DOI 10.31891/2307-5732-2019-275-4-72-79
УДК 004.85, 004.89, 004.9

П.Б. ВІТИНСЬКИЙ, Р.О. ТКАЧЕНКО, І.В. ІЗОНІН, Н.О. КУСТРА
Національний університет «Львівська політехніка»

АНСАМБЛІ НЕЙРОПОДІБНИХ СТРУКТУР МПГП З RBF РОЗШИРЕННЯМ ВХОДІВ ДЛЯ ЗАДАЧ РЕГРЕСІЇ ТА КЛАСИФІКАЦІЇ

У роботі описано архітектурні принципи створення, топології і особливості застосування ансамблів штучних нейронних мереж з неітеративним навчанням. Подаються і обґрунтовуються основні кроки побудови гібридних нейроподібних структур, що лежать в основі функціонування ансамблів, висвітлені базові структури зав'язків між елементами. Обґрунтовано доцільність їх використання. Описано метод побудови ансамблю на основі дихотомії. Надано алгоритмічну реалізацію методу, наведено блок-схему його роботи для обох базових нейроподібних структур, які покладено в основу ансамблів. Моделювання роботи обох ансамблів проведено на реальних даних задачах прогнозування. Описано основні атрибути вибірки даних, наведено її візуалізацію. Надано результати експериментального дослідження щодо точності та швидкості роботи обох розроблених ансамблів. Встановлено, що ансамбль на основі комбінованої нейроподібної структури з додатковим RBF шаром забезпечує вищі показники точності роботи при тривалішій процедурі навчання. Здійснено порівняння роботи розроблених ансамблів з рядом існуючих методів. Встановлено найвищу точність роботи ансамблю на основі комбінованої нейроподібної структури з додатковим RBF шаром. Розроблені ансамблі можна використовувати для розв'язання задач регресії і класифікації в різних областях.

Ключові слова: ансамбль, неітеративні алгоритми навчання, нейроподібні структури, машинне навчання, прогнозування, класифікація.

P.B. VITYNSKIY, R.O. TKACHENKO, I.V. IZONIN, N.O. KUSTRA
Lviv Polytechnic National University

THE ENSEMBLES OF SGTM NEURAL-LIKE STRUCTURES WITH RBF LAYER FOR REGRESSION AND CLASSIFICATION TASKS

The developed ensembles of non-iterative artificial neural networks are described. The basic steps of constructing hybrid neural-like structures which are the basic of ensembles and their topology are given. Their usage is grounded. The method of constructing an ensemble based on dichotomy is described. The algorithmic implementation of method and block-scheme for both neural-like structures underlying ensembles are presented. Modelling of both ensembles performed on real-life forecasting problem. Attributes of the dataset are described and visualized. The results of experimental research including accuracy and speed for both developed ensembles are presented. The ensemble based on a combined neural-like structure of SGTM with RBF layer provides higher accuracy with the longer procedure of training. Comparison between developed ensembles and existing methods is made. The ensemble based on the combined neural-like structure with RBF shows the highest accuracy. The developed ensembles can be used for regression and classification tasks in different in various areas.

Keywords: ensemble, non-iterative training algorithms, neural-like structures, machine learning, forecasting, classification.

Вступ

Завдання ефективного опрацювання даних великих обсягів, тобто пошук методів та способів автоматичного та швидкого аналізу великих об'ємів даних, є важливою в багатьох сферах індустрії. Серед ефективних підходів до її розв'язання являється застосування методів машинного навчання. Ці методи дозволяють знаходити в наборах даних початково невідомі, складні взаємозалежності та закономірності [1].

Однією із задач, які доволі часто виникають у різних прикладних областях, є задача регресії. Задача регресії або навчання за прецедентами, полягає у побудові та навчанні моделі, що буде прогнозувати значення цільової змінної на основі набору вхідних даних. Існуючі методи машинного навчання не завжди забезпечують [2]:

точний результат;

достатні генералізаційні властивості;

швидке опрацювання даних великих обсягів.

Саме тому існує потреба у розробленні нових методів опрацювання великих наборів даних, які нівелюватимуть описані вище недоліки.

Існує чимало підходів до опрацювання даних великих обсягів на основі ансамблів [3]. Згідно з [4] останні можна розділити на дві групи: послідовні та паралельні. Прикладом першого може бути метод Adaboost, прикладом другого – Random Forest [5]. Побудова ансамблю може відбуватися за різними алгоритмами [6]. У цій роботі ми використовуємо ідею кооперації між елементами ансамблю, де навчання кожного з елементів ансамблю проходить паралельно. Для цього необхідне розбиття загальної великої вибірки на менші складові. Опрацювання окремих вибірок даних може забезпечити підвищення точності роботи методу загалом.

Аналіз попередніх досліджень

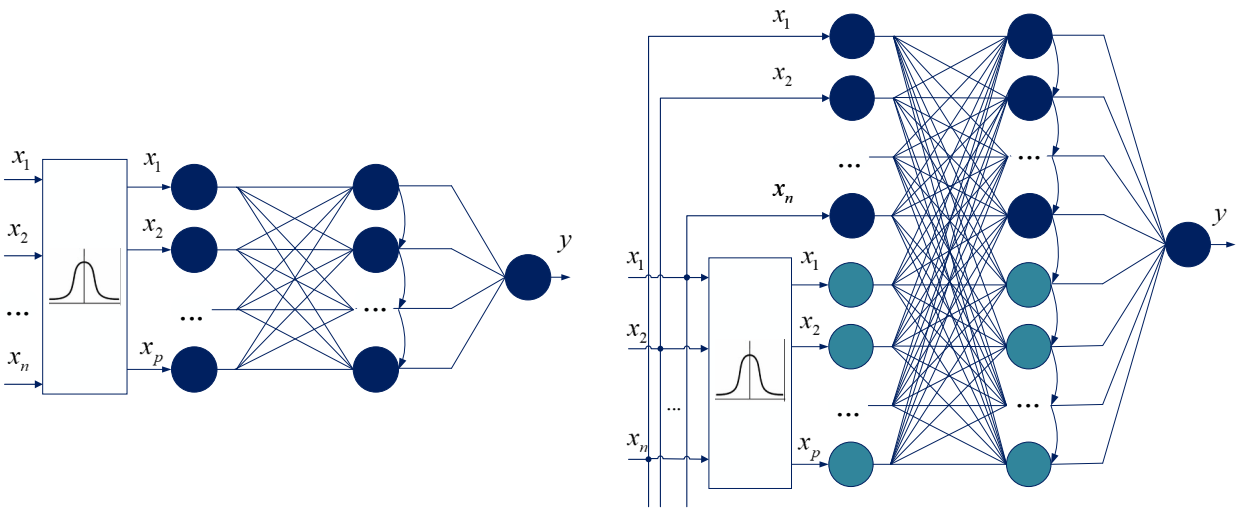
У [7] описано новий неітераційний інструмент штучного інтелекту для розв'язання задач регресії. Парадигма побудови штучних нейронних мереж, закладена в його основі (Модель послідовних геометричних перетворень – МПГП), використовує жадібний алгоритм навчання. Нейроподібні структури, побудовані з використанням цієї моделі, забезпечують високу швидкодію навчання за хороших показників щодо точності роботи [8]. Основні переваги цієї моделі перед існуючими нейронними мережами прямого

поширення, а також деталі алгоритмів її навчання і застосування описано у [9].

У [10] з метою підвищення точності розв'язання задачі регресії розроблено дві гібридні структури цього інструменту обчислювального інтелекту (рис. 1):

нейроподібна структура МПГП з додатковим RBF шаром;

комбінована нейроподібна структура МПГП з додатковим RBF шаром.



а) **б)**

Рис. 1. Топології гібридних нейроподібних структур МПГП з додатковим RBF шаром

Основна ідея гібридизації обох полягає у розширенні початкових входів із використанням додаткового RBF шару. Метою такого розширення є збільшення розмірності простору вхідних даних задачі для отримання більш точного результату. Враховуючи неітераційність роботи обраного інструменту, таке розширення з точки зору швидкодії роботи є доречним.

Для побудови такої структури з рис. 1*а* необхідно [11]:

1. обрати P довільно розподілених точок з навчальної вибірки, вхідні компоненти яких будуть координатами центрів RBF;

2. обрати параметр гаусівської функції активації σ;

3. обчислити евклідові відстані від кожного i-го вектора навчальної та тестової вибірки до кожного центра RBF:

$$R_{p,i} = \sqrt{\sum_{j=1}^{n} (xc_{p,j} - x_{i,j})^2}, \quad (1)$$

де $p = \overline{1,\mathrm{P}}$ є центром; $i = \overline{1,N}$.

обчислити гаусівську функції для усіх (1):

$$F_{p,i} = \exp(-\frac{R_{p,i}^{\ 2}}{\sigma^2}) \quad (2)$$

В результаті виконання таких перетворень початкові входи повністю замінюються на розширені. Розмірність вхідних даних після розширення залежить від кількості RBF центрів, обраних користувачем. Основною відмінністю другого гібриду є те, що він сумісно використовує як початкові входи задачі, так і розширення на основі RBF (рис. 1*б*). Ця комбінація дозволяє підвищити екстраполяційні властивості гібридної структури, чим збільшити точність розв'язання задач класифікації та регресії. З рис. 1*б* видно, що вхідний шар формується із кількості нейронів, яка є рівною сумі вхідних початкових входів задачі та розширених входів із використанням RBF генератора.

Метою роботи є підвищення точності розв'язання задач класифікації та регресії шляхом побудови ансамблів нейроподібних структур моделі послідовних геометричних перетворень з додатковим RBF шаром.

Побудова ансамблів

У роботі пропонується метод побудови ансамблю із використанням дихотомії. Для цього, існуючу вибірку даних розбиваємо на частини згідно з запропонованим алгоритмом. Основними кроками алгоритму для побудови ансамблю нейроподібних структур МПГП з додатковим RBF шаром є такі:

розділити існуючу вибірку на навчальну та тестову;

обчислити середнє значення вихідної змінної **Ошибка! Объект не может быть создан из кодов полей редактирования.** в обох вибірках;

здійснити навчання обраної нейроподібної структури (рис. 1а чи 1б);

застосувати як навчальну, так і тестову вибірки в режимі тестування для отримання поверхні

відгуку **Ошибка! Объект не может быть создан из кодов полей редактирования.** для кожної з вибірок;

розділити як навчальну, так і тестову вибірки на дві підвибірки. До першої увійдуть вектори даних, для яких **Ошибка! Объект не может быть создан из кодов полей редактирования.**, до другої – усі інші;

повторити алгоритм спочатку з кожною з двох отриманих підвибірок до заданих критеріїв зупинки.

Блок-схема алгоритму для поділу навчальної вибірки на дві підвибірки представлено на рис. 2. Таким самим чином відбувається поділ тестової вибірки на частини. В результаті виконання наприклад одного кроку описаного алгоритму ми отримаємо дві підвибірки для двох нейроподібних структур із рис. 2. Для двох кроків поділу ми отримаємо 4 нейроподібні структури і т.д. Поділ продовжуємо за заданих критеріїв зупинки. Слід зазначити, що під *Нейроподібною структурою МПГП* на рис. 2 розуміється або перший, або другий її гібрид залежно від ансамблю, що будується.

Після виконання алгоритму ми отримаємо ансамбль з k нейроподібних структур, кожна з яких буде опрацьовувати якусь унікальну частину із загальної вибірки даних.

Застосування такого ансамблю дозволить зменшити похибки прогнозування [12] та збільшити генералізуючі властивості нейроподібних структур, а використання паралельних обчислень, зокрема на багатопроцесорному ПК зменшити час процедури опрацювання даних великих обсягів [13]. Окрім цього можлива ефективна апаратна реалізація такого підходу [14].

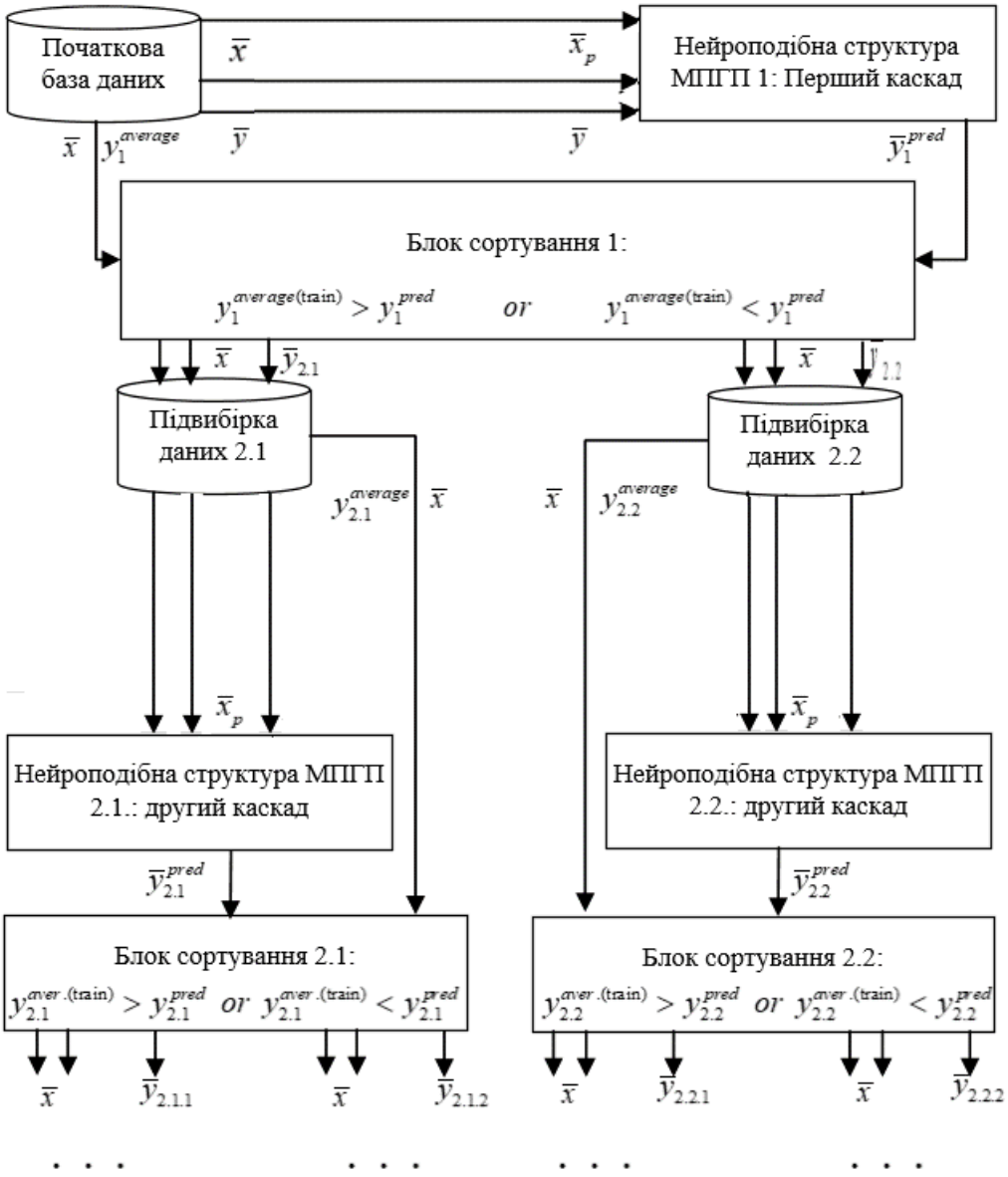


Рис. 2. Блок-схема розробленого алгоритму

Моделювання

Моделювання роботи методу відбувалося на 1338 спостереженнях щодо величини страхових виплат у чотирьох областях США. Цю задачу прогнозування на реальних даних взято із [15]. Вона містила 6 факторних змінних, деякі з яких набували текстових значень. З метою моделювання, вибірку модифіковано наступним чином (Таблиця 1): замість 6 факторних ознак утворено 11 числових факторних змінних [7], які

візуалізовано з використанням Orange [15] на рис. 3. Кольором на рис. 3 виділено курців (синій) та не курців (червоний колір). Різні фігури позначають чоловіків та жінок, а величина фігура визначає розмір страхових виплат.

Таблиця 1

Характеристики вибірки даних

Назва атрибута	Характеристика
Вік особи	Мінімальний – 18 років; максимальний – 64 років; середній – 39.2
Чоловіки	676
Жінки	662
Індекс маси тіла, kg/m2	Мінімальний – 15.96; максимальний – 53.13; середній – 30.66
Кількість дітей	Мінімальна – 0; максимальна – 5; середня – 1.095
Курці-чоловіки	517
Курці-жінки	115
Область проживання: північний захід (у США)	325
Область проживання: південний схід (у США)	364
Область проживання: північний схід (у США)	324
Область проживання: південний захід (у США)	325

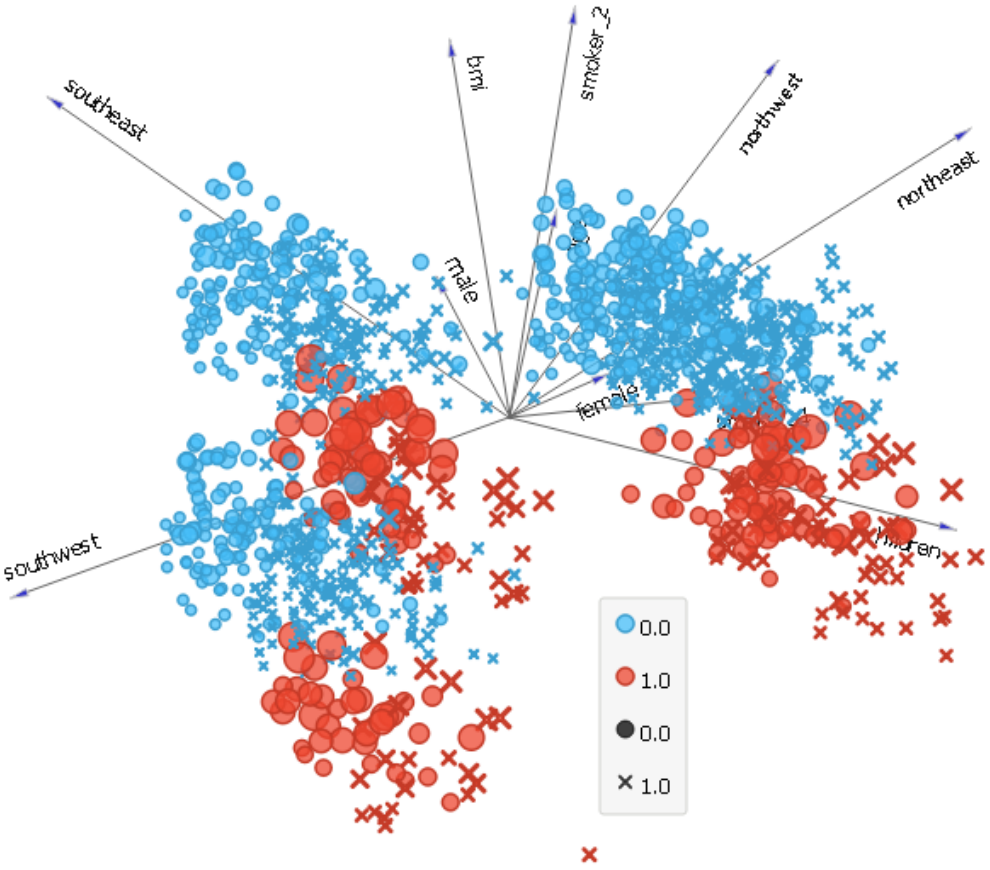


Рис. 3. Візуалізація вибірки даних

Вибірку даних розділено у співвідношенні 1070 до 268 векторів відповідно для навчальної та тестової. Основними параметрами алгоритму є: один крок поділу вибірки на дві підвибірки, тобто лише дві нейроподібні структури комітету, 100 центрів RBF, $\sigma = 5$ (як для першої, так і для другої гібридної структури). Відповідно до обраних параметрів, кожна із структур ансамблю містила;

100 нейронів у вхідному та прихованому шарах для нейроподібної структури МПГП з додатковим RBF шаром (рис. 1*а*) (ансамбль на основі першого гібриду);

111 нейронів у вхідному та прихованому шарах для комбінованої нейроподібної структури МПГП з додатковим RBF шаром (рис. 1*б*) (ансамбль на основі другого гібриду). Варто зазначити, що кількість нейронів вхідного та прихованого шарів в цьому випадку складалася з двох частин: 100 з них це розширення входів із використанням RBF, а решта 11 – це початкові входи вибірки даних із табл. 1.

Результати похибок прогнозування з використанням нейроподібних структур обох ансамблів для кожної підвибірки даних, а також їх зважене значення надано у табл. 2 та табл. 3. Слід зазначити, що

результати роботи розроблених ансамблів подано для обох режимів його роботи (навчання та тестування).

Таблиця 2

Ефективність роботи ансамблю на основі першого гібриду

Вибірки	***Розмірність, векторів***	**Показники ефективності**		
		MAPE, *%*	**RMSE**	**SMAPE**
Режим навчання				
Підвибірка 1	682	36,050815	5765,6463	0,141925
Підвибірка 2	388	32,641979	4657,3945	0,115016
Зважене значення	*1070*	34,814714	5363,7755	0,132167
Режим тестування				
Підвибірка 1	176	41,594660	4500,0046	0,200104
Підвибірка 2	92	18,203836	6668,2257	0,094108
Зважене значення	*268*	33,564974	5244,3193	0,164046

Таблиця 3

Ефективність роботи ансамблю на основі другого гібриду

Вибірки	***Розмірність, векторів***	**Показники ефективності**		
		MAPE, *%*	**RMSE**	**SMAPE**
Режим навчання				
Підвибірка 1	729	30,859413	4904,61659	0,117006
Підвибірка 2	341	32,294283	4940,75105	0,116870
Зважене значення	*1070*	*31,316694*	*4916,13235*	*0,116963*
Режим тестування				
Підвибірка 1	186	35,190399	4404,82332	0,173066
Підвибірка 2	82	18,765715	7063,32974	0,094458
Зважене значення	*268*	*30,164936*	*5218,24693*	*0,149014*

У Таблиці 4 подано значення тривалості процедур навчання для: усієї вибірки даних, двох підвибірок, що були утворені внаслідок одного кроку поділу та загальний час навчання (для одного кроку поділу) кожного з ансамблів. Останній визначався як сума часу навчання усієї вибірки та найбільшого часу навчання однієї із двох підвибірок даних. Таку формулу було обрано у зв'язку із можливістю використання паралельної обробки обох підвибірок, оскільки вектори даних у них не повторюються.

Таблиця 4

Час тривалості процедур навчання обох ансамблів

Вибірка	**Час навчання ансамблю на основі першого гібриду, секунд**	**Час навчання ансамблю на основі другого гібриду, секунд**
Ціла вибірка	0,073339939	0,0987722873687744
Підвибірка 1	0,037760496	0,0533866882324218
Підвибірка 2	0,050544262	0,0689921379089355
Загальний час навчання ансамблю для одного кроку поділу	*0,123884201*	*0,167764425277710*

Саме зважене значення похибок в режимі тестування роботи ансамблю із табл. 2 та табл. 3 та загальний час навчання ансамблю з табл. 4 бралися до уваги при порівнянні роботи розроблених ансамблів з існуючими методами.

Порівняння та обговорення

Для порівняння роботи розроблених ансамблів використано традиційні регресійні методи (багатошаровий перцептрон, нейронна мережа узагальненої регресії, алгоритм AdaBoost та лінійна регресія на основі стохастичного градієнтного спуску), а також існуючі неітеративні розробки (нейроподібна структура МПГП, нейроподібна структура МПГП з RBF шаром (рис. 1*а*) та комбінована нейроподібна структура МПГП з RBF шаром (рис. 1*б*)).

У роботі проведено експериментальне оцінювання похибок як режиму навчання, так і застосування усіх досліджуваних методів (рис. 4), а також тривалість їхнього навчання (рис. 5).

Як видно з рис. 4 існуючі ітеративні методи (багатошаровий перцептрон, алгоритм AdaBoost та лінійна регресія на основі Стохастичного градієнтного спуску) показують точність роботи розв'язання задачі прогнозування страхових виплат меншу за 50 %. Окрім цього, багатошаровий перцептрон демонструє найнижчу швидкість роботи (рис. 5). Алгоритм AdaBoost та лінійна регресія на основі стохастичного градієнтного спуску демонструють одні з найкращих результатів щодо часу їх навчання, проте, зважаючи на показники точності, це незадовільний результат.

Рис. 4. Порівняння точності роботи усіх досліджуваних методів

Рис. 5. Порівняння часу навчання усіх досліджуваних методів

Нейронна мережа узагальненої регресії демонструє високу точність роботи: більш ніж на 1,5% нижчу MAPE за один із розроблених ансамблів. Проте, ця нейронна мережа демонструє один із найгірших показників щодо часу процедури застосування (оскільки навчання як такого вона не має). Вона займає передостаннє місце щодо цього критерію із усіх досліджених методів і працює в 3 рази повільніше за розроблений ансамбль на основі другого гібриду. Окрім цього, для отримання такого високого результату щодо точності, оптимальний параметр її роботи σ $(\sigma = 0.4)$ підбирався експериментальним шляхом $(\sigma \in [0.1,1.5], \Delta = 0.1)$, де для кожного σ із заданого інтервалу затрачався приблизно такий же час ($\approx 0{,}533577$ seconds) для реалізації процедури навчання. З огляду на це, результат роботи цієї мережі також

не є оптимальним.

Якщо розглядати роботу найбільш подібних, неітераційних методів, то один із найвищих показників щодо швидкості роботи демонструє нейроподібна структура МПГП. Проте точність роботи цього інструменту сягає близько 59%, чого також недостатньо для розв'язання поставленої задачі. Результати роботи двох її гібридів з рис. 1 показують у 2,7 та 2,9 разів відповідно меншу швидкість реалізації процедури навчання. Це пояснюється збільшенням кількості входів цих неітеративних інструментів з 11 у базового інструменту до 100 та 111 у першого та другого гібриду відповідно. Проте як перша, так і друга гібридна нейроподібні структури показують значне збільшення точності роботи під час розв'язання поставленої задачі: на більш ніж 5% та 6% відповідно.

Ансамбль, побудований на основі першого гібриду, як і очікувалося, демонструє як збільшення точності (на 3,3% вища точність на основі MAPE), так і часу навчання (у 1,68 разів довше) в порівнянні з роботою однієї базової структури (рис. 1*а*).

Ансамбль побудований на основі використання другого гібриду демонструє найвищі значення у точності роботи як згідно MAPE так і RMSE. Він показує більше ніж на 5% збільшення точності в порівнянні із роботою однієї базової нейроподібної структури (рисю 1*б*) при збільшенні швидкості процедури навчання лише у 2 рази.

Розроблений ансамбль на основі другого гібриду за рахунок комбінованого використання як початкових входів, так і їх розширення з використанням RBF значно збільшив вимірність кожного вхідного вектору даних (з 11 до 111). Проте це сприяло значному підвищенню екстраполяційних властивостей кожної нейроподібної структури ансамблю. В порівнянні, ансамбль на основі першого гібриду використовував лише замінені початкові входи (11) на RBF входи (100). Це не забезпечило значного збільшення точності. Зокрема, ансамбль на основі другого гібриду забезпечує на 3,4% вищу точність прогнозу при збільшенні часу навчання в 1,34 рази в порівнянні із ансамблем на основі першого гібриду. Це може сприяти його використанню для розв'язання практичних задач у сфері медичного страхового бізнесу, електронної комерції, медицини, економіки, матеріалознавства тощо.

Висновки

Авторами розроблено ансамблі неітеративних нейроподібних засобів штучного інтелекту підвищеного рівня точності і генералізації. Їх принципи побудови базуються на використанні набору гібридних нейроподібних структур МПГП. Гібридизація полягала у сумісному або комбінованому використанні початкових та RBF входів. В результаті використання такої комбінації, простір вхідних даних суттєвим чином розширювався. Це впливало як на точність роботи кожної окремої мережі ансамблю та її екстраполятивні властивості так і на швидкодію процедур навчання. З огляду на неітеративність характеру процедури навчання, значне розширення простору вхідних даних задачі для підвищення точності має сенс.

Моделювання роботи обох ансамблів відбувалося на задачі прогнозування страхових виплат. Експериментальним чином встановлено суттєве зменшення похибок роботи ансамблів у порівнянні із однією базовою нейроподібною структурою. Похибка зменшилася на 5% при невеликому підвищення часу процедури навчання.

Порівняння роботи розроблених ансамблів з існуючими методами обчислювального інтелекту, ітераційного та неітераційного типів підтвердили ефективність їх використання. Розроблений ансамбль на основі другого гібриду забезпечує найменшу похибку на основі MAPE та RMSE. Окрім цього він демонструє задовільні часові характеристики навчання про використанні двох структур у ансамблі. Розроблений ансамбль в програмному чи апаратному варіантах можна застосовувати для розв'язання задач регресії та класифікації на вибірках даних великих обсягів в різних галузях промисловості.

Література

1. Zhernova P. Y. Adaptive Kernel Data Streams Clustering Based on Neural Networks Ensembles in Conditions of Uncertainty about Amount and Shapes of Clusters / P. Y. Zhernova, A. O. Deineko, Y. V. Bodyanskiy, V. O. Riepin VO // 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2018, pp. 7–12. doi: 10.1109/DSMP.2018.8478616

2. Bodyanskiy Y. V. An evolving connectionist system for data stream fuzzy clustering and its online learning / Y. V. Bodyanskiy, O. K. Tyshchenko, D. S. Kopaliani // Neurocomputing, vol. 262, 2018, pp. 41–56. doi: 10.1016/j.neucom.2017.03.081

3. Rokach L. Taxonomy for characterizing ensemble methods in classification tasks: A review and annotated bibliography / Computational Statistics & Data Analysis, vol. 53, 2009, pp. 4046–4072. doi: 10.1016/j.csda.2009.07.017

4. Smolyakov V. Ensemble Learning to Improve Machine Learning Results / Stats and Bots., 2017. URL : https://blog.statsbot.co/ensemble-learning-d1dcd548e936. Accessed 24 Feb 2019

5. Ensemble Methods: Foundations and Algorithms. In: CRC Press. URL : https://www.crcpress.com/Ensemble-Methods-Foundations-and-Algorithms/Zhou/p/book/9781439830031. Accessed 24 Feb 2019

6. Sharkey AJC. Types of Multinet System. Roli F, Kittler J (eds) Multiple Classifier Systems. Springer Berlin Heidelberg, 2002, pp. 108–117.

7. Tkachenko R. Development of the Non-Iterative Supervised Learning Predictor Based on the Ito Decomposition and SGTM Neural-Like Structure for Managing Medical Insurance Costs / R. Tkachenko, I. Izonin, P. Vitynskyi, N. Lotoshynska, and O. Pavlyuk // Data, vol. 3, no. 4, p. 46, Oct. 2018.

8. Doroshenko A. Piecewise-Linear Approach to Classification Based on Geometrical Transformation Model for Imbalanced Dataset / 2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP), 2018, pp. 231–235.

9. Tkachenko R. Geometrikal data modelling / R. Tkachenko, P. Tkachenko, O. Tkachenko, and J. Schmitz, // Proceedings of the international conference on Intelligent systems of making decisions and applied aspects of information technology, Eupatoria, vol. 2, 2016, pp. 279–285.

10. Tkachenko R. Non-iterative Neural-like Predictor for Solar Energy in Libya / R. Tkachenko, H. Kutucu, I. Izonin, A. Doroshenko, and Y. Tsymbal // Proceedings of the 14th International Conference on ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer. Volume I: Main Conference, Kyiv, Ukraine, May 14–17, 2018, 2018, vol. 2105, pp. 35–45.

11. Babichev S. A Fuzzy Model for Gene Expression Profiles Reducing Based on the Complex Use of Statistical Criteria and Shannon Entropy / S. Babichev, V. Lytvynenko, A. Gozhyj, M. Korobchynskyi, and M. Voronenko // Advances in Computer Science for Engineering and Education, 2018, pp. 545–554.

12. Shakhovska N. Big data federated repository model / N. B. Shakhovska, Y. J. Bolubash, and O. M. Veres // The Experience of Designing and Application of CAD Systems in Microelectronics, 2015, pp. 382–384.

13. Tsmots I. Structure and Software Model of a Parallel-Vertical Multi-Input Adder for FPGA Implementation / I. Tsmots, O. Skorokhoda, V. Rabyk // Computer Sciences and Information Technologies - Proceedings of the 11th International Scientific and Technical Conference, CSIT 2016, 2016, pp. 158–160.

14. “Medical Cost Personal Datasets”. URL : https://www.kaggle.com/mirichoi0218/insurance. Accessed: 08-Dec-2018.

15. Demšar et al. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research, vol. 14, pp. 2349–2353, 2013.

Рецензія/Peer review : 24.05.2019 р. Надрукована/Printed :17.7.2019 р.
Рецензент: д.т.н., проф. Цмоць І. Г.