УДК 519.85

О.І. Чумаченко, к.т.н., доц.
В.С. Горбатюк, аспір.
Національний технічний університет України «КПІ»

КОМПЛЕКСУВАННЯ ДЕКІЛЬКОХ АЛГОРИТМІВ ПІД ЧАС РОЗВ'ЯЗАННЯ ЗАДАЧІ ПРОГНОЗУВАННЯ

Задача прогнозування є, безсумнівно, однією з найважливіших для людства, але водночас стоїть поряд з найскладнішими, оскільки немає жодної гарантії, що поведінка прогнозованого процесу не зміниться кардинально у майбутньому.

Метою роботи є розробка нового методу прогнозування, що використовує підхід комплексування декількох моделей, та перевірка його якості на наборі тестових даних. Для цього були розглянуті основні існуючі методи прогнозування, а саме: штучні нейронні мережі, метод групового урахування аргументів та лінійна регресія. Для комплексування прогнозуючих моделей був обраний підхід bagging.

Запропонований метод було реалізовано у програмному середовищі Matlab і перевірено (разом з декількома існуючими методами) на 11 тестових наборах даних. Наукова новизна та практична значущість дослідження полягає в тому, що серед методів, що тестувалися, запропонований метод показав найкращі результати, що свідчить про можливість його успішного застосування на практиці.

Ключові слова: *прогнозування часових рядів; комплексування моделей; штучні нейронні мережі; метод групового урахування аргументів.*

Постановка проблеми у загальному вигляді. Підхід комплексування декількох алгоритмів успішно застосовується для вирішення великого спектра задач машинного навчання протягом багатьох років. Дійсно, поєднання одразу декількох методів зазвичай дозволяє досягти кращого результату, ніж при використанні кожного методу окремо.

Більшість методів комплексування можна поділити на 2 групи: ті, що використовують підхід bagging [1, с. 123–140], і ті, що використовують підхід boosting [2, с. 512–518].

Підхід bagging передбачає генерацію m підвибірок розміру n' із вихідної вибірки шляхом семплювання із заміною. Після цього кожна модель, що комплексується, навчається на окремій підвибірці, а прогноз для нових даних виконується шляхом усереднення виходів усіх моделей з певними усереднюючими коефіцієнтами.

Згідно з підходом boosting, кожна модель навчається на повній вибірці, але при цьому навчається корегувати помилки усіх попередніх моделей – тобто моделі навчаються по черзі, і навчальні приклади, на яких поточний набір моделей помиляється найбільше, будуть мати найбільшу вагу при навчанні наступної моделі.

Постановка задачі прогнозування. Нехай задано n значень часового ряду: $\vec{x} = [x_1, \ldots, x_n]$. Тоді задача прогнозування, що розглядається у цій роботі, полягає в побудові деякої моделі прогнозованого об'єкта, яка б залежала від k послідовних значень часового ряду $[x_i, \ldots, x_{i+k}]$ та видавала значення прогнозу для значення x_{i+k+t}: $F(x_i, \ldots, x_{i+k}) = \hat{x}_{i+k+t}$.

Викладення основного матеріалу. Штучні нейронні мережі (ШНМ) [3, с. 386–408]. ШНМ є системою з'єднаних і взаємодіючих між собою штучних нейронів. ШНМ не програмуються в звичному сенсі цього слова, вони навчаються. У процесі навчання нейронна мережа здатна виявляти складні залежності між вхідними та вихідними даними, а також виконувати узагальнення. Здібності нейронної мережі до прогнозування безпосередньо випливають з її здатності до узагальнення і виділення прихованих залежностей між вхідними і вихідними даними. Після навчання мережа здатна прогнозувати майбутні значення деякої послідовності на основі декількох попередніх значень і/або якихось існуючих зараз чинників [4, с. 402–409; 5, с. 47–55; 6, с. 495–506]. Зазвичай, для навчання ШНМ використовують певну модифікацію алгоритму зворотного поширення похибки [7, с. 533–536], головні рівняння якого мають вигляд:

$$\frac{\partial E}{\partial y_i \in l} = \sum_{y_j \in l+1} \frac{\partial y_j}{\partial y_i} * \frac{\partial E}{\partial y_j}, \tag{1}$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial y_j}{\partial w_{ij}} * \frac{\partial E}{\partial y_j}, \tag{2}$$

тобто:

- похідна функції помилки за виходом певного нейрону y_i, що знаходиться у шарі за номером l, дорівнює сумі за всіма нейронами y_j наступного шару $l+1$ похідних виходу цього нейрону за виходом y_i помноженим на похідну функції помилки за виходом y_j;
- похідна функції помилки за певним ваговим коефіцієнтом w_{ij} дорівнює похідній виходу нейрону, що є вхідним для зв'язку w_{ij}, за цим ваговим коефіцієнтом, помноженій на похідну функції помилки за виходом цього нейрону.

Метод групового урахування аргументів (МГУА) [8, с. 43–53] – це набір алгоритмів прогнозування (а точніше математичного моделювання), який ґрунтується на розбитті вихідних даних на дві вибірки: навчальну і перевірочну, і використанні опорних функцій деякого вигляду, параметри яких знаходяться на навчальній вибірці, а перевірка того, наскільки добре вони моделюють заданий ряд, виконується на перевірочній вибірці. Зазвичай, як опорні функції використовують поліному Колмогорова–Габора певної складності:

$$y(x_1,...,x_n) = a_0 + \sum_i a_i x_i + \sum_i \sum_j a_{ij} x_i x_j + ... \tag{3}$$

Лінійна регресія з динамічними вагами (ЛРДВ) [9, с. 4–8]. Даний метод є «вдосконаленням» звичайної лінійної регресії [10, с. 169–174], що дозволяє знаходити декілька наборів ваг лінійної моделі для певних підвибірок (причому ці ваги знаходяться з урахуванням не лише прикладів відповідної підвибірки, але й усіх прикладів вибірки) і подальшого прогнозування шляхом усереднення усіх ваг, де коефіцієнти усереднення залежать від близькості прикладів до відповідної підвибірки (кластеру). Вектори ваг для підвибірок знаходяться з використанням наступних функцій помилок:

$$E_k = \alpha * \left[\sum_{x_i \in X_k} \left(\sum_j w_{kj} * x_{ij} - y_i \right)^2 \right] + \beta * \sum_j (w_{kj} - w_j^{global})^2 , \tag{4}$$

де $\vec{w}^{global}$ – вагові коефіцієнти регресії, що знайдені на усій вибірці X; X_k – підвибірка під номером k; $\vec{w}_k$ – вектор ваг для підвибірки під номером k; $\alpha, \beta > 0, \alpha + \beta = 1$ – коефіцієнти, що регулюють наскільки близько до вектора ваг $\vec{w}^{global}$, знайденого на усій вибірці, повинний бути вектор ваг для підвибірки за номером k та наскільки «сильно» потрібно зменшити помилку регресії для підвибірки k при використанні окремого вектора ваг.

У цій роботі для вирішення задачі прогнозування пропонується використання підходу bagging для комплексування моделей, отриманих за допомогою 3 методів: ШНМ, МГУА та ЛРДВ. Для оцінки усереднюючих коефіцієнтів моделей використаємо зовнішній критерій, а саме – помилку моделі на окремій, валідаційній вибірці.

Таким чином, алгоритм побудови прогнозуючої моделі має наступний вигляд:

1. Попередня обробка даних.
2. Генерація підвибірок. Згідно з підходом bagging, генеруються 3 підвибірки розміру $n' \le n$ шляхом семплювання з заміною.
3. На кожній підвибірці навчається модель з використанням певного методу: ШНМ, МГУА чи ЛРДВ.
4. Знаходяться усереднюючі ваги. Для цього спочатку розраховуються середньоквадратичні помилки кожної моделі на валідаційній вибірці.

Після цього усереднюючі коефіцієнти моделей розраховуються як:

$$\begin{aligned} &\vec{e} = [e_1, e_2, e_3], \\ &e_i = \sum_{j=1}^{m} (F_i(\vec{x}_j^{(v)}) - y_j^v)^2, \\ &\alpha_i = \frac{1}{2} * (1 - \frac{e_i}{\sum_i e_i}), i = 1,2,3. \end{aligned} \tag{5}$$

5. Отже, «фінальна» прогнозуюча модель буде мати вигляд:

$$F(\vec{x}) = \sum_{i=1}^{3} \alpha_i * F_i(\vec{x}). \tag{6}$$

Для перевірки запропонованого методу було використано 11 наборів даних, що знаходяться у відкритому доступі в мережі Інтернет ([11, с. 1; 12, с. 2]). Нормалізована середньоквадратична помилка методу порівнювалась з відповідними помилками таких методів, як МГУА, ШНМ та ЛРДВ. Під час навчання усім методам використовувались наступні спільні параметри:

- розмірність вкладення часових рядів $\Leftrightarrow$ кількість вхідних змінних $m = 5$;
- прогнозування виконувалось на 2 значення наперед;
- у навчальну вибірку відбиралось 50 % усіх прикладів.

Для оцінки якості моделі використовувалася нормалізована середньоквадратична помилка, що розраховувалась на усій вихідній вибірці:

$$E = \frac{\sum_j (F(x_{j1},...,x_{jn}) - y_j)^2}{\sum_j y_j^2} \qquad (7)$$

Отримані результати наведено у таблиці 1.

Таблиця 1

Нормалізована середньоквадратична помилка моделей на усій вибірці

Вибірка	ШНМ	ЛРДВ	МГУА	Запропонований метод
Виробництво електроенергії в Австралії	0.1148	0.0804	0.0575	0.0237
Тест CATS (Competition on Artificial Time Series – Конкурс з штучних часових рядів) [13, с. 1615–1620]	0.1321	0.1092	0.0785	0.0336
Курс долару до євро	0.1589	0.2495	0.1270	0.2693
Курс долару до фунту	0.2920	0.2496	0.0840	0.0516
Індекс CPI (індекс споживчих цін)	0.1867	0.1477	0.1533	0.1337
Споживання електроенергії в Іспанії	0.1314	0.0875	0.0603	0.0289
Середні відсоткові ставки в Іспанії	0.1432	0.1390	0.0905	0.0362
Індекс фондової біржі в Іспанії	0.1153	0.1610	0.1358	0.1293
Кількість плям на сонці	0.2273	0.0575	0.0525	0.0071
Виробництво літаків в США	0.1396	0.0548	0.0612	0.0192
Зимовий індекс NAO (Північно-атлантичного коливання)	0.4770	0.3368	0.1736	0.0777
Сумарна помилка	2.1182	1.6730	1.0742	0.8103

Наприклад, такий вигляд має прогноз для вибірки «Курс долару до євро», отриманий з використанням ЛРДВ:

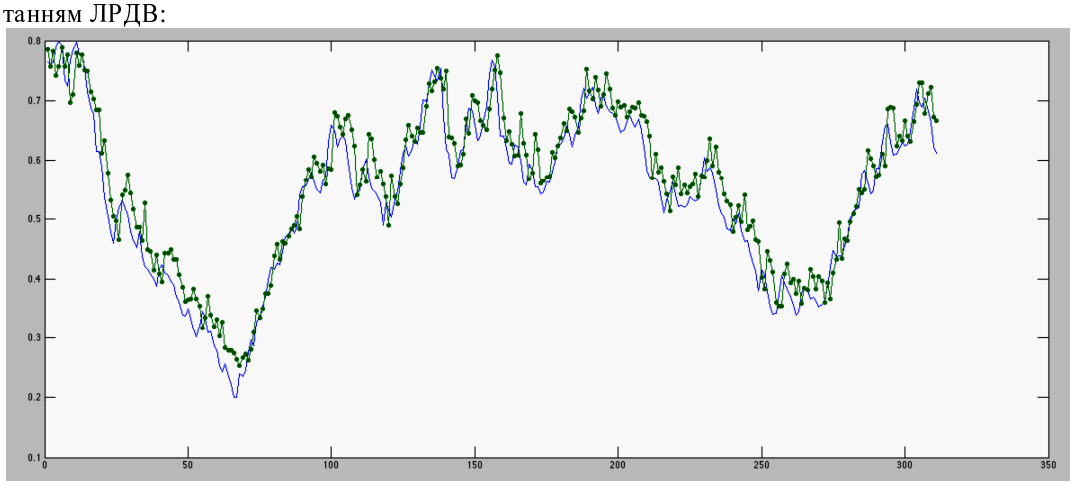

Рис. 1. Прогноз, отриманий з використанням ЛРДВ

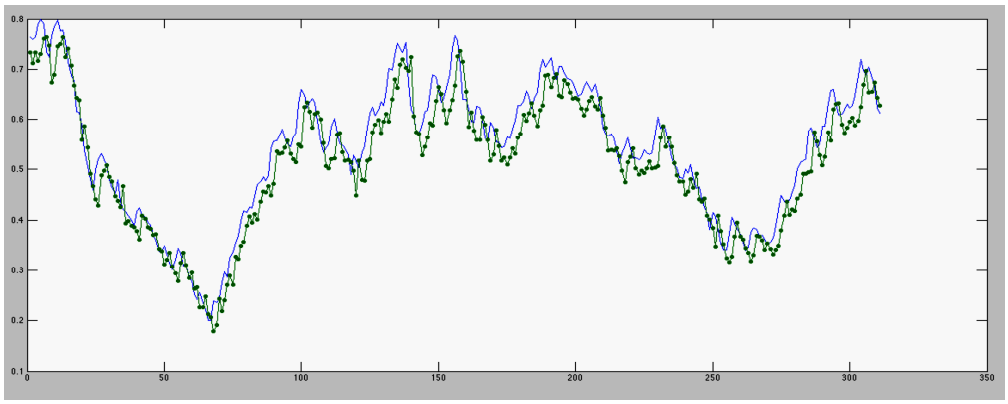

Рис. 2. Прогноз,
отриманий за допомогою застосування ШНМ

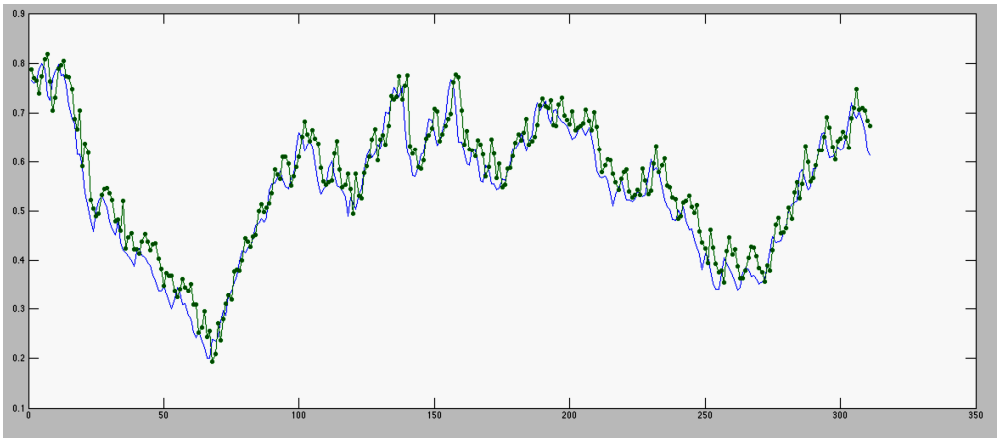

Рис. 3. Прогноз, отриманий з використанням МГУА

І, нарешті, наведений прогноз, отриманий при комплексуванні усіх 3 моделей (рис. 4).

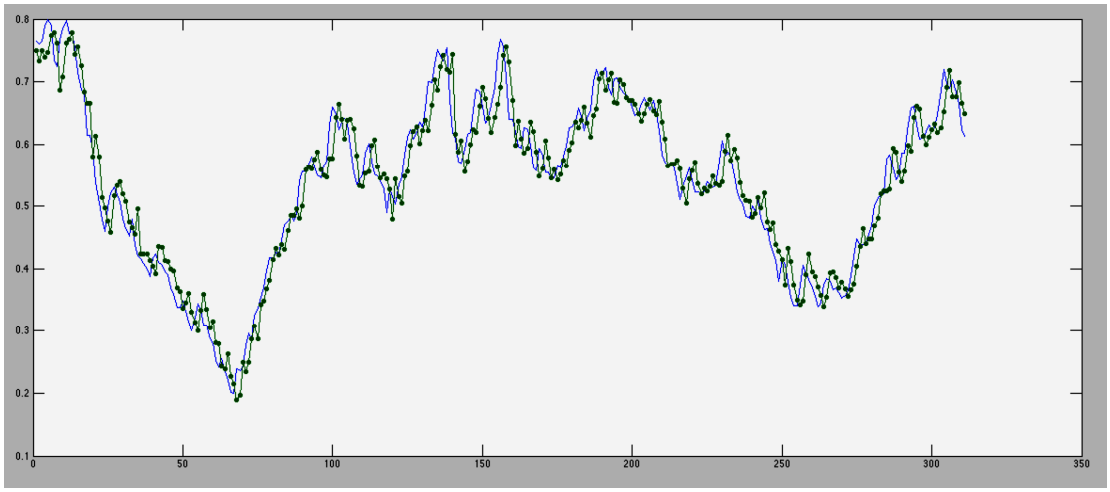

Рис. 4. Прогноз, отриманий в результаті комплексування усіх 3 моделей

Візуально помітно, що «розкид» прогнозу, отриманого за допомогою комплексування, значно менший, ніж у прогнозів, отриманих при використанні ШНМ, МГУА чи ЛРДВ поодинці.

Висновки з даного дослідження та перспективи подальших розвідок. У цій роботі запропоновано метод вирішення задач прогнозування на основі комплексування наступних алгоритмів: ШНМ, МГУА та ЛРДВ, де усереднюючі коефіцієнти моделей знаходяться на валідаційній вибірці. Поєднання цих підходів дозволяє на практиці знайти більш адекватну модель прогнозованого об'єкта. Серед методів, які

тестувалися, запропонований метод показав найкращі результати, що свідчить про можливість його успішного застосування на практиці. Перспективним напрямком подальших досліджень описаної задачі автори вважають можливість комплексування інших алгоритмів прогнозування для отримання ще точніших результатів.

Список використаної літератури:

1. *Breiman L.* Bagging predictors / *L.Breiman //* Machine Learning. – 1996. – Vol. 24, № 2. – Pp. 123–140.
2. Boosting algorithms as gradient descent function space / *L.Mason, J.Baxter, P.Bartlett, M.Frean //* Advances in neural information processing systems. – 2000. – Vol. 12. – Pp. 512–518.
3. *Rosenblatt F.* The perceptron: a probabilistic model for information storage and organization in the brain / *F.Rosenblatt //* Psychological review. – 1958. – Vol. 65, № 6. – Pp. 386–408.
4. *Gioqinang Z.* Neural network forecasting of the British pound/US dollar exchange rate / *Z.Gioqinang, M.Y. Hu //* International journal of management science. – 1998. – Vol. 26. – Pp. 495–506.
5. *Bodyanskiy Ye.* Adaptive wavelet–neuro–fuzzy network in the forecasting and emulation tasks / *Ye.Bodyanskiy, I.Pliss, O.Vynokurova //* Int. journal on information theory and applications. – 2008. – Vol. 15, № 1. – Pp. 47–55.
6. *Amir F.A.* A comparison between neural–network forecasting techniques – case study: river flow forecasting / *F.A Amir, I.S. Samir //* IEEE Transactions on neural networks. – 1999. – Vol. 10, № 2. – Pp. 402–409.
7. *Rumelhart D.E.* Learning representations by back–propagating errors / *D.E. Rumelhart, G.E. Hinton, R.J. Williams //* Nature. – 1986. – Vol. 323, № 6088. – Pp. 533–536.
8. *Stepashko V.S.* GMDH algorithms as basis of modeling process automation after experimental data / *V.S. Stepashko //* Sov. journal of automation and information sciences. – 1988. – Vol. 21, № 4. – Pp. 43–53.
9. *Sineglazov V.* A method for building a forecasting model with dynamic weights / *V.Sineglazov, O.Chumachenko, V.Gorbatiuk //* Eastern–European journal of enterprise technologies. – 2014. – Vol. 2, № 4. – Pp. 4–8.
10. *Cook R.D.* Influential observations in linear regression/ *R.D. Cook //* Journal of the American Statistical Association. – 1979. – № 74. – Pp. 169–174.
11. U.S. General Aviation Aircraft Shipments and Sales. Barr Group Aerospace & AeroWeb. – 2014 [Електронний ресурс]. – Режим доступу : http://www.bga–aeroweb.com/database/Data3/US–General–Aviation–Aircraft–Sales–and–Shipments.xls.
12. Data Sets for Time–Series Analysis. Evolutionary and Neural Computation for Time Series Prediction Mini–site. – 2005 [Електронний ресурс]. – Режим доступу : http://tracer.uc3m.es/tws/TimeSeriesWeb/repo.html.
13. Time series prediction competition: The CATS Benchmark / *A.Lendasse, E.Oja, O.Simula, M.Verleysen //* International joint conference on neural networks, Budapest (Hungary), IEEE. – 2004. – Pp. 1615–1620.

References:

1. Breiman, L. (1996), "Bagging predictors", *Machine Learning*, Vol. 24 (2), pp. 123–140.
2. Mason, L., Baxter, J., Bartlett, P. and Frean, M. (2000), "Boosting algorithms as gradient descent function space", *Advances in neural information processing systems*, Vol. 12, pp. 512–518.
3. Rosenblatt, F. (1958), "The perceptron: a probabilistic model for information storage and organization in the brain", *Psychological review*, No. 65 (6), pp. 386–408.
4. Gioqinang, Z. and Hu, M.Y. (1998), "Neural network forecasting of the British pound/US dollar exchange rate", *International journal of management science*, Vol. 26, pp. 495–506.
5. Bodyanskiy, Ye., Pliss, I. and Vynokurova, O. (2008), "Adaptive wavelet–neuro–fuzzy network in the forecasting and emulation tasks", *Int. journal on information theory and applications*, Vol. 15 (1), pp. 47–55.
6. Amir, F.A. and Samir, I.S. (1999), "A comparison between neural–network forecasting techniques – case study: river flow forecasting", *IEEE Transactions on neural networks*, Vol. 10, No. 2, pp. 402–409.
7. Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986), "Learning representations by back–propagating errors", *Nature*, Vol. 323 (6088), pp. 533–536.
8. Stepashko, V.S. (1988), "GMDH algorithms as basis of modeling process automation after experimental data", *Sov. journal of automation and information sciences*, No. 21 (4), pp. 43–53.
9. Sineglazov, V., Chumachenko, O. and Gorbatiuk, V. (2014), "A method for building a forecasting model with dynamic weights", *Eastern–European journal of enterprise technologies*, No. 2 (4), pp. 4–8.

10. Cook, R.D. (1979), “Influential observations in linear regression”, *Journal of the American Statistical Association*, No. 74, pp. 169–174.
11. Barr Group Aerospace & AeroWeb (2014), “U.S. General Aviation Aircraft Shipments and Sales”, available at: www.bga–aeroweb.com/database/Data3/US–General–Aviation–Aircraft–Sales–and–Shipments.xls (accessed 10.05.2016).
12. Data Sets for Time–Series Analysis (2005), “Evolutionary and Neural Computation for Time Series Prediction Mini–site”, available at: http://tracer.uc3m.es/tws/TimeSeriesWeb/repo.html (accessed 10.05.2016).
13. Lendasse, A., Oja, E., Simula, O. and Verleysen, M. (2004), “Time series prediction competition: The CATS Benchmark”, *International joint conference on neural networks*, IEEE, Budapest, Hungary, pp. 1615–1620.

ЧУМАЧЕНКО Олена Іллівна – кандидат технічних наук, доцент кафедри технічної кібернетики, факультет інформатики та обчислювальної техніки Національного технічного університету України «КПІ».

Наукові інтереси:
- моделювання та управління гнучкими комп'ютерними системами;
- автоматизовані системи підтримки прийняття рішень;
- штучний інтелект.

Тел.: +3800634598487.
E-mail: lobach21@mail.ru.

ГОРБАТЮК Владислав Сергійович – аспірант кафедри технічної кібернетики, факультет інформатики та обчислювальної техніки Національного технічного університету України «КПІ».

Наукові інтереси:
- штучні нейронні мережі;
- алгоритми прогнозування часових рядів;
- методи глибокого навчання;
- гібридні алгоритми.

Тел.: +3680934356892.
E-mail: vladislav.horbatiuk@gmail.com.

Стаття надійшла до редакції 06.04.2016