

М. М. Герцюк

Державна установа «Інститут геохімії навколишнього середовища НАН України», Київ, Україна
Державний університет телекомунікацій, Київ, Україна

МОДЕЛЬ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ НАСЛІДКІВ ЗАБРУДНЕННЯ ВОДОЙМ РІЧОК З ВИКОРИСТАННЯМ НЕЙРОННОЇ МЕРЕЖІ, ЩО БАЗУЄТЬСЯ НА ОСНОВІ ЗАДАЧ РЕГРЕСІЇ

Анотація. У статті описаний метод коригування результатів прогнозування забруднення водойм річок, як частина інтелектуального опрацювання результатів, основаних на серії емпіричних гідрологічних рівнянь Харві Джобсона. Обґрунтовані основні принципи, логіка функціонування та процес контекстуальної валідації методу. Для вирішення поставленої мети використані можливості нейронної мережі із застосуванням задачі регресії, як метод визначення коефіцієнту, що коригує основний результат на основі похибки. В основі методу коригування похибки лежить накладення коефіцієнту на деякий результат значення характеристики піку концентрації, виведений для конкретної точки. Моделлю передбачено використання серії емпіричних гідрологічних рівнянь Харві Джобсона, як основного методу прогнозування наслідків забруднення водойм, що є емпіричним та не вимагає деталізованих вхідних даних русла для проведення обчислень. Даний метод є складовою інтелектуальної обробки результату, а не проведення прогнозувань. З використанням нейронної мережі, що функціонує на основі задачі регресії розроблений метод коригування похибки результатів прогнозування забруднення водойм. Визначена можливість використання нейронної мережі з іншими методами, що обчислюють пік концентрації в певній точці в конкретний час.

Ключові слова: оцінка якості водойм, нейронна мережа, русло річки, регресія, прогнозування наслідків забруднення водойм, математичне моделювання.

Вступ

Питання прогнозування наслідків забруднення якості водойм русла річок небезпечними та токсичними речовинами розглядаються в існуючих програмних моделях.

Зазвичай, дані рішення базуються на емпіричних та фундаментальних підходах. В переважній кількості випадків, емпіричні методи створюються при врахуванні обмежень вхідних та вихідних даних. Етап калібрування моделі відбувається на реальних річках та ситуаціях. В результаті, в модель можуть бути введені деякі емпіричні коефіцієнти, що обчислюються при дослідженні закономірності між різними характеристиками. Тому, перевагою емпіричних методів є робота в заданих умовах, на відміну від фундаментальних, що моделюють процеси максимально деталізовано.

Такими емпіричними методами є математична модель, описана в роботі Д.П. Лукаса та Е. ван Біка [0], автоматизовані моделі SWToolbox [2], Visual Plumes [3], MMSOILS [4], RESRAD [4], MEPAS [4].

Деякі методи були розроблені на основі принципів нейронних мереж та машинного навчання. Такі моделі описуються в роботах Yafra Khan, Chai Soo See [5], Md. Saikat Islam Khanad, Nazrul Islam [6], Ali Najah Ahmeda, Faridah Binti Othman [7], Saber Kouadri, Ahmed Elbeltagi, Abu Reza Md. [8] та інші.

Зазвичай, такі моделі є достатньо точними у своїх розрахунках та успішно розробляються для заданих умов. Однак, ці моделі вирішують конкретні, вузькі завдання.

При використанні емпіричних методів існує можливість створення деякої закономірної похибки, що

виникає в результаті використання моделі на конкретному місці річки.

Вирішенням даної проблеми може бути коригуючий результат коефіцієнт, що виводиться для конкретної точки на річці. Таке значення може бути вирішене шляхом створення та навчання нейронної мережі.

Даною моделлю передбачено використання серії емпіричних гідрологічних рівнянь Харві Джобсона [9], як основного методу прогнозування наслідків забруднення водойм, що є емпіричним та не вимагає деталізованих вхідних даних русла для проведення обчислень. Звісно, точність результатів при використанні моделі в визначеному місці русла річки може мати закономірну похибку. Для даної моделі є пікова концентрація (C_p) в певному місці, дистанція (D), час (T_p), через який концентрація набуде свого піку в певній точці водойма річки.

Мета статті. Для вирішення проблеми необхідно розробити нейронну мережу, що коригує основний результат C_p на основі вивченої похибки. Даний метод є методом складової інтелектуальної обробки результату, а не проведення прогнозувань. В основі архітектури нейронної мережі закладена задача регресії, як метод обчислення значення коефіцієнту.

Метод дослідження

В основі методу коригування похибки лежить накладення коефіцієнту на деякий результат значення характеристик піку концентрації C_p , виведений для конкретної точки.

Таким чином, справедливою є така формула:

$$C_{pCor} = C_p * k,$$

де Cp – пік концентрації в конкретній точці, k – виведений коефіцієнт похибки для конкретної точки, Cp_{cor} – кориговане, або очікуване значення піку концентрації в конкретній точці.

Значення Cp є вхідним значенням до даного методу.

Коефіцієнт k може виводитись шляхом навчання проєктованої нейронної мережі.

При розробці такої нейронної мережі беруться до уваги наступні параметри:

- фактор локації, як комплекс можливих факторів, що впливають з особливостей самої точки;
- фактор пори року, як комплекс факторів, що впливають від погодних умов що характерні для конкретного сезону року.

Також, можливе врахування фактору речовини, як додаткового комплексу, що включає визначення фактору закономірності похибки від речовини та впливає на значення похибки. Однак, такий підхід ускладнює логіку нейронної мережі, що означає необхідність формування обширного масиву ситуацій, як для кожної точки та пори року, так і для кожної речовини. Формування такої бази може бути недоречним через можливі обмеження в механізмах збору даних або фактичну відсутність інформації, щодо наявності певних речовин в конкретній точці.

Вхідними даними для навчання нейронної мережі є коефіцієнт похибки, виведений для конкретної ситуації в заданій точці в певну пору року. Формула для обчислення необхідного значення має наступний вигляд:

$$k_{cor} = \frac{Cp}{Cp_{cor}},$$

де k_{cor} - коефіцієнт похибки для конкретної ситуації в заданій точці в певну пору року. Випадок $k_{cor} = 1$ означає відсутність необхідності коригування похибки. На практиці такий випадок є достатньо рідким. Значення піку концентрації Cp_{cor} означає реальну величину зазначеної характеристики для конкретної ситуації в заданій точці. Характеристика Cp є очікуваним значенням, отриманим в результаті проведення обчислень.

Таким чином, формується база даних для навчання, що включає такі характеристики:

- умовний ідентифікатор точки русла річки;
- очікуване значення піку концентрації, отримане в шляхом проведення обчислень;
- реальне значення піку концентрації;
- дата настання надзвичайної ситуації.

Структуруючи цю базу знань, отримуємо серію масивів, що групуються за двома ознаками:

- точка обчислення;
- пора року.

Таким чином, ідеальним результатом є 4 масиви даних для конкретної точки.

Інформація містить значення наступних характеристик:

- очікуване значення піку концентрації;
- реальне значення піку концентрації.

Вищеописана інформація дозволяє розробити архітектуру нейронної мережі.

Вхідними характеристиками є:

- X_1 – фактор локації;
- X_2 – фактор пори року.

Дані змінні можуть приймати значення 0 або 1, що означає необхідність врахування даного фактору при роботі нейронної мережі.

Перцептрон, що працює з описаною метою при описаних обмеженнях показано на рис. 1.

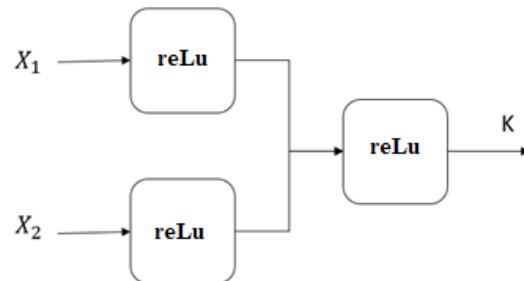


Рис. 1. Перцептрон для виведення коефіцієнту похибки

В основі даної нейронної мережі лежить задача регресії. Нейронна мережі є двошаровою. Перший шар приймає рішення, ґрунтуючись на окремій характеристиці. Другий – сумує значення, отримане від попередніх характеристик та приймає своє рішення, тобто виводить коефіцієнт похибки. В основі функції, що приймає рішення лежить ReLU [10]. Така функція активації має наступну формулу:

$$W = \max(0, w * x),$$

де x – деяка характеристика, w – деяка вага, тобто значимість даного фактору, W – результат рішення функції активації.

Таким чином, результатом нейрона першого рівня є деяке значення коефіцієнту w або 0 – що означає, що даний фактор не потрібно враховувати у випадку, якщо $x = 0$.

Нейрон другого рівня, що обчислює коефіцієнт регресії в основі функції активації, також використовує ReLU. Приймаючи рішення на основі результату шарів першого рівня шляхом використання суматора, що описаний наступною формулою:

$$W_n = \sum W.$$

Результуючим значенням, а отже коефіцієнтом похибки є наступна функція:

$$k = \sum W_n.$$

Зазначені коефіцієнти w_1 та w_2 для факторів мають два шляхи виведення власних значень:

- отримання значень на основі експертної оцінки;
- навчання нейронної мережі. В цьому випадку, нейронна мережі розширяється додатковим функціоналом коригування власних результатів, тобто навчанням. Очевидним методом, що дозволяє проводити дану операцію є метод оберненого розповсюдження помилки.

Таким чином, загальна архітектура нейронної мережі має схему, що показано на рис. 2.

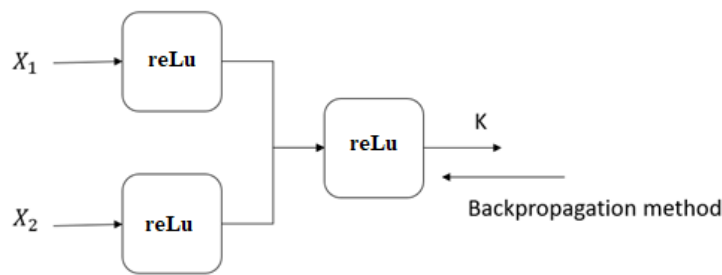


Рис. 2. Загальна архітектура нейронної мережі для виведення коефіцієнту похибки

Валідація результатів

Точність роботи даного методу передбачає правильне визначення коефіцієнту на основі існуючих даних. На основі заданої інформації, нейронна мережа має згенерувати значення, близьке до очікуваного. Таким чином, база даних має містити набір значень, що зберігають основну закономірність локації та пори року. Аномальні, або нетипові значення не варто застосовувати при роботі з даним методом, так як існує ризик збільшення значення похибки і, відповідно, зменшення ефективності роботи мережі. Таким чином, масив даних для навчання варто попередньо аналізувати та фільтрувати за необхідності.

Мережа потребує 3 масиви даних:

- тестовий – масив, що підтверджує працездатність мережі;
- тренувальний – масив, що сформовано на основі попередньо підібраних значень для реальних точок. Це дає можливість оцінки ефективності алгоритму;
- кінцевий – масив, сформований на основі визначених даних у реальних ситуаціях. Обчислені коефіцієнти повинні застосовуватись для реальних ситуацій.

Цей підрозділ має на меті доведення правильності роботи методу. Тому, існує необхідність в розробці тестового масиву даних, на основі якого мережа має сформувати очікуваний результат.

Нехай, еталонним коефіцієнтом є 0.8. Створимо 5 умовних точок, що мають 5 тестових значень. Для тренування достатньо використати 10000 епох. Масив значень згенерованих значень для характеристик реальних та обчислених характеристик описано у табл. 1-5.

Етапи проведення навчання нейронної мережі за заданими тестовими масивами показали результат, показаний в табл. 6.

Визначенні результати мають достатньо близькі значення до еталонного коефіцієнту. В заданому масиву можна виділити результат, що було виведено для 5-ї точки, що близьких до значення 0,76.

Таблиця 1 – Значення для точки № 1

| Обчислений пік концентрації | Реальний пік концентрації |
|-----------------------------|---------------------------|
| 28,71 | 22,968 |
| 31,98 | 25,584 |
| 360 | 288,024 |
| 14,35 | 11,48 |
| 0,58 | 0,464 |

Таблиця 2 – Значення для точки № 2

| Обчислений пік концентрації | Реальний пік концентрації |
|-----------------------------|---------------------------|
| 0,8 | 0,64 |
| 0,98 | 0,784 |
| 9,66 | 7,728 |
| 0,41 | 0,328 |
| 0,016 | 0,0128 |

Таблиця 3 – Значення для точки № 3

| Обчислений пік концентрації | Реальний пік концентрації |
|-----------------------------|---------------------------|
| 0,71 | 0,568 |
| 0,91 | 0,728 |
| 8,57 | 6,856 |
| 0,38 | 0,304 |
| 0,015 | 0,012 |

Таблиця 4 – Значення для точки № 4

| Обчислений пік концентрації | Реальний пік концентрації |
|-----------------------------|---------------------------|
| 0,255 | 0,204 |
| 0,33 | 0,264 |
| 3,08 | 2,464 |
| 0,13 | 0,104 |
| 0,05 | 0,04 |

Таблиця 5 – Значення для точки №5

| Обчислений пік концентрації | Реальний пік концентрації |
|-----------------------------|---------------------------|
| 0,174 | 0,1392 |
| 0,21 | 0,168 |
| 2,09 | 1,672 |
| 0,09 | 0,072 |
| 0,03 | 0,024 |

Таблиця 6 – Значення коефіцієнту похибки для різних точок

| № точки | Коефіцієнт похибки |
|---------|--------------------|
| 1 | 0,799974132805534 |
| 2 | 0,7951551855680876 |
| 3 | 0,7965094022097843 |
| 4 | 0,7933086155135414 |
| 5 | 0,7627572392513363 |

Така поведінка може означати надмірну недостатню кількість епох для навчання.

Однак, підвищуючи кількість епох можна отримати результат надмірної чутливості нейронної

мережі до будь-яких змін. Реальні результати, що мають деяку «дисперсію» значень можуть створити дану ситуацію, що приведе до меншого значення похибки для одних точок та до більшого значення для інших точок. Таким чином, правильність роботи методу може бути порушена.

Масив реальних результатів має бути навченим за наступними етапами:

1. групування масиву даних за точками;
2. групування масивів даних за порами року.

Таким чином, для кожної точки повинно обчислитись 4 результати. У випадку відсутності даних для точки, заданий коефіцієнт має значення 1. Такий підхід дозволить відобразити обчислений результат без урахування похибки. При використанні графічного виведення результату, аналіз діаграми дозволить дослідити ситуацію в цілому, спираючись на кориговані результати сусідніх точок.

Підсумовуючи вищеписані результати та обмеження можна зробити висновок, що для моделі математичного моделювання наслідків забруднення водним нейронна мережа виконує поставлені завдання.

Висновки

Таким чином, використовуючи нейронну мережу, що функціонує на основі задачі регресії розроблено метод коригування похибок результатів прогнозування забруднення водним.

Зокрема, результатом є коригування піку концентрації в певній точці в конкретний час.

Метод використовується, як частина інтелектуальної обробки результату, що працює в парі з серією емпіричних гідрологічних рівнянь Харві Джобсона.

Також, існує можливість використання описаної нейронної мережі з іншими методами, що обчислюють пік концентрації в певній точці в конкретний час.

СПИСОК ЛІТЕРАТУРИ

1. Loucks D.P., van Beek E. Water Resource Systems Planning and Management, 2017. DOI:10.1007/978-3-319-44234-1.
2. SWToolbox: A Surface-Water Toolbox for Statistical Analysis of Streamflow Time Series, Chapter 11 of Section A, Statistical Analysis, Book 4, Hydrologic Analysis and Interpretation, Virginia, USA, 2018.
3. Frick W.E., Roberts P.J.W., Davis L.R., Keyes J., Baumgartner D.J., George K.P. Dilution Models for Effluent Discharges. 4-th Edition (Visual Plumes). Athens, Georgia, 2003, pp. 148. [Online]. Available: <https://www.epa.gov/sites/production/files/documents/VP-Manual.pdf>.
4. Laniak G. F., Droppo J. G., Faillace E.R., Gnanapragasam E.K., Mills W.B., Strenge D.L., Whelan G., Yu C. An Overview of a Multimedia Benchmarking Analysis for Three Risk Assessment Models: RESRAD, MMSOILS, and MEPAS. Risk Analysis. [Online]. Available: <https://pubmed.ncbi.nlm.nih.gov/9202489/>.
5. Yafra K., Chai S. S., Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model. 2016 IEEE Long Island Systems, Applications and Technology Conference. 29 April 2016. Farmingdale, NY, USA. DOI: 10.1109/LISAT.2016.7494106. Available: <https://ieeexplore.ieee.org/document/7494106>.
6. Md. Saikat I. K., Islam N., Uddin J. et al. Water quality prediction and classification based on principal component regression and gradient boosting classifier approach. Journal of King Saud University. Computer and Information Sciences. 2021. Riyadh, Saudi Arabia. DOI: 10.1016/j.jksuci.2021.06.003. Available: <https://www.sciencedirect.com/science/article/pii/S1319157821001361>.
7. Ali Najah Ahmeda, Faridah Binti Othman, Haitham Abdulmohsin Afan, Rusul Khaleel Ibrahim, Chow Ming Fai, Md Shabbir Hossain, Mohammad Ehteram, Ahmed Elshafie. Machine learning methods for better water quality prediction. Journal of Hydrology, Volume 578, November 2019. DOI: 10.1016/j.jhydrol.2019.124084. Available: <https://www.sciencedirect.com/science/article/abs/pii/S0022169419308194?via%3Dihub>.
8. Kouadri, S., Elbeltagi, A., Islam, A.R.M.T. et al. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl Water Sci, 11, 2021. DOI: 10.1007/s13201-021-01528-9. Available: <https://doi.org/10.1007/s13201-021-01528-9>.
9. Jobson H. E. Prediction of Travel time and Longitudinal Dispersion in Rivers and Streams, 1996, pp. 69.
10. A Gentle Introduction to the Rectified Linear Unit (ReLU). Available: <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>.

Received (Надійшла) 22.02.2022

Accepted for publication (Прийнята до друку) 06.04.2022

A mathematical modeling model of river's water pollution consequences with use of neural network, which based on regression problem

M. Gertsyuk

Abstract. This article describes a method of river's water pollution forecast results adjusting, as a part of results intellectual processing, based on series of Harvey Jobson empirical hydrological equations. Basic principles, functioning logic and a process of contextual methods validation were reasoned. The capabilities of the neural network using the regression problem as a method of determining the coefficient that corrects the main result based on error were used to solve this goal. The method of error correction is based on the imposition of a coefficient on some result of peak concentration characteristic value, measured for a particular point. The model provides the use of series of empirical hydrological Harvey Jobson equations, as main method for forecasting the effect of water pollution, which has empirical nature, and doesn't require detailed input data for providing measurements. This method is part of the intellectual result processing, rather than making predictions. With use of neural network, which functions on regression problem, method of forecasting water pollution results error correcting. A possibility of neural network use with another methods, which measures peak concentration in certain location on specific time.

Keywords: water quality assessment, neural network, riverbed, regression, forecasting the consequences of water pollution, mathematical modeling.