



П. Б. Вітинський, Р. О. Ткаченко, І. В. Ізонін

Національний університет "Львівська політехніка", м. Львів, Україна

АНСАМБЛЬ МЕРЕЖ GRNN ДЛЯ РОЗВ'ЯЗАННЯ ЗАДАЧ РЕГРЕСІЇ З ПІДВИЩЕНОЮ ТОЧНІСТЮ

Розроблено метод ансамблювання нейронних мереж узагальненої регресії для підвищення точності розв'язання задачі прогнозування. Описано базові положення функціонування нейронної мережі узагальненої регресії. На основі цього подано алгоритмічну реалізацію розробленого ансамблю. Аналітично доведено можливість підвищення точності прогнозу із використанням розробленого ансамблю. Із використанням бібліотек мови Python, розроблено програмне рішення для реалізації описаного методу. Проведено експериментальне моделювання роботи методу на реальних даних задачі регресії. Встановлено високу ефективність розв'язання поставленої задачі із застосуванням розробленого методу на основі як середньої абсолютної похибки у відсотках, так і з використанням середньоквадратичної похибки. Здійснено порівняння роботи методу із наявними: апроксимацією поліномом Вінера на основі Стохастичного Градієнтного спуску, нейронною мережею узагальненої регресії та модифікованим алгоритмом AdaBoost. Експериментальним шляхом доведено найвищу точність розв'язання поставленої задачі розробленим методом на основі обох показників точності серед усіх розглянутих у роботі методів. Зокрема, він забезпечує більш ніж на 3,4, 4,3 та 8,3 % (MAPE) вищу точність порівняно із наявними методами відповідно. Розроблений метод можна використовувати для отримання розв'язків підвищеної точності під час вирішення прикладних завдань електронної комерції, медицини, матеріалознавства, бізнес-аналітики та інших.

Ключові слова: регресія; генералізаційні властивості; підвищення точності; нейронні мережі узагальненої регресії.

Вступ. Нейронну мережу узагальненої регресії (GRNN) представлено Donald F. Specht у 1991 p. (Specht, 1991). Цю мережу можна використовувати для моделювання дуже нерегулярних, істотно нелінійних поверхонь відгуку. З часу свого створення вона або її гібриди набули широкого застосування для розв'язання різноманітних практичних задач. Зокрема, у роботі (Ivanets, Bukricieva & Dvornik, 2011) розглянуто модель прогнозування економічних процесів, у роботі (Terekovskiy, 2011) – метод діагностики стану комп'ютерної мережі із використанням GRNN. У роботах (Bodianskyi et al., 2017; Deineko, Bodianskyi & Pliss, 2012) розроблено методи нечіткої кластеризації та еволюційної нейрофазі системи, в основі яких закладено використання мережі GRNN.

Відомо низку переваг нейромереж GRNN перед нейронними мережами інших типів (Izonin et al., 2019). Серед них відсутність процедури навчання, потреба налаштування єдиного параметра нейромережі та високі генералізаційні властивості.

До недоліків мереж GRNN відносять відносно неви-

соку точність і певні часові затримки в режимі застосування (Izonin et al., 2019). Враховуючи швидкісні параметри сучасних комп'ютерних засобів, а також можливість застосовувати кластерні технології для розв'язку задач нейромережею цього типу за окремими кластерами (Specht, 1991), основним недоліком мереж GRNN, значення якого бажано мінімізувати, вважають значні похибки функціонування.

Нейронна мережа узагальненої регресії: базові положення. Проаналізуємо деякі базові особливості алгоритму функціонування GRNN. Нехай задано набір спостережень певного явища/об'єкта. Кожне спостереження містить вектор незалежних змінних \bar{x} та залежну – y . Для певної кількості спостережень із набору відомо значення шуканої компоненти, інші – не містять значень, через різноманітні причини. Задача полягає у здійсненні передбачень значень невідомої залежної компоненти конкретного спостереження:

$$y = f(\bar{x}) \quad (1)$$

із використанням нейронної мережі.

Якщо набір спостережень подати у матричній формі

Інформація про авторів:

Вітинський Павло Богданович, аспірант, кафедра інформаційних технологій видавничої справи.

Email: pavlo.vitynsky@gmail.com; <https://orcid.org/0000-0002-3183-3596>

Ткаченко Роман Олексійович, д-р техн. наук, професор, завідувач кафедри інформаційних технологій видавничої справи.

Email: roman.tkachenko@gmail.com; <https://orcid.org/0000-0002-9802-6799>

Ізонін Іван Вікторович, канд. техн. наук, асистент кафедри інформаційних технологій видавничої справи.

Email: ivanizonin@gmail.com; <https://orcid.org/0000-0002-9761-0096>

Цитування за ДСТУ: Вітинський П. Б., Ткаченко Р. О., Ізонін І. В. Ансамбль мереж GRNN для розв'язання задач регресії з підвищеною точністю. Науковий вісник НЛТУ України. 2019, т. 29, № 8. С. 120–124.

Citation APA: Vitynskiy, P. B., Tkachenko, R. O., & Izonin, I. V. (2019). GRNN ensemble for increasing the accuracy of regression tasks. *Scientific Bulletin of UNFU*, 29(8), 120–124. <https://doi.org/10.36930/40290822>

$X = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_N)^T$, то продукування відгуку y_k на основі відповідних \bar{x}_k , з урахуванням відомої частини набору \bar{x}_i та y_i , можна виконати за допомогою методу GRNN.

Це передбачає виконання таких кроків (Izonin et al., 2019):

Пошук евклідових віддалей від вхідного вектора з компонентами $x_{k,j}$ до наявних векторів з відомими значеннями виходів, які вважаємо опорними $x_{i,j}$:

$$E_{k,i} = \sqrt{\sum_{j=1}^n (x_{k,j} - x_{i,j})^2}, \quad (2)$$

де: $i, i = \overline{1, N}$ – номер опорного вектора (спостереження), для якого завжди відомі значення виходів y ; $j = \overline{1, n}$ – номер вхідної ознаки вектора кожного спостереження; $k, k = \overline{1, \dots}$ – номер вхідного вектора (спостереження), для якого невідомі значення виходів y .

Обчислення гаусівських функцій від евклідових віддалей (2):

$$G_{k,i} = \exp\left(-\frac{(E_{k,i})^2}{\sigma^2}\right), \quad (3)$$

де σ – коефіцієнт розмаху гаусівської функції.

Обчислення значення шуканої величини y_k відповідно до робочої формули методу GRNN:

$$y_k^{pred} = \sum_{i=1}^N y_i G_{k,i} / \sum_{i=1}^N G_{k,i}, k = \overline{1, \dots} \quad (4)$$

Метою роботи є підвищення точності функціонування мережі GRNN способом моделювання формули похибки методу для корекції результатів прогнозування.

Об'єктом дослідження є процеси розв'язання задач регресії в умовах частково пропущених даних за допомогою методу машинного навчання.

Предметом дослідження є методи побудови ансамблів штучних нейронних мереж на основі нейропарадигми узагальненої регресії даних.

Задачею досліджень є розроблення методу побудови ансамблю нейронних мереж узагальненої регресії як засобу з найвищим рівнем генералізації даних, що до цього використовувалися в режимі розділеного застосування.

Для досягнення поставленої мети потрібно: проаналізувати складові похибки формування вихідного сигналу окремою мережею GRNN, розробити основи методу побудови ансамблю з двох мереж цього типу, його програмну реалізацію та провести експериментальний аналіз і порівняння його ефективності.

Складові похибки формування вихідного сигналу GRNN. Проаналізуємо складову методичної похибки формування вихідного сигналу мережі GRNN. Для цього запишемо очевидну тотожність:

$$\frac{\sum_{i=1}^N (y_k - y_i) G_{k,i}}{\sum_{i=1}^N G_{k,i}} = y_k - \frac{\sum_{i=1}^N y_i G_{k,i}}{\sum_{i=1}^N G_{k,i}}. \quad (5)$$

Введемо позначення:

$$z_{k,i} = (y_k - y_i) G_{k,i}. \quad (6)$$

Враховуючи позначення (6), формулу (5) можна подати так:

$$y_k = \frac{\sum_{i=1}^N y_i G_{k,i}}{\sum_{i=1}^N G_{k,i}} + \frac{\sum_{i=1}^N z_{k,i} G_{k,i}}{\sum_{i=1}^N G_{k,i}}. \quad (7)$$

Перший член правої частини рівності (7) відповідає формулі (4) обчислень вихідного сигналу мережею GRNN. Логічно припустити, що якщо рівності (6)-(7) точні, то другий член формули відображає похибку методу GRNN:

$$\Delta_k = \sum_{i=1}^N z_{k,i} G_{k,i} / \sum_{i=1}^N G_{k,i}, k = \overline{1, \dots} \quad (8)$$

Відому методичну складову похибки, як різницю між точним і знайденим за формулою (4) значеннями, можна також обчислити за допомогою (8), але лише для кожного з N опорних векторів. Однак ця формула показує, що поверхня відгуку похибки є достатньо гладкою, і отже, може бути певним способом промодельована в локальній області простору вхідних змінних. Як підтвердили експерименти, застосування ще однієї мережі GRNN зі зменшеною величиною коефіцієнта розмаху σ забезпечує задовільне наближення похибки методу. Візьмемо до уваги, що для підвищення точності обчислень величини похибки згідно з формулою:

$$\Delta_k^{pred} \approx \frac{\sum_{i=1}^N (y_i^{pred} - y_i) G_{k,i}}{\sum_{i=1}^N G_{k,i}} \quad (9)$$

необхідно обирати значно менші значення коефіцієнта розмаху σ , ніж при їх застосуванні для формули (4), що пояснюють відмінностями рельєфів поверхонь відгуків.

Ансамбль на основі двох нейронних мереж узагальненої регресії. З описаного вище впливає метод підвищення точності розв'язання задачі регресії, побудований на ансамблі GRNN з двох елементів, із використанням загальної концепції застосування мереж цього типу. Він складається із двох основних етапів: підготовка даних та процедура застосування.

1. Процедура попередньої підготовки даних передбачає виконання таких кроків:

- по черзі для кожної i -ої опорної точки $i = \overline{1, N}$ відносно $N - 1$ точок ($l = \overline{1, N - 1}$), що залишилися, обчислюємо відгук методом GRNN:

$$y_i^{pred} = \sum_{l=1}^{N-1} y_l G_{i,l} / \sum_{l=1}^{N-1} G_{i,l}, i = \overline{1, N}; \quad (10)$$

- обчислюємо величини відхилень між точними і обчисленими значеннями:

$$\Delta_i = y_i - y_i^{pred}; \quad (11)$$

2. Процедура застосування ансамблю мереж GRNN для рухомого k -го вектора передбачає виконання таких кроків:

- застосовуючи (4), обчислюємо y_k^{pred} ;
- для прогнозування величини похибки повторно застосовуємо формулу GRNN:

$$\Delta_k^{pred} = \sum_{i=1}^N \Delta_i G_{k,i} / \sum_{i=1}^N G_{k,i}, k = \overline{1, \dots}; \quad (12)$$

- остаточний результат роботи методу отримуємо згідно з формулою:

$$y_k \approx y_k^{pred} + \Delta_k^{pred}, k = \overline{1, \dots} \quad (13)$$

Моделювання та результати

Опис даних. Моделювання роботи розробленого методу відбувалося на даних задачі відновлення пропусків у даних екологічного моніторингу стану повітряного середовища. Вибірка містила реальний набір даних про хімічний склад повітря біля італійського міста

(Repository, 2019; Vito et al., 2008). Її було зібрано пристроєм інтернету речей. Сенсори пристрою збирали інформацію про щогодинний хімічний склад атмосферного повітря та інших чинників на основі набору показників, поданих на рис. 1 (Izonin et al., 2019). На рис. 1 подано коротку характеристику кожної змінної: назва показника хімічного складу повітря, його середнє, мінімальне та максимальне значення обчислені для усієї вибірки даних.



Рис. 1. Характеристика вибірки даних

Враховуючи те, що найбільше пропущених значень мів стовпець CO, моделювання проводилося саме для відновлення втрачених даних цього показника (Mishchuk & Tkachenko, 2019).

Оскільки набір даних містив дуже багато пропусків, усі вектори із пропущеними значеннями хоча би одного із 12 показників було вилучено для реалізації процедур прогнозування. Таким способом моделювання проводилося на наборі, розмірністю 6950 векторів (Mishchuk & Tkachenko, 2019). Його було розділено на дві вибірки у співвідношенні 80 до 20 %. Першу вибірку даних використано для навчання, другу – для тестування.

Для аналізу результатів роботи розробленого методу, у роботі використано такі показники (Molnár et al., 2014; Kaczor & Kryvinska, 2013):

$$MAPE = \frac{1}{N} \sum_{i=1}^N \left| \frac{y_i^p - y_i}{y_i} \right| 100; \quad (14)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i^p - y_i)^2}, \quad (15)$$

де y_i – фактичне, а y_i^p – отримане значення для кожного i вектора.

Підбір оптимальних параметрів роботи розробленого методу. Нейронна мережа узагальненої регресії характеризується єдиним параметром налаштування мережі – коефіцієнтом розмаху функції активації σ . Відповідно, запропонований метод на основі ансамблю із двох GRNN також буде залежати від величини цього параметра. У роботі проведено оптимізацію методом перебору (Dubrovyn et al., 2003) для визначення значень розмаху (σ) для відповідних гаусівських функцій обох мереж.

Позначимо через σ_1 – параметр розмаху функції активації основної нейромережі, а σ_2 – відповідно додаткової. Експеримент відбувався при зміні значень $\sigma_1 (\sigma_1 \in [0.05, 1.45], \Delta\sigma_1=0.05)$ та $\sigma_2 (\sigma_2 \in [0.05, 1.45], \Delta\sigma_2=0.05)$ для обчислення MAPE та RMSE розробленого методу. Кількість входів GRNN – 11, один вихід. Отримані результати як для (14), так і для (15) візуалізовано на рис. 2. На цьому рисунку на осі ox подано різні значення розмаху функції активації основної мережі GRNN σ_1 . На осі oy відображено значення розмаху функції активації додаткової мережі GRNN σ_2 . Вісь oz відповідає значенням похибок MAPE (див. рис. 2, а) та RMSE (див. рис. 2, б) для різних комбінацій σ_1 та σ_2 . Як видно з обох поверхонь, найменші похибки отримано при значенні $\sigma_2 = 0.05$. Для деталізації представлення отриманого результату, на рис. 3 наведено значення MAPE (див. рис. 3, а) та RMSE (див. рис. 3, б) при $\sigma_2 = 0.05$ на всьому проміжку $\sigma_1 (\sigma_1 \in [0.05, 1.45], \Delta\sigma_1=0.05)$.

Графіки на рис. 3 можна умовно розбити на три інтервали: перший – при $\sigma_1 \in [0.05, 0.1]$, другий – при $\sigma_1 \in [0.15, 0.25]$ та третій – $\sigma_1 \in [0.3, 1.45]$. Перший з них характеризується різким спадом похибок застосування розробленого методу при мінімальній зміні значення σ_1 . Другий інтервал демонструє стадію насичення (помаранчевим кольором виділено відповідні кружечки на обох графіках), а третій – стадію росту обох похибок при зміні σ_1 за інших рівних умов.

Для точнішого визначення оптимального значення розмаху функції активації основної GRNN з другого інтервалу, на рис. 4 наведено деталізацію зміни значень обох індикаторів точності при $\sigma_1 \in [0.1, 0.29]$, де $\Delta\sigma_2=0.01$ за інших рівних умов.

Як видно з рис. 4 (помаранчевим кольором виділено відповідні кружечки на обох графіках), оптимальні результати $MAPE = 18,689 \%$, $RMSE = 0,461$ отримано при $\sigma_2 = 0.05$ та $\sigma_1 = 0.24$.

Порівняння. Точність роботи розробленого методу на основі (14) та (15) порівнювали із результатом роботи відомих методів:

- нейронною мережею узагальненої регресії (Izonin et al., 2019);
- модифікованим алгоритмом AdaBoost (Izonin et al., 2019a);
- апроксимацією поліномом Вінера на основі Стохастичного Градієнтного спуску (Izonin et al., 2019b).

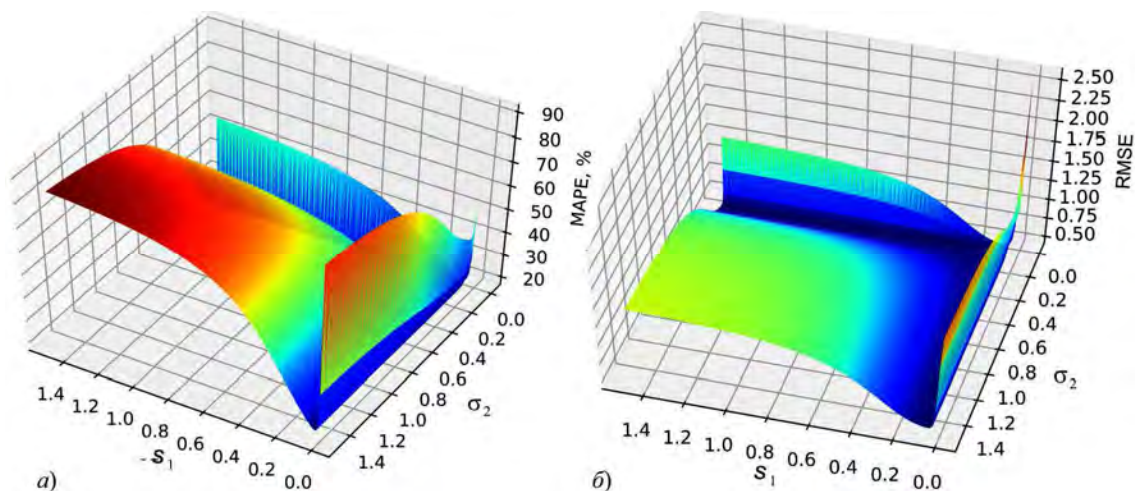


Рис. 2. Результати роботи розробленого методу при різних комбінаціях значень розмаху функції активації обох мереж: а) MAPE, %; б) RMSE

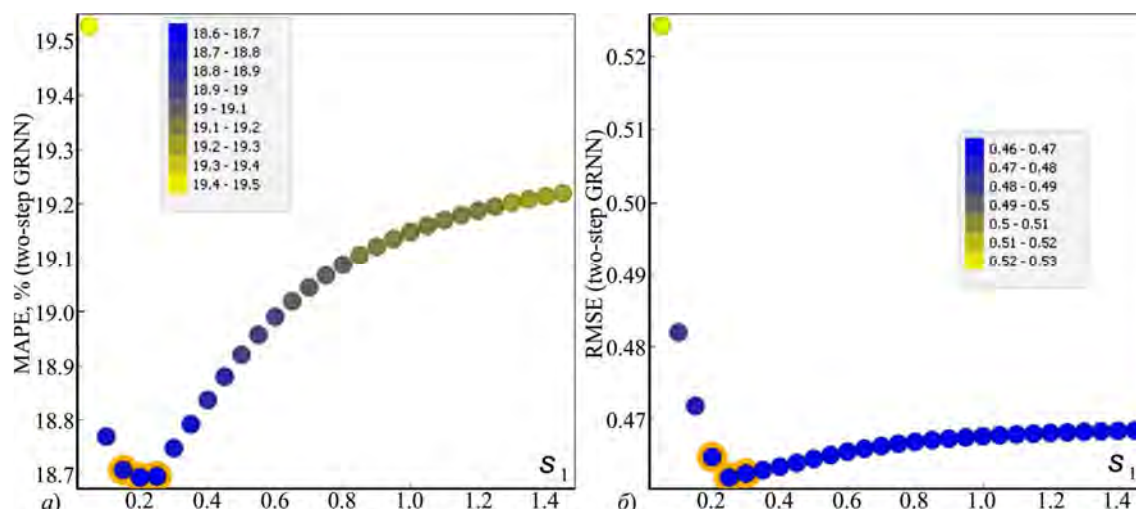


Рис. 3. Деталізація результатів роботи методу при $\sigma_2 = 0.05$ та зміні значень $\sigma_1 (\sigma_1 \in [0.05, 1.45], \Delta\sigma_1 = 0.05)$: а) MAPE, %; б) RMSE

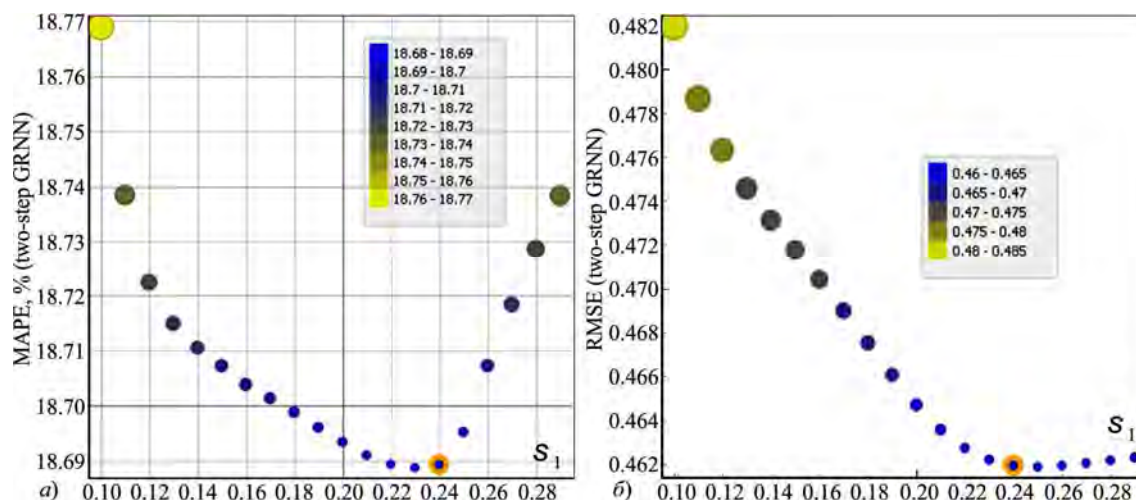


Рис. 4. Результати роботи методу при $\sigma_2 = 0.05$ та зміні значень $\sigma_1 (\sigma_1 \in [0.1, 0.29], \Delta\sigma_1 = 0.01)$: а) MAPE, %; б) RMSE

Моделювання роботи відомих методів відбувалося із використанням авторського програмного забезпечення. Результати порівняння подано у таблиці.

Як видно з таблиці, розроблений метод забезпечує найточніший результат серед усіх розглянутих у роботі методів.

Висновки. У роботі розроблено ансамбль з двох нейронних мереж узагальненої регресії для розв'язання задач прогнозування підвищеної точності. Описано ба-

зові положення процедури функціонування нейронної мережі узагальненої регресії. Розглянуто складові похибки формування вихідного сигналу нейромережею цього типу. На основі цього, аналітично доведено можливість апроксимації і часткового усунення методичної похибки роботи GRNN. Подано загальну алгоритмічну реалізацію розробленого методу.

Моделювання роботи ансамблю відбувалося з використанням реальних даних. Подано розв'язок задачі за-

повнення пропущених значень у наборах даних моніторингу складу повітряного середовища. Експериментальним шляхом встановлено ефективність застосування розробленого методу під час розв'язання цієї задачі. Здійснено порівняння роботи розробленого методу з низкою відомих. Встановлено найвищу точність роботи розробленого методу як на основі середньої абсолютної похибки у відсотках так і з використанням середньоквадратичної похибки.

Таблиця. Порівняння ефективності розробленого методу з ефективністю за наявних методів

№ з/п	Метод	Базові параметри	MAPE, %	RMS E
1	Розроблений метод	$\sigma_1 = 0.24$, $\sigma_2 = 0.05$	18.689	0.461
2	Апроксимація поліномом Вінера на основі Стохастичного Градієнтного спуску (Izonin et al., 2019b)	тип функції втрат – "elipson intensive", степінь полінома = 2	22.119	0.507
3	Нейронна мережа узагальненої регресії (Izonin et al., 2019)	$\sigma_2 = 0.1$	23.00	0.483
4	Модифікований алгоритм AdaBoost (Izonin et al., 2019a)	максимальна глибина дерева = 4, кількість дерев = 300, степінь полінома = 2	27.011	0.537

Перелік використаних джерел

- Bodianskyi, Ye. V., Deineko, A. O., Zhemova, P. Ye., Zolotukhin, O. V., & Khaustova, Ya. V. (2017). Consecutive nuclear fuzzy clustering of large data sets based on hybrid computational intelligence system. (Ser. Information Systems and Networks). *Bulletin of the National University of Lviv Polytechnic*, 872, 20–24. [In Ukrainian].
- Deineko, A. O., Bodianskyi, Ye. V., & Pliss, I. P. (2012). Combined learning of the evolutionary neuro-phase system. *Radioelektronika. Informatyka. Upravlinnia*, 1, 86–92. [In Ukrainian].
- Dubrovyn, V. Y., Subbotyn, S. A., Bohuslaev, A. V., & Yatsenko, V. K. (2003). *Intelligent means of diagnostics and prediction of reliability of aircraft engines*. Zaporizhzhia: OAO "Motor-Sych", 279 p. [In Russian].
- Ivanets, O. B., Bukrieva, O. V., & Dvornik, M. V. (2011). Construction of prediction models using artificial neural networks. *Elektronika ta systemy upravlinnia*, 4(30), 139–142. [In Ukrainian].
- Izonin, I., Greguš, M., Tkachenko, R., Logoida, M., Mishchuk, O., & Kynash, Yu. (2019b). SGD-based Wiener Polynomial Approximation for Missing Data Recovery in Air Pollution Monitoring Dataset. *Lecture Notes in Computer Science*, 11506, 781–793.
- Izonin, I., Kryvinska, N., Tkachenko, R., & Zub, K. (2019a). An approach towards missing data recovery within IoT smart system. *Procedia Computer Science* (in print).
- Izonin, I., Kryvinska, N., Vitynskyi, P., Tkachenko, R., & Zub, K. (2019). GRNN Approach Towards Missing Data Recovery between IoT Systems. *Advances in Intelligent Systems and Computing*, 1035, 445–453.
- Kaczor, S., & Kryvinska, N. (2013). It is all about Services – Fundamentals, Drivers, and Business Models. *Journal of Service Science Research*, 5(2), 125–154.
- Mishchuk, O. S., & Tkachenko, R. O. (2019). Methods of processing and filling of missing parameters in ecological monitoring data. *Scientific Bulletin of UNFU*, 29(6), 119–122. <https://doi.org/10.15421/40290623>
- Molnár, E., Molnár, R., Kryvinska, N., & Greguš, M. (2014). Web Intelligence in practice. *Journal of Service Science Research*, 6(1), 149–172.
- Repository. (2019). *UCI Machine Learning Repository: Air Quality Data Set*. Retrieved from: <http://archive.ics.uci.edu/ml/datasets/air+quality>. [Accessed: 09.09.2019].
- Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*, 2(6), 568–576.
- Terekovskiy, A. I. (2011). Optimization of neural network architecture for diagnostics of computer network status. *Upravlinnia rozvytkom skladnykh system*, 6, 155–158. [In Ukrainian].
- Vito, S. De, Massera, E., Piga, M., Martinotto, L., & Francia, D. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2), 750–757.

P. B. Vitynskiy, R. O. Tkachenko, I. V. Izonin

Lviv Polytechnic National University, Lviv, Ukraine

GRNN ENSEMBLE FOR INCREASING THE ACCURACY OF REGRESSION TASKS

The effective solution of regression problems is an important task for e-commerce, medicine, business analytics and for many other industries. In recent years, the demand to use artificial intelligence methods for solving regression problems has been rapidly increased. This can be explained by the need to work with large datasets, or the complex interrelationships between multiple independent variables. General regression neural network is one of the options for solving this problem. However, the use of this computational intelligence method does not provide high accuracy of the result, which imposes a number of limitations. The new method based on a general regression neural network ensemble for increasing prediction task accuracy is developed. The main advantages and disadvantages of neural networks of this type are described in detail. A brief description of the operation of the general regression neural network is given. An algorithmic implementation of the developed ensemble is provided. An increased prediction accuracy using developed ensemble has been received. A software solution for the implementation of the described method with use libraries of Python programming language is developed. Experimental modeling of the method is conducted on real data of the regression problem. High efficiency of solving the problem is established using the developed method on the basis of both the mean absolute error in percentage and using the standard error. The method is compared with the existing ones: the Wiener polynomial approximation based on Stochastic Gradient Descent, the general regression neural network, and the modified AdaBoost algorithm. The highest accuracy of the solution of the problem by the developed method is proved experimentally based on both indicators of accuracy among all the methods considered in the work. In particular, it provides accuracy more than 3.4 %, 4.3 % and 8.3 % (MAPE) compared to existing methods. The method developed can be used to obtain high-precision solutions for solving applications of e-commerce, medicine, materials science, business analytics and others. The plan for further researches is to develop a hybrid high-speed computational intelligence system based on the combination of the developed method and the successive geometric transformations model (SGTM).

Keywords: regression; generalization properties; increase of accuracy; generalized regression neural networks.