються. Про це свідчить велика кількість публікацій не тільки наукового, але й науково-популярного характеру. Дані моделі застосовуються для рішення широкого кола наукових і практичних задач, серед яких можна виділити класифікацію, кластеризацію й апроксимацію функцій. Стосовно останньої вважається [1], що застосування моделі штучної нейронної мере-

жі (ШНМ) дозволяє апроксимувати функції довільної складності. Зазначена задача зазвичай зводиться до отримання математичного опису об'єктів і процесів, інформація про структуру або параметри яких неповна або повністю відсутня. Для цього можуть використовуватися й інші структури, такі як авторегресійні моделі AR, MA, ARMA, ARX, OE [2], моделі на базі фільтрів з кінцевою (FIR) і безкінечною імпульсною характеристикою (IIR) [3] або моделі на основі систем ортонормованих функцій (OBF) [4, 5] та їх нелінійні інтерпретації. Проте вища швидкодія обчислювальних процесів ШНМ, котра особливо важлива при апаратно-програмній реалізації [6], у порівнянні з іншими зумовлює широке застосування саме нейромережевих структур.

При побудові моделей на базі ШНМ проблемними питаннями є їх структурна та параметрична ідентифікація.

У загальному випадку ШНМ містить три типи шарів: вхідний, вихідний та прихований. При цьому на відміну від двох перших, прихованих шарів може бути більше одного. Кожен шар містить певну кількість вузлів (нейронів). І якщо кількість нейронів на вхідному й вихідному шарах незмінна та визначається поставленою задачею, то кількість вузлів прихованих шарів може бути довільною. Таким чином, при структурній ідентифікації можна виділити дві змінні величини – кількість прихованих шарів та кількість нейронів у них. При цьому збільшення числа шарів і нейронів призводить з одного боку до підвищення точності нейронної мережі, а з іншого до зниження швидкодії параметричної ідентифікації.

Разом з цим також слід враховувати, що обчислювальні ресурси апаратного забезпечення обмежені, а необхідна точність наближення виходу моделі до тестових даних може бути недосяжна. Отже при виборі оптимальної архітектури повинен бути забезпечений компроміс між якістю отриманої моделі та часом її навчання.

Таким чином, необхідність розробки методу визначення оптимальної конфігурації ШНМ для рішення задачі апроксимації функцій довільної складності зумовлює актуальність проведеного дослідження.

**2. Аналіз літературних даних та постановка проблеми**

На теперішній час проведений ряд досліджень, присвячених розробці методів визначення оптимальної структури ШНМ. У роботі [7] методи структурної ідентифікації розділено на три групи. Перша – це методи, які здійснюють спрощення нейронної мережі шляхом виключення її окремих елементів (вузлів та зв'язків), що суттєво не впливають на вихід моделі. Друга – методи, що здійснюють послідовне збільшення кількості числа прихованих шарів та кількості вузлів у них до досягнення необхідної точності моделі. Третя – методи, що використовують еволюційні алгоритми.

У роботі [8] наведений метод спрощення нейронної мережі за рахунок виключення зв'язків значення ваг яких нижче деякого встановленого граничного значення (magnitude-based pruning – MP). Процес оптимізації структури моделі складається з трьох етапів. Спочатку формується початкова структура мережі, що складається з великої кількості прихованих шарів зі значним числом вузлів

у кожному з них, та проводиться її навчання. Далі визначаються й усуваються зв'язки з нижчими значеннями ваг. На завершальному етапі здійснюється оцінка параметрів спрощеної нейронної структури. Процес ітеративно повторюється до моменту досягнення допустимого рівня точності моделі. Незважаючи на простоту реалізації автори не визначають на якому рівні необхідно встановлювати поріг, що вказує на необхідність виключення зв'язку. Також не наведена точність до якої доцільно здійснювати визначення архітектури мережі. Запропонований метод передбачає здійснення повного циклу навчання після кожної зміни структури, що призводить до високого обчислювального навантаження.

Основною особливістю підходу, котрий базується на виключенні ліній зв'язку з нижчими значеннями ваг, є суттєве зростання похибки моделі на кожній ітерації.

У роботі [9] розглядається метод Optimal Brain Damage (OBD), котрий передбачає визначення тих ліній зв'язку та їх кількості, виключення яких не призведе до значного підвищення загальної похибки нейронної мережі. Для цього розраховується матриця Гессе, елементами якої виступають другі похідні похибки мережі за параметрами $w_{ij}^{(k)}$. Враховуючи, що обчислювальне навантаження при розрахунку такої матриці дуже суттєве, тому автори запропонували спростити її до діагональної форми, що призвело до зниження якості методу. Параметри, а отже і лінії зв'язку, що відповідають елементам матриці з низькими значеннями другої похідної, виключаються з мережі. Процес оптимізації архітектури відбувається до моменту досягнення допустимого рівня похибки моделі. OBD-метод як і MP-метод реалізується ітеративно, проте дозволяє видалити зв'язки, відсутність яких суттєво не позначиться на точності моделі. Незважаючи на це, необхідність після кожного виключення зв'язків проводити повний цикл навчання мережі та розрахунок матриці Гессе, призводить до зниження швидкості структурної ідентифікації.

Ключовою особливістю процедури OBD є визначення Гессіану після збіжності процесу параметричної ідентифікації, що суттєво впливає на час визначення структури нейронної мережі. У дослідженні [10] запропоновано визначати значущість зв'язків до моменту досягнення локального мінімуму функції похибки при прямому проходженні нейронної мережі – метод Early Brain Damage (EBD). Авторами представлений критерій значущості EBD, який являє собою другу похідну від різниці функцій похибки для значення ваги при збіжності алгоритму навчання та для нульового значення ваги. Проте у роботі не встановлене достатнє число ітерацій навчання мережі при якій розрахунок критерію дозволить правильно оцінити значущість параметрів її ліній зв'язку. Відмінність процедури оптимізації архітектури полягає також у виключенні зі структури половини зв'язків з нижчими значеннями критерію EBD. Проте доцільність виключення саме такої кількості зв'язків авторами не пояснюється.

Метод Optimal Brain Surgeon (OBS) [11] розвиває принципи методу OBD. Він також використовує матрицю Гессе для оцінки значущості ваги зв'язку. У якості критерію використовується відношення квадрату значення ваги до подвійного значення діагонального елементу оберненого Гессіану, що їй відпові-

дає. Виключенню підлягають ваги з найменшим значенням критерію. Перевага методу полягає у тому, що він вимагає лише одного циклу прямого проходження нейронної мережі до збіжності. Далі розраховується критерій значущості. Після виключення окремого зв'язку його вага та діагональний елемент оберненої матриці Гессе, що їй відповідає, використовується для обчислення оновлених значень ваг, котрі залишилися. У результаті при оцінці значущості не спрощується матриця Гессе, що підвищує якість оцінки. Особливістю методу є велике обчислювальне навантаження, що зумовлене необхідністю розрахунку оберненого Гессіану.

Іншим шляхом оптимізації архітектури є виключення не зв'язків, а нейронів (вузлів). Такий підхід дозволяє значно спростити нейронну мережу, враховуючи, що усунення одного вузла у прихованому шарі призводить до видалення всіх вхідних і вихідних зв'язків, що з ним пов'язані.

У роботі [12] запропонований метод NoiseOut, який дозволяє об'єднувати між собою нейрони з високим рівнем корелювання активацій. Для визначення такої пари нейронів на вихідні значення тестової вибірки накладається адитивна завада. Ідентифікація структури здійснюється у процесі навчання моделі. Проте автори не визначають ступінь кореляції між двома нейронами, коли один з них може бути виключений. Наводиться лише ідеальний випадок, коли кореляція складає одиницю, що у реальних процесах, особливо при додаванні шумів випадкового характеру, не можливий. Також не встановлено, який з двох вузлів підлягає видаленню та як виключення певних зв'язків вплине на процес визначення значень ваг зв'язків, котрі залишилися, враховуючи, що форма функцій похибок вихідного та прихованих шарів може кардинально змінитися. Визначення структури ШНМ завершується при зниженні точності мережі нижче встановленого граничного значення, проте яким воно повинно бути не пояснюється.

У роботі [13] розроблено метод виключення вузлів, що базується на оцінці значущості нейрона за трьома критеріями. Перший оцінює вузол спираючись на функцію ентропії його важливості, яка залежить від кількості елементів тестової вибірки, які призвели до активації або деактивації нейрона. Під активацією вузла автори розуміють встановлення на його виході значення більше нуля при застосуванні сигмоїдної активаційної функції. Два інші критерії являють собою середні значення ваг вхідних та вихідних зв'язків вузла. Оцінка нейронів відбувається після завершення процесу навчання з наступним виключенням вузлів з нижчими величинами критерію значущості. Разом з вузлом видаляються всі його вхідні та вихідні зв'язки, що призводить, як вже було зазначено, до значного погіршення точності нейронної моделі. Для підвищення якості автори пропонують повторно виконувати процедуру оцінки параметрів. Особливостями методу є необхідність виконання циклів навчання до та після видалення нейрона, оцінку значущості вузла окремо за одним з трьох запропонованих критеріїв, а не у комплексі, невизначеність граничного рівня, при якому нейрон вважається активованим. Також авторами не наводиться конкретна умова видалення нейрона.

Робота [14] розвиває ідеї, запропоновані у дослідженні [13]. Автори вводять комплексну функцію оцінки значущості нейрона, яка поєднує функцію ен-

тропії важливості вузла та функції ентропії важливості вхідних та вихідних зв'язків цього нейрона, що вводяться замість середніх значень відповідних ваг входу-виходу. На базі сигмоїдної функції, для якої у якості аргументу використано розроблений комплексний критерій, визначено області перетину окремих ентропій, при яких доцільно видаляти вузол прихованого шару. Метод, як і у попередньому випадку, перед початком процедури визначення структури та після видалення окремого вузла передбачає проведення навчання до збіжності алгоритму. Також автори не обґрунтовують граничний рівень активації нейрона при визначенні функції ентропії його важливості. Запропонований комплексний критерій значущості не враховує час навчання мережі.

У роботі [15] запропонований метод побудови нейронної мережі, що передбачає визначення прихованого шару в який може бути доданий вузол. При цьому нейрон може бути введений в існуючий або новостворений шар без урахування кількості вузлів, що у них вже знаходяться. Кількість нейронів у окремих шарах може не співпадати. Розроблений критерій комплексний і містить два вирази: різницю похибок нейронної мережі попередньої і поточної епох навчання й абсолютне значення середньої різниці між вихідними значеннями двох попередньо доданих нейронів для кожного відліку тестової вибірки. Для кожної умови встановлені граничні значення. У залежності від того, яке з цих значень досягнуто, визначається місце додання нового нейрону в структурі нейронної мережі.

При застосуванні розглянутого методу навчання моделі відбувається у два етапи. Спочатку ваги ліній зв'язку для доданого нейрона ініціалізуються нульовими значеннями, а вже існуючих – випадковими величинами. Потім здійснюється наближення значень параметрів цього вузла до оптимальних за алгоритмом зворотного розповсюдження похибки. Процес визначення параметрів зупиняється при досягненні граничного значення першого виразу критерію, який розглянутий вище. Вважається, що при цьому досягається локальний мінімум функції похибки. На завершальному етапі на значення ваг з'єднань існуючих вузлів накладається адитивний шум Гауса з нульовим середнім значенням і одиничним коефіцієнтом варіації. Після цієї операції здійснюється оцінка значень ваг методом зворотного розповсюдження похибки. До особливостей методу слід віднести необхідність проведення перенавчання моделі після додання кожного нового вузла, недостатню обґрунтованість виразів критерію й умов визначення місця додавання вузла у структурі мережі, а також граничних рівнів цих виразів.

Авторами [16] розроблений генетичний алгоритм для визначення оптимальної архітектури нейронної мережі з одним прихованим шаром. Структура моделі при цьому представляється бінарним рядком біти якого розділено на три групи. Перша – це біти, що визначають межі зміни значень ваг при ініціалізації та проведенні процесу навчання. Друга – біти, котрі встановлюють кількість входів мережі, що використовуються при навчанні. Третя – біти, що визначають кількість вузлів у прихованому шарі. З набору випадкових бінарних рядків формується початкова популяція. Далі методом спряжених градієнтів здійснюється навчання кожної нейронної мережі з популяції до моменту досягнення

мінімуму середньоквадратичної похибки. На базі цієї функції розроблена цільова функція пристосованості, яка використовується для відбору рядків, що підлягають репродукції (селекція) і наступному схрещуванню. Процес схрещування реалізується шляхом отримання пари нащадків за рахунок обміну частинами бінарних рядків у парі батьків. При операції мутації випадковим чином визначається кількість бітів, що входять до частин, якими обмінюються пари батьків. Відбір двох екземплярів, котрі пройшли початкову селекцію, для схрещування також здійснюється випадково. Далі проводиться навчання екземплярів нової популяції, з використанням функції пристосованості робиться селекція та повторюються процедури схрещування та мутації. Визначення структури завершується коли всі екземпляри у популяції сходяться до однієї архітектури. Особливістю такого підходу є велике обчислювальне навантаження зумовлене необхідністю проведення процесу навчання для кожного екземпляру нейронної мережі у популяціях при здійсненні селекції на кожній ітерації алгоритму. Запропонований варіант бінарного кодування структури дозволяє визначити кількість вузлів лише в одному прихованому шарі, а розроблена функція пристосованості не враховує час оцінки параметрів моделі. Також алгоритм орієнтований на визначення необхідної кількості вузлів без можливості оптимізації з'єднань між нейронами окремих шарів шляхом усунення ліній зв'язку з низьким рівнем значущості. Стверджується, що застосування розробленого методу дозволить отримати глобально оптимальну архітектури мережі.

У роботі [17] розглядається генетичний алгоритм, котрий дозволяє оптимізувати структуру внутрішніх з'єднань нейронної мережі. У цьому випадку бінарний рядок складається з кодів окремих ліній зв'язку. Вага кожної лінії описується чотирибітним числом. При нульовому значенні вважається, що з'єднання між нейронами відсутнє. Під час визначення архітектури зв'язки можуть як усуватися зі структури моделі, так і створюватися або поновлюватися. Безпосередня реалізація етапів алгоритму приведена у статті у закритому вигляді. До недоліків методу слід віднести складність бінарного рядка. Так для кодування нейронної мережі, що включає три шари з двома нейронами на вхідному та прихованому шарах та одним на вихідному, і шістьома лініями зв'язку використовується двадцятичотирибітний рядок. Також при такому кодуванні значення ваг представляються тільки цілими позитивними числами, що ускладнює пошук мінімуму функції похибки при навчанні, та у окремих випадках зовсім не забезпечує збіжність алгоритму оцінки параметрів моделі. Процедура структурної ідентифікації не передбачає визначення оптимального числа вузлів. Як і у методі, запропонованому в роботі [16], для здійснення процесу селекції на всіх ітераціях алгоритму необхідно виконувати навчання кожного екземпляру нейронної мережі у початковій та нових популяціях.

У роботі [18] автори поєднали всі три розглянуті вище підходи для визначення структури нейронної мережі з декількома прихованими шарами. Зокрема, метод кодування архітектури моделі для наступної оптимізації з використанням еволюційного алгоритму, який на відміну від розглянутого у [17] передбачає представлення ваг ліній зв'язку не бінарними, а дійсними числами. При цьому екземпляр нейронної мережі також представляється рядком. З'єднання з нульо-

вим значенням ваги вважається виключеним зі складу моделі. Процес структурної ідентифікації здійснюється наступним чином. На початку формується початкова популяція з визначеної кількості елементів. Базова архітектура складається з одного нейрону в прихованому шарі та з однією лінією зв'язку між цим вузлом і одним нейроном вхідного шару, який вибирається випадковим чином. Потім за методом зворотного розповсюдження похибки здійснюється навчання нейронних мереж, що відповідають елементам початкової популяції, протягом фіксованої кількості епох. З використанням функції пристосованості вибирається пара елементів для схрещування. У якості функції пристосованості прийнята середньоквадратична похибка мережі. Процес схрещування полягає у об'єднанні структур двох мереж у одну загальну. Наприклад, якщо початкові мережі містять: перша – один прихований нейрон і три зв'язки, друга – два нейрони та п'ять зв'язків, то мережа-нащадок буде складатися з трьох нейронів та восьми ліній зв'язку. Після схрещування здійснюється мутація популяції нащадків шляхом додавання в кожній моделі одного з'єднання, що обирається випадково. Після цього проводять навчання отриманих мереж. На наступному етапі здійснюється оцінка рівня значущості нейронів прихованого шару в отриманих у результаті мутацій структурах моделей. Для цього застосовують критерій, котрий розраховується як корінь квадратний з модуля значення ваги лінії зв'язку між окремими вузлами прихованого та вихідного шарів. Вузол прихованого шару з найнижчим значенням видаляється з мережі, а інші розділяються на дві групи з більш високими та меншими значеннями відповідно. Для кожного нейрона у останній групі генерується випадкове число з рівномірним розподілом ймовірностей. Якщо воно менше 0,5, то такий вузол також видаляється.

На останньому етапі з отриманих структур з використанням коефіцієнту виживання вибираються найбільш придатні для подальшого схрещування екземпляри моделей і починається наступна ітерація алгоритму. Процес оптимізації архітектури завершується вибором найкращої за критерієм середньоквадратичної похибки моделі після проходження фіксованої кількості ітерацій (поколінь) еволюційного алгоритму. При цьому, процес спрощення структури нейронної мережі не обґрунтовується. Зокрема, не визначена умова видалення нейрона за критерієм значущості, не встановлене граничне значення для розділення вузлів на групи та не зазначено навіщо необхідно проводити таку диференціацію. Також залишається незрозумілим ймовірнісний підхід до видалення нейронів з групи меншої значущості, так як у результаті цієї операції можуть бути виключені з мережі вузли, що мають вищі значення критерію ніж ті, що залишилися.

Особливо слід відзначити обмеження кількості епох при навчанні мережі та кількості ітерацій еволюційного алгоритму. При підвищенні розміру мережі зростає її обчислювальна складність. Тому можливий випадок, коли процес оцінки параметрів моделі з незначною кількістю вузлів продемонструє швидку збіжність і кращу точність за фіксовану кількість епох ніж модель з більшим числом вузлів. У результаті для подальшого схрещування вибирається не якісніша модель з вихідної сукупності. Обмеження кількості ітерацій еволюційного алгоритму може забезпечити отримання локально, а не глобально оптимальної структури мережі.

Таким чином, аналіз проведених досліджень продемонстрував відсутність методу оптимізації архітектури ШНМ з використанням верифікованого критерію, котрий пов’язував би точність отриманої структури з часом оцінки параметрів моделі.

**3. Мета та задачі дослідження**

Мета роботи полягає у розробці методу визначення оптимальної структури ШНМ з використанням вартісного підходу, що базується на співставленні складності конфігурації структури нейронної мережі, часу її навчання та точності отриманої моделі.

Для досягнення мети було поставлено наступні задачі:

– визначити прогностичні оцінки точності ШНМ, часу її навчання та складності архітектури моделі;

– визначити підхід до формування величин експертних оцінок для вхідних та вихідних інформаційних продуктів ШНМ;

– розробити метод визначення оптимальної структури ШНМ з використанням верифікованого показника ефективності використання ресурсів.

**4. Розробка та дослідження методу структурної оптимізації штучних нейронних мереж**

**4. 1. Сутність методу**

Проведений огляд показав, що в даний час існують різні методи визначення структури ШНМ і методи навчання.

Етап синтезу структури ШНМ, особливо її внутрішніх шарів, слабо пов'язаний з особливостями функціонування досліджуваного об'єкта і здійснюється без належного теоретичного обґрунтування, часто методом проб і помилок. Тому кожна зміна структури ШНМ вимагає додаткового обґрунтування позитивного впливу запропонованих змін структури ШНМ. Таке обґрунтування здійснюється шляхом виконання чисельного експерименту, пов'язаного з навчанням ШНМ і подальшої оцінки прийнятих показників якості навчання. Серед таких в даний час розглядається тривалість навчання для отримання деякої точності роботи моделі на тестовій вибірці.

Таким чином, в даний час структурна оптимізація ШНМ є процесом пошукової оптимізації, тобто предметом спеціальних досліджень, і може займати надзвичайно тривалий час.

Ітераційний процес такої структурної оптимізації НС представлений на рис. 1.

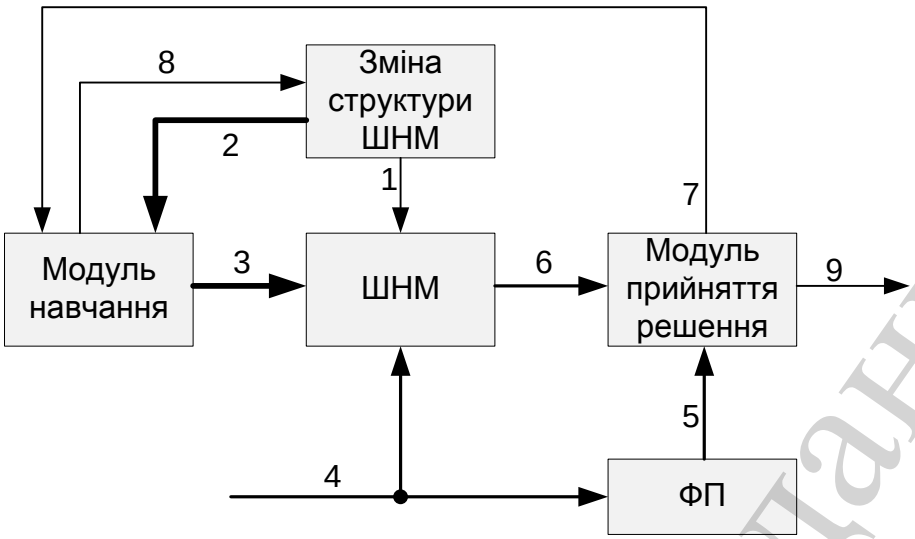


Рис. 1 Структурна схема існуючого методу визначення структури ШНМ: 1 – канал подачі нової структури ШНМ; 2 – ваги і коефіцієнти нелінійності; 3 – значення ваг і коефіцієнтів нелінійності; 4 – подача вектора тестових змінних; 5 – вектор еталонних значень; 6 – вектор відгуку ШНМ; 7 – канал передачі сигналу «рішення не знайдено»; 8 – канал подачі сигналу про необхідність зміни структури ШНМ; 9 – канал передачі сигналу «рішення знайдено»; ФП – функціональний перетворювач

Вибір структури ШНМ зумовлює кількість ваг і коефіцієнтів нелінійності, значення яких необхідно визначити. Цим і займається модуль навчання.

На виході модуля навчання формуються значення змінних структури, які встановлюються в ШНМ.

Після цього на вхід ШНМ і досліджуваного функціонального перетворювача (ФП) подаються змінні, а на виході ШНМ і ФП з'являються результат роботи ШНМ і еталон.

У блоці порівняння здійснюється прийняття рішення про досягнення необхідної точності. Якщо в ході циклу навчання задана точність не досягнута, змінюється структура ШНМ і процес повторюється. При цьому підхід до вибору відповідної структури ШНМ є ітераційним.

Аналіз показує, що такий процес не може привести до вибору оптимальної структури ШНМ. Не може в тому сенсі, що оптимальна структура – це найкраща структура, виходячи з визначення критерію оптимізації. Однак критерій «точність» є одним з показників процесу, але не критерієм найкращого рішення. Точність і далі можна підвищувати.

З іншого боку, в процесі руху до заданої точності може виявитися, що для її досягнення необхідна неприпустима тривалість обчислювального процесу. Це означає, що вимоги до точності необхідно знизити. Також це означає, що показник «точність» є не єдиним параметром, на який орієнтуються в процесі прийняття рішення.

Таким чином, ітераційний процес структурної оптимізації є неформалізованим, оскільки спирається на суб'єктивний підхід до прийняття рішення. Це

пов'язано з тим, що дослідники не використовують в явному вигляді всю необхідну інформацію для прийняття рішення.

Суть пропонованого методу ґрунтується на явному використанні, в тому числі, показника «час операції» для прийняття рішення (рис. 2).

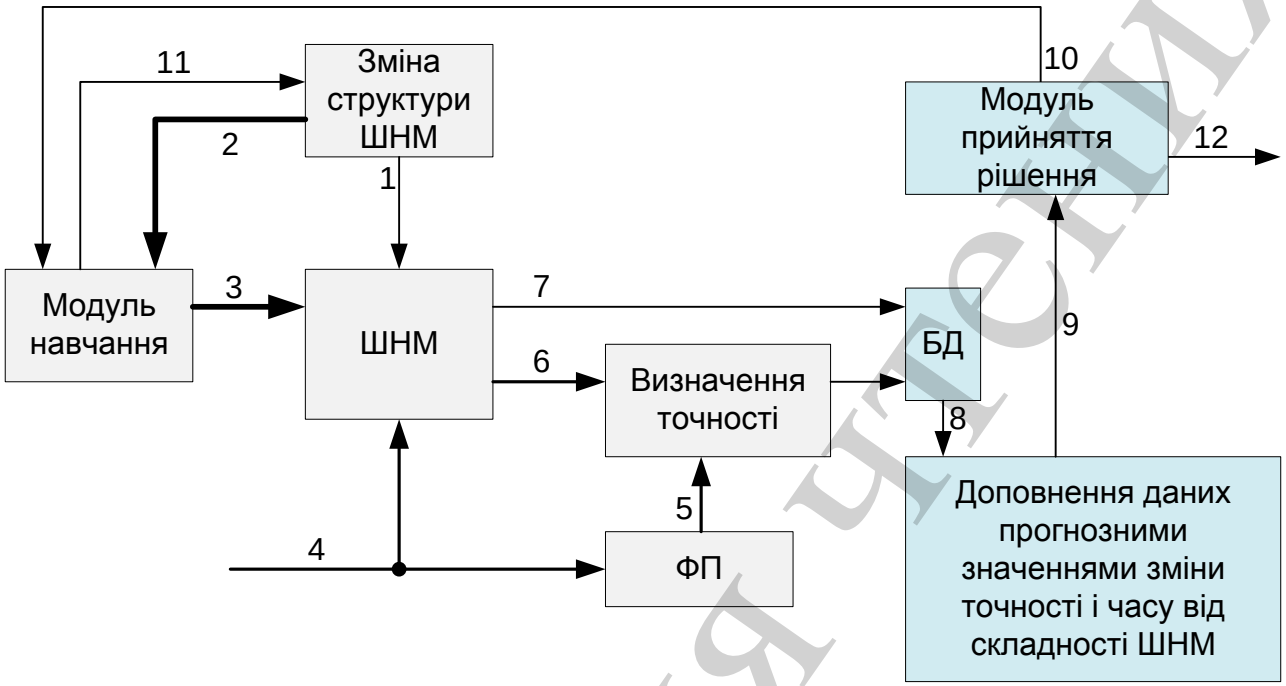


Рис. 2 Структурна схема пропонованого методу визначення структури ШНМ: 1 – канал подачі нової структури ШНМ; 2 – ваги і коефіцієнти нелінійності; 3 – значення ваг і коефіцієнтів нелінійності; 4 – подача вектора тестових змінних; 5 – вектор еталонних значень; 6 – вектор відгуку ШНМ; 7 – час операційного процесу; 8 – передача пакета експериментальних даних; 9 – дані дослідження з урахуванням прогнозних значень; 10 – канал передачі сигналу «рішення не знайдено»; 11 – канал подачі сигналу про необхідність зміни структури ШНМ; 12 – канал передачі сигналу «рішення знайдено»; ФС – функціональний перетворювач; БД – база даних

Крім того, немає необхідності у використанні ітераційного підходу при визначенні структури НС близькою до оптимальної. Це пов'язано з тим, що підвищення складності мережі призводить до прогнозованого підвищення точності і зростання часу розрахунку (рис. 3).

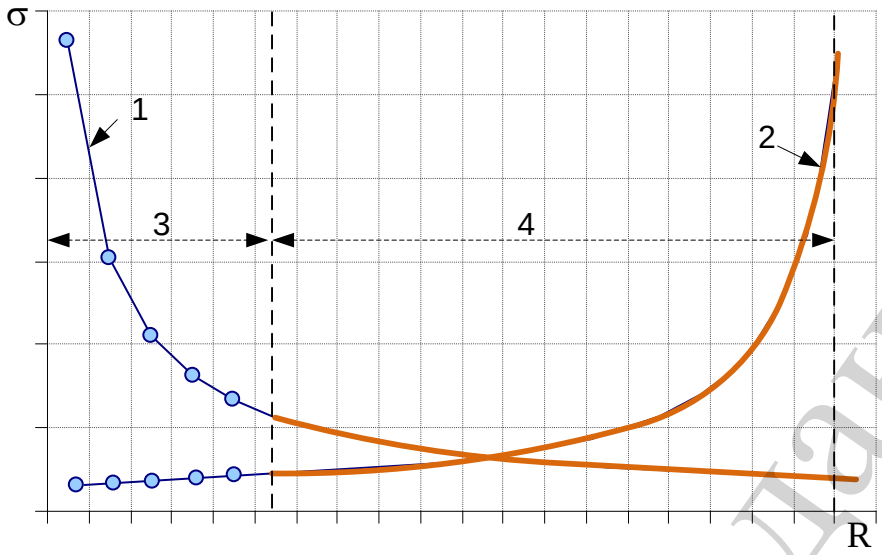


Рис. 3. Експериментальні дані та дані екстраполяції показників «точність» і «час розрахунку» від складності ШНМ: 1 – точність розрахунку; 2 – час розрахунку; 3 – експериментальні дані; 4 – прогностична оцінка зміни точності і часу розрахунку

Пропонований метод (рис. 2) ґрунтується на зборі експериментальних даних в процесі послідовного підвищення складності ШНМ, використанні методу технічного прогнозування та критерію оптимізації, який дозволяє комплексно оцінювати співвідношення складності, точності і часу розрахунку.

**4.2 Реалізація методу ідентифікації ШНМ**

Процес побудови та використання штучної нейронної мережі (ШНМ) вимагає використання обчислювальних ресурсів апаратного забезпечення. При цьому існує функціональна залежність між часом параметричної ідентифікації і розрахунку вихідного значення моделі, що можна визначити як процес навчання нейромережевої структури, та її якістю.

Підвищення точності моделі ШНМ вимагає збільшення кількості прихованих шарів і числа вузлів у них, а також збільшення часу оцінки параметрів. Таким чином, підвищення цінності отриманого результату супроводжується зростання складності структури ШНМ та збільшенням обчислювального навантаження.

Виникає задача співставлення між собою експертної оцінки ресурсів необхідних для проведення навчання ШНМ (RE), часу навчання (TO) і експертної оцінки отриманого результату (PE).

У цьому випадку визначення найкращої архітектури та параметрів ШНМ зводиться до задачі оптимізації за критерієм максимально ефективного використання ресурсів $E=f(\mathrm{RE}, \mathrm{PR}, \mathrm{TO})$.

У свою чергу наукова задача полягає у визначенні значень складових критерію (RE, TO, PE) для забезпечення можливості виконання порівняльної оцінки різних варіантів архітектури ШНМ.

З цією метою, у рамках проведеного дослідження, здійснювалася якісна оцінка ШНМ, у формі моделі багатошарового персептрона з одним прихованим шаром. У якості еталонної функції, що застосовувалася для апроксимації, використовувалась нелінійна функція виду $y=1/x$. При цьому на кожному етапі дослідження зростала складність нейромережевої структури за рахунок збільшення числа нейронів прихованого шару (рис. 4).

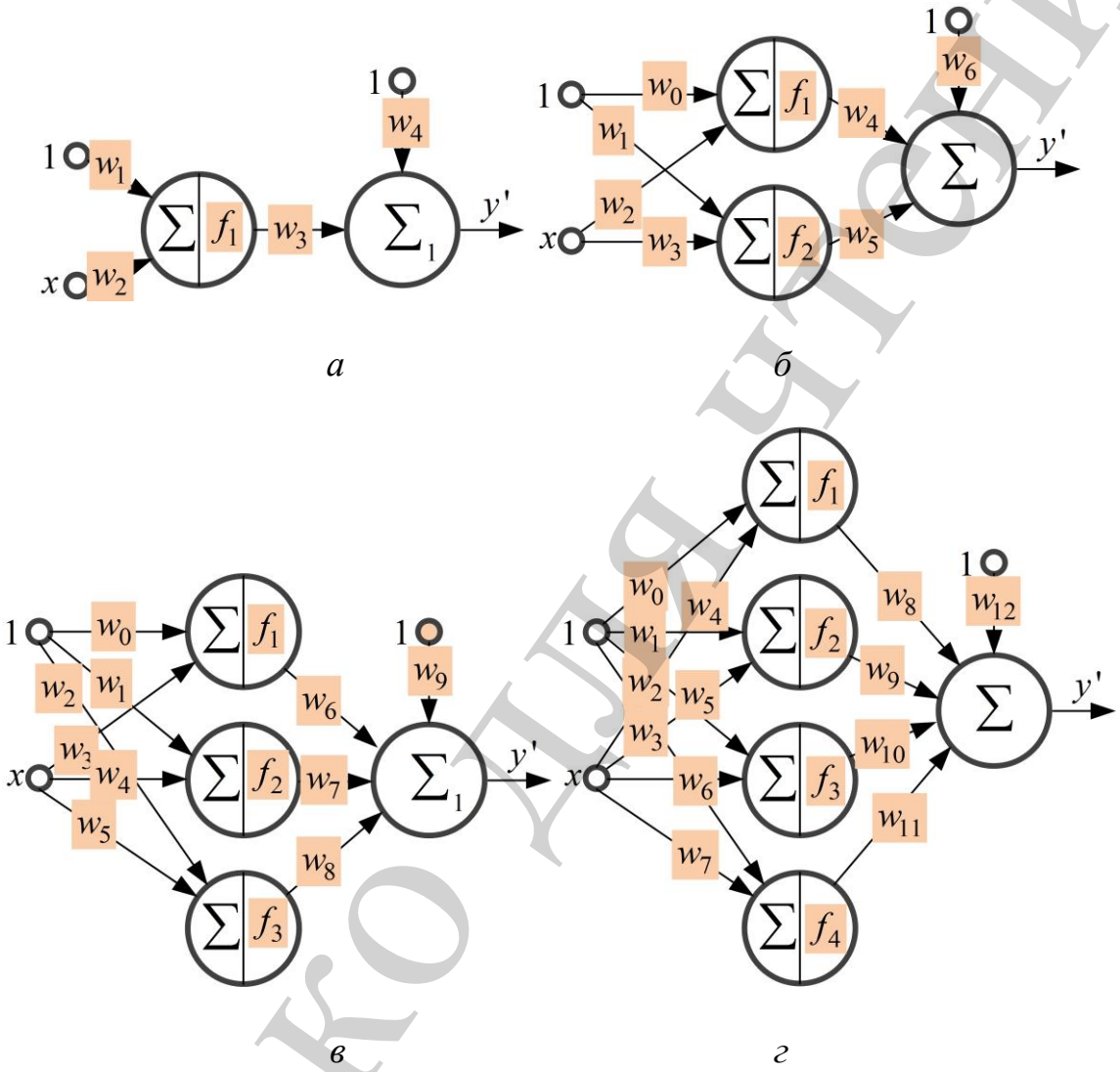


Рис. 4. Збільшення числа параметрів моделі при підвищенні кількості нейронів прихованого шару: *а* – п’ять параметрів при одному нейроні; *б* – дев’ять параметрів при двох нейронах; *в* – тринадцять параметрів при трьох нейронах; *г* – сімнадцять параметрів при чотирьох нейронах

Видно (рис. 1), що кількість параметрів ШНМ, котрі необхідно визначити у процесі навчання лінійно зростає при ускладненні конфігурації.

Для проведення порівняльного аналізу з використанням критерію ефективності використання ресурсів було запропоновано відмовитися від традиційного методу ідентифікації параметрів з використанням алгоритму зворотного роз-

повсюдження похибки та використати метод рівномірного пошуку. Такий підхід дозволяє точніше інтерпретувати результати оцінки часу навчання ШНМ.

Для визначення оптимальних параметрів моделі необхідно встановити інтервали зміни значень ваг ліній зв'язку між вузлами $\left\{ w_{ij}^{(k)} \in \mathrm{N} \mid w_{ij\,\min}^{(k)} \le w_{ij}^{(k)} \le w_{ij\,\max}^{(k)} \right\}$ та коефіцієнту форми $f_i$ нелінійних функцій активації нейронів прихованого шару.

Перед проведенням обчислювальних експериментів було визначено інтервали зміни параметрів порівнюваних моделей. Вони вибиралися таким чином, щоб процес рівномірного пошуку виконувався з кроком, котрий забезпечує перебір значень проміжку за рівну кількість ітерацій. Тобто крок повинен бути кратний діапазону зміни значення параметру.

У результаті процес навчання ШНМ здійснюється шляхом повного перебору значень параметрів моделі з різним кроком, але за однакове число ітерацій.

Після завершення процедури оцінки параметрів здійснювалося визначення часу навчання та значення середньоквадратичного відхилення отриманих значень виходу моделі від тестових даних.

Оскільки складність нейронної мережі зростає лінійно, то експертна оцінка вхідного продукту операції навчання ШНМ (RE) визначається за кількістю параметрів моделі, що підлягають оцінці. Складова RE відображає експертну оцінку поставленої задачі, енергетичних витрат та апаратного забезпечення, котре задіяне у обчислюваному процесі. Складова РЕ визначається експертними оцінками вивільнених ресурсів апаратного забезпечення й експертної оцінки якості апроксимації вихідної функції (АЕ).

Першим етапом отримання складової АЕ критерію ефективності є визначення функції інтерполяції для зміни похибки моделі ШНМ (табл. 1), а отримання прогнозованих значень при підвищенні складності структури моделі здійснюється шляхом екстраполяції функції (рис. 5).

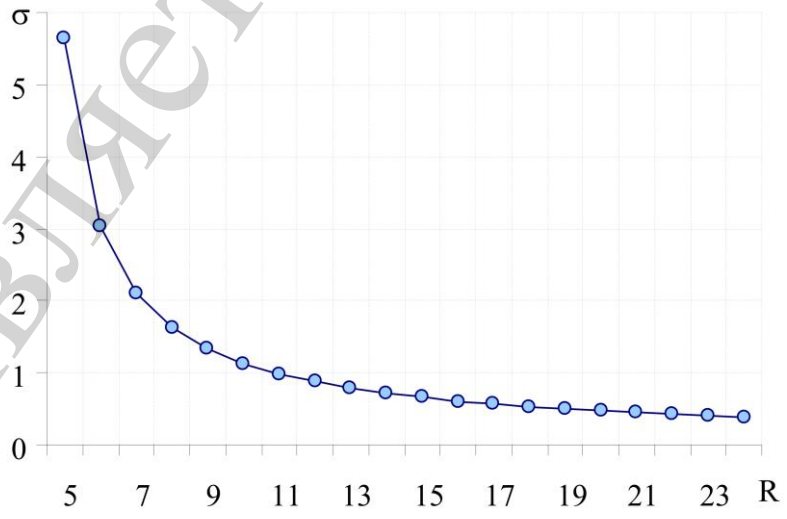


Рис. 5. Зниження похибки моделі при підвищенні складності структури ШНМ

У стовпчику ЕЗ (табл. 1) одиницею позначені дані, котрі отримані експериментально.

Приймемо, що похибка моделі з мінімальною конфігурацією, тобто з одним нейроном у прихованому шарі, має нульову вартість. Підвищення вартості результату апроксимації функції штучною нейронною мережею при підвищенні складності структури моделі можна визначити за наступним виразом:

$$AE = \left(\sigma_1 - \sigma(R)\right) f\left(\mathrm{A}, \mathrm{TO}\right) = \left(\sigma_1 - \sigma(R)\right)\left(\frac{1}{1+e^{-A\cdot\ln(TO+D)}}\right),$$

де $\sigma_1$ – похибка моделі ШНМ з мінімальною конфігурацією; $\sigma(R)$ – функція зміни похибки від складності нейромережевої структури; $f\left(\mathrm{A}, \mathrm{TO}\right)$ – нелінійна функція експертної оцінки точності моделі; А – коефіцієнт форми нелінійної функції; ТО – час навчання, секунд модельного часу; D – зміщення лінеарізованої ділянки часу.

При проведенні обчислювальних експериментів було прийнято, що 10 тис. секунд модельного часу займає 1 секунду реального.

Таблиця 1

Розрахункові дані отримані по точкам інтерполяції (1-13) і екстраполяції (14-20)

| N | R | ТО, с | Ln(TO) | Похибка | Точність | f(R, D) | RE | PE | AE | R | E | ЕЗ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5 | 1,13E-05 | 0 | 5,641 | 0,359 | 0,018 | 5 | 5,01 | 0,006 | 292 | 2,21E-05 | 1 |
| 2 | 6 | 1,28E-04 | 2,04 | 3,032 | 2,968 | 0,047 | 6 | 6,14 | 0,141 | 544 | 0,000259 | 0 |
| 3 | 7 | 1,45E-03 | 4,47 | 2,108 | 3,891 | 0,119 | 7 | 7,46 | 0,464 | 1123 | 0,000413 | 0 |
| 4 | 8 | 1,64E-02 | 6,89 | 1,629 | 4,370 | 0,269 | 8 | 9,17 | 1,176 | 1484 | 0,000792 | 0 |
| 5 | 9 | 1,8E-01 | 9,32 | 1,334 | 4,666 | 0,5 | 9 | 11,33 | 2,333 | 1899 | 0,001228 | 1 |
| 6 | 10 | 2,1 | 11,75 | 1,133 | 4,867 | 0,731 | 10 | 13,56 | 3,558 | 2629 | 0,001353 | 0 |
| 7 | 11 | 2,39E+01 | 14,17 | 0,987 | 5,013 | 0,881 | 11 | 15,41 | 4,415 | 3858 | 0,001144 | 0 |
| 8 | 12 | 2,71E+02 | 16,60 | 0,876 | 5,124 | 0,952 | 12 | 16,88 | 4,881 | 5719 | 0,000853 | 0 |
| 9 | 13 | 3,07E+03 | 19,03 | 0,788 | 5,212 | 0,982 | 13 | 18,12 | 5,118 | 8332 | 0,000614 | 1 |
| 1о | 14 | 3,47E+04 | 21,46 | 0,717 | 5,283 | 0,993 | 14 | 19,25 | 5,247 | 11820 | 0,000444 | 0 |
| 11 | 15 | 3,93E+5 | 23,88 | 0,658 | 5,341 | 0,997 | 15 | 20,33 | 5,328 | 16322 | 0,000326 | 0 |
| 12 | 16 | 4,46E+06 | 26,31 | 0,609 | 5,391 | 0,999 | 16 | 21,38 | 5,386 | 21990 | 0,000245 | 0 |
| 13 | 17 | 5,05E+07 | 28,74 | 0,567 | 5,433 | 0,999 | 17 | 22,43 | 5,431 | 28993 | 0,000187 | 1 |
| 14 | 18 | 5,72E+08 | 31,16 | 0,530 | 5,469 | 0,999 | 18 | 23,47 | 5,469 | 37513 | 0,000146 | 0 |
| 15 | 19 | 6,48E+09 | 33,60 | 0,499 | 5,501 | 1,0 | 19 | 24,50 | 5,501 | 47748 | 0,000115 | 0 |
| 16 | 20 | 7,34E+10 | 36,02 | 0,471 | 5,529 | 1,0 | 20 | 25,53 | 5,529 | 59904 | 9,23E-05 | 0 |
| 17 | 21 | 8,32E+11 | 38,45 | 0,446 | 5,554 | 1,0 | 21 | 26,55 | 5,554 | 74204 | 7,48E-05 | 0 |
| 18 | 22 | 9,42E+12 | 40,87 | 0,424 | 5,576 | 1,0 | 22 | 27,58 | 5,576 | 90881 | 6,14E-05 | 0 |
| 19 | 23 | 1,07E+14 | 43,30 | 0,404 | 5,596 | 1,0 | 23 | 28,60 | 5,596 | 110179 | 5,08E-05 | 0 |
| 20 | 24 | 1,21E+15 | 45,73 | 0,385 | 5,614 | 1 | 24 | 29,61 | 5,614 | 132357 | 4,24E-05 | 0 |

Тут нелінійна функція експертної оцінки вартості (рис. 6) враховує той факт, що суттєве підвищення точності моделі на початковому етапі підвищення складності ШНМ призводить до незначного зростання вартості результату апроксимації. Потім вартість результату швидко зростає, та подальше зниження похибки моделі не призводить до пропорційного підвищення вартості результату.

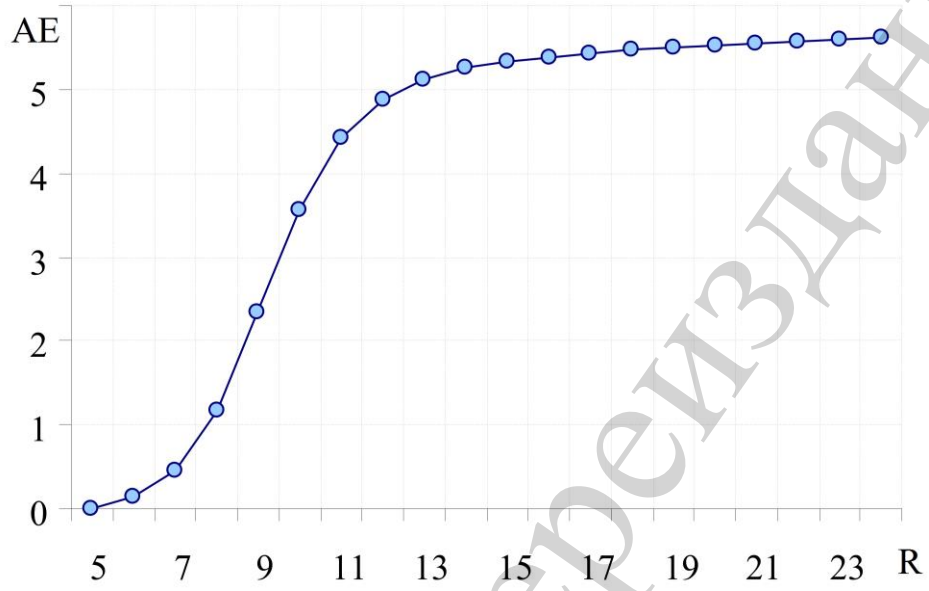


Рис. 6. Нелінійна функція зміни експертної оцінки точності ШНМ

На наступному етапі визначається вид функції інтерполяції для отримання прогнозованих значень часу навчання ШНМ при підвищенні складності структури моделі (рис. 7).

Для коректного визначення критерію ефективності необхідно виконати узгодження його складових. Враховуючи, що складність моделі зростає лінійно, а швидкодія процесу параметричної ідентифікації при збільшенні кількості параметрів – експоненціально, то для формування точки екстремуму доцільно провести лінеаризацію функції залежності часу навчання від числа параметрів.

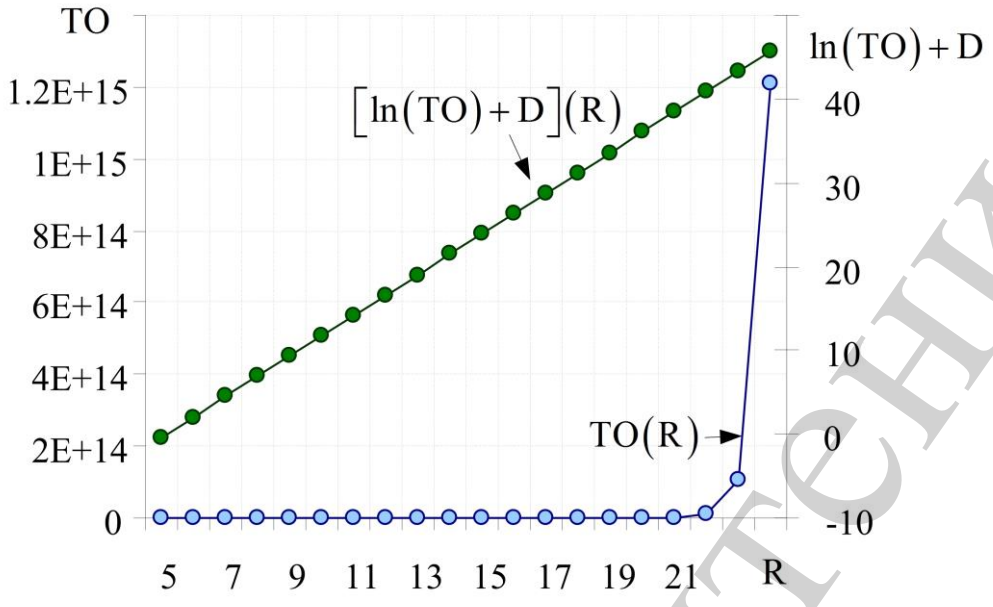


Рис. 7. Лінеаризація часової залежності

У якості критерію оптимізації використано оціночний показник [19], який пройшов верифікацію на предмет його використання у якості критерію ефективності [20–22]:

$$E = \frac{\left(PE - RE\right)^2}{RE \cdot PE \cdot \left[\ln\left(TO\right) + D\right]^2}.$$

Обробка результатів обчислювальних експериментів дозволяє побудувати залежність ефективності застосування ШНМ для апроксимації нелінійної функції виду $y = 1/x$ при підвищенні складності моделі (рис. 8).

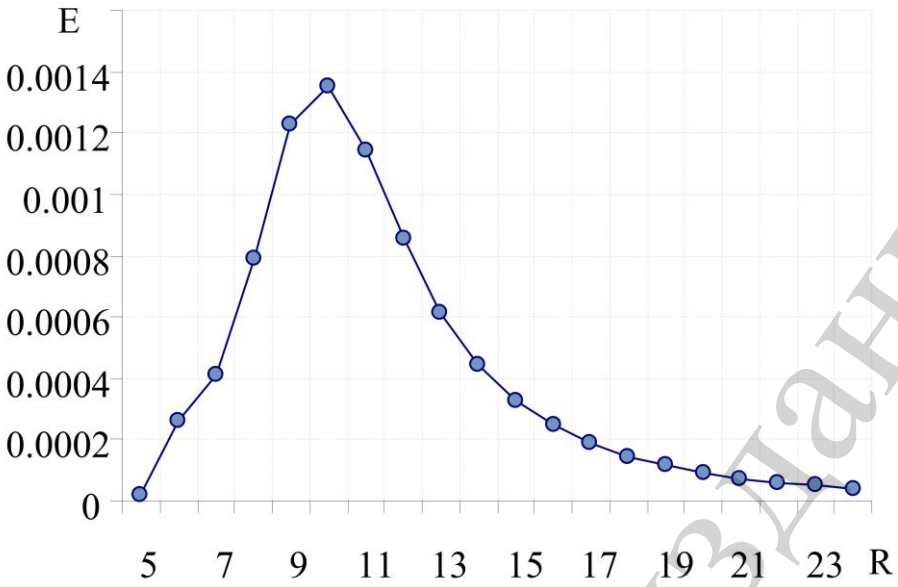


Рис. 8. Зміна ефективності ШНМ від складності її конфігурації

Як видно з рис. 8 оптимальна кількість вузлів прихованого шару ШНМ для рішення розглянутої задачі апроксимації складає 10 нейронів.

**5. Обговорення результатів досліджень, що пов’язані з визначенням структури нейронної мережі**

Штучні нейронні мережі створювалися як обчислювальні об'єкти, що моделюють процеси функціонування людського мозку. Однак створення ШНМ це, швидше, спроба відтворення механізму перетворення інформації, ніж повноцінної структури, здатної самостійно визначати свою архітектуру, у залежності від особливостей задачі, що розв’язується.

Архітектура ШНМ на даний час визначається дослідним шляхом, здебільшого в залежності від області її застосування. Причому одним з найбільш складних завдань є визначення інтервалів зміни параметрів моделі, зокрема ваг ліній зв’язку та коефіцієнтів форми нелінійних функцій активації.

Запропонований підхід дозволяє формалізувати найбільш відповідальну процедуру – вибору складності архітектури ШНМ з урахуванням точності моделі і часу її навчання. При цьому використовується ціннісний підхід, який природним чином пов'язує такі параметри, як складність конфігурації багатошарової мережі з одним прихованим шаром, час навчання і точність отриманої моделі.

Але немає ніяких обмежень у застосуванні методу для прогнозування ефективності функціонування біль складних структур, наприклад, при збільшені кількості нейронів не тільки по вертикалі (рис. 9).

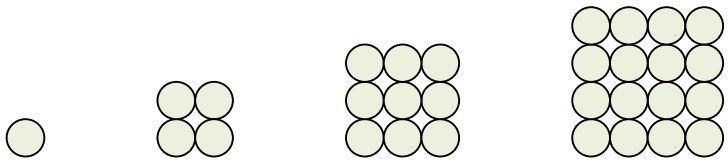

Рис. 9. Схематичне зображення методу зміни структури ШНМ на кожному наступному кроці при збільшені кількості прихованих шарів та одночасному зростанні їх вертикалі

Таким чином, відривається можливість для розширення кола досліджень.

У біологічних нейронних мережах питання структурної і параметричної оптимізації, очевидно, вирішуються з використанням недосліджених на сьогодні технологій, оскільки зачіпають такі аспекти людської діяльності як абстрактне мислення. При цьому швидкість біологічних процесів також є недосяжним еталоном, незважаючи на істотний прогрес у цій галузі.

**6. Висновки**

1. Визначена проблема вибору оптимальної структури ШНМ, котра зумовлена необхідністю узгодження точності отриманої при параметричній ідентифікації моделі та нелінійного зростання часу навчання. Для рішення цієї задачі запропонований підхід, що базується на отриманні прогностичних оцінок, які б пов’язували зростання часу навчання при підвищенні складності архітектури ШНМ з точністю отриманої моделі.

2. Розроблено ціннісний підхід до визначення величин експертних оцінок вхідних та вихідних інформаційних продуктів ШНМ, що забезпечило можливість здійснити узгодження в часі складності структури моделі з рівнем відхилення виходу моделі від тестових даних. Суть запропонованого походу полягає у визначенні експертної оцінки складності розв'язуваної задачі і експертної оцінки цінності результату, що має певну точність. При цьому цінність отриманого рішення нелінійно пов'язана з показником «точність розрахунку».

3. Розроблено метод визначення оптимальної структури ШНМ у вигляді моделі багатошарового персептрону з одним прихованим шаром, що базується на співставленні прогностичних оцінок ефективності використання ресурсів. При цьому вихідними даними для отримання таких оцінок є: експертне значення складності конфігурації мережі, час навчання та експертне значення точності отриманої моделі.

**Література**

1. Горбань А. Н. Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сибирский журнал вычислительной математики. 1998. Т. 1, № 1. С. 11–24.

2. Nelles O. Nonlinear System Identification. From Classical Approaches to Neural Networks and Fuzzy Models. Springer, 2001. 785 p. doi: https://doi.org/10.1007/978-3-662-04323-3

3. Diniz P. S. R. Adaptive Filtering: Algorithms and Practical Implementation. Springer, 2008. doi: https://doi.org/10.1007/978-0-387-68606-6

4. Mykhailenko O. Research of adaptive algorithms of laguerre model parametrical identification at approximation of ore breaking process dynamics // Metallurgical and Mining Industry. 2015. Issue 6. P. 109–117.

5. Mykhailenko O. Ore Crushing Process Dynamics Modeling using the Laguerre Model // Eastern-European Journal of Enterprise Technologies. 2015. Vol. 4, Issue 4 (76). P. 30–35. doi: https://doi.org/10.15587/1729-4061.2015.47318

6. Haykin S. Neural Networks and Learning Machines. 3rd ed. Pearson, 2009. 938 p.

7. A structure optimization algorithm of neural networks for large-scale data sets / Yang J., Ma J., Berryman M., Perez P. // 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). 2014. doi: https://doi.org/10.1109/fuzz-ieee.2014.6891662

8. Learning both Weights and Connections for Efficient Neural Network / Han S., Pool J., Tran J., Dally W. // Proceedings of Advances in Neural Information Processing Systems. 2015.

9. Liu C., Zhang Z., Wang D. Pruning deep neural networks by optimal brain damage // INTERSPEECH 2014. 2014. P. 1092–1095.

10. Tresp V., Neuneier R., Zimmermann H. G. Early Brain Damage // Proceedings of the 9th International Conference on Neural Information Processing Systems NIPS96. 1996. P. 669–675.

11. Optimal Brain Surgeon on Artificial Neural Networks in Nonlinear Structural Dynamics / Christiansen N. H., Job J. H., Klyver K., Hogsbrg J. // In Proceedings of 25th Nordic Seminar on Computational Mechanics. 2012.

12. Babaeizadeh M., Smaragdis P., Campbell R. H. NoiseOut: A Simple Way to Prune Neural Networks // Proceedings of 29th Conference on Neural Information Processing Systems (NIPS 2016). Barcelona, 2016.

13. Reshaping deep neural network for fast decoding by node-pruning / He T., Fan Y., Qian Y., Tan T., Yu K. // 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2014. doi: https://doi.org/10.1109/icassp.2014.6853595

14. Takeda R., Nakadai K., Komatani K. Node Pruning Based on Entropy of Weights and Node Activity for Small-Footprint Acoustic Model Based on Deep Neural Networks // Interspeech 2017. 2017. P. 1636–1640. doi: https://doi.org/10.21437/interspeech.2017-779

15. A New Adaptive Merging and Growing Algorithm for Designing Artificial Neural Networks / Islam M., Sattar A., Amin F., Yao X., Murase K. // IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2009. Vol. 39, Issue 3. P. 705–722. doi: https://doi.org/10.1109/tsmcb.2008.2008724

16. Arifovic J., Gençay R. Using genetic algorithms to select architecture of a feedforward artificial neural network // Physica A: Statistical Mechanics and its Applications. 2001. Vol. 289, Issue 3-4. P. 574–594. doi: https://doi.org/10.1016/s0378-4371(00)00479-9

17. Finding Optimal Neural Network Architecture using Genetic Algorithms / Fiszelew A., Britos P., Ochoa A., Merlino H., Fernández E., García-Martínez R. // Advances in Computer Science and Engineering Research in Computing Science. 2007. Vol. 27. P. 15–24.

18. Yang S.-H., Chen Y.-P. An evolutionary constructive and pruning algorithm for artificial neural networks and its prediction applications // Neurocomputing. 2012. Vol. 86. P. 140–149. doi: https://doi.org/10.1016/j.neucom.2012.01.024

19. Lutsenko I. Definition of efficiency indicator and study of its main function as an optimization criterion // Eastern-European Journal of Enterprise Technologies. 2016. Vol. 6, Issue 2 (84). P. 24–32. doi: https://doi.org/10.15587/1729-4061.2016.85453

20. Development of a verification method of estimated indicators for their use as an optimization criterion / Lutsenko I., Fomovskaya E., Oksanych I., Koval S., Serdiuk O. // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 2, Issue 4 (86). P. 17–23. doi: https://doi.org/10.15587/1729-4061.2017.95914

21. Development of test operations with different duration in order to improve verification quality of effectiveness formula / Lutsenko I., Fomovskaya O., Vihrova E., Serdiuk O., Fomovsky F. // Eastern-European Journal of Enterprise Technologies. 2018. Vol. 1, Issue 4 (91). P. 42–49. doi: https://doi.org/10.15587/1729-4061.2018.121810

22. Development of the method for modeling operational processes for tasks related to decision making / Lutsenko I., Oksanych I., Shevchenko I., Karabut N. // Eastern-European Journal of Enterprise Technologies. 2018. Vol. 2, Issue 4 (92). P. 26–32. doi: https://doi.org/10.15587/1729-4061.2018.126446