https://doi.org/10.31649/1997-9266-2020-153-6-76-83

УДК 004.9+556

М. В. Дратований[1]
О. М. Козачко[1]
О. Л. Мельник[1]
І. В. Варчук[1]

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОПТИМІЗАЦІЇ ПАРАМЕТРІВ АНСАМБЛЮ МОДЕЛЕЙ ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ПРОГНОЗУВАННЯ НАЯВНОСТІ ОПАДІВ ЗА ДАНИМИ МЕТЕОМОНІТОРИНГУ

[1]Вінницький національний технічний університет

Прогнозування даних — це тривіальна задача системного аналізу, існують різні види прогнозів та передбачення. Одним з них є бінарний прогноз, який відповідає на питання: «Відбудеться подія чи ні?». Одним з питань метеорології є питання прогнозування наявності опадів, а також яка точність буде у такого прогнозу.

В роботі розглянуто задачу прогнозування наявності опадів за даними метеорологічного моніторингу та запропонована інформаційна технологія оптимізації параметрів ансамблю таких моделей машинного навчання, як моделі градієнтного бустингу та логістичної регресії, що побудовані на основі набору інформативних ознак. Запропонована інформаційна технологія дозволяє виконати розвідувальний аналіз вхідних даних та визначити оптимальний набір інформативних ознак, а за рахунок алгоритму, який на кожному кроці визначає оптимальні одно-, дво-, три-, … елементні набори ознак, максимізувати точність прогнозування. Побудовано графіки впливу ознак на точність використаних моделей машинного навчання. Для кожного типу моделей визначено свій набір ознак. Для побудови інформаційної технології взято дані, надані Вінницьким центром з гідрометеорології. Це дані моніторингу атмосфери м. Вінниця за останні 10 років, які включають: температуру повітря, вологість повітря, точку роси, хмарність та швидкість вітру.

Для підвищення точності прогнозування запропоновано математичну модель, яка базується на оптимальному визначенні ваг ансамблю моделей градієнтного бустингу та логістичної регресії. Проведено експерименти, які показали достатньо точний результат. Точність запропонованої інформаційної технології показала 80 %. Це підвередило достовірність запропонованої технології.

Ключові слова: інформаційна технологія, моделі штучного інтелекту, прогнозування наявності опадів, інформативні ознаки.

Вступ

Будь-яка задача аналізу та прогнозування даних з використанням методів машинного навчання розв’язується за декілька етапів [1]:

– цензурування даних (виявлення аномальних та помилкових даних);
– розвідувальний аналіз даних;
– формування набору інформативних ознак;
– вибір оптимальної моделі прогнозування даних.

Важливою задачею метеорологів є задача прогнозування погоди, а саме прогноз наявності опадів. Це тривіальна задача для короткочасного прогнозу, але при довгостроковому прогнозі виникає велика похибка. Над проблематикою довгострокового прогнозу метеорологічних даних працюють науковці Європи, США, Китаю. Світовим лідером в цій галузі наразі є European Centre for medium-range weather forecasts (ECMWF). Але вони досліджують проблеми клімату на рівні континентів за допомогою суперкомп’ютерів. [2]—[4]. Але в Східній Європі питанню прогнозування погоди приділяють недостатньо уваги. Тому для дослідження клімату на території України, зокрема і Вінни-

цької області, виникає завдання бінарного прогнозування опадів у м. Вінниця та підвищення точності таких прогнозів. Якщо для прогнозування даних достатньо методів машинного навчання, то для розв'язання задачі підвищення точності прогнозу потрібно розробити інформаційну технологію оптимізації параметрів ансамблю моделей штучного інтелекту.

Мета дослідження — розробити метод підвищення точності прогнозування наявності опадів у м. Вінниця на основі інформаційної технології оптимізації параметрів ансамблю моделей штучного інтелекту.

Аналіз вхідних даних моделей

Вхідними даними для прогнозування наявності опадів є набір даних зі статистикою щоденних погодних умов протягом останніх 10 років в місті Вінниця, отриманих від Вінницького обласного центру з гідрометеорології. Інформацію про ознаки, на основі яких прогнозується наявність опадів подано в таблиці.

Інформація про ознаки для прогнозування наявності опадів

№	Позначення	Назва фактора	Мінімальне значення	Максимальне значення
1	d	Дата спостереження	—	—
2	avgTemp	Середня температура повітря, °C	–23,4	28
3	maxTemp	Максимальна температура повітря, °C	–19,4	37,3
4	minTemp	Мінімальна температура повітря, °C	–28,5	22,9
5	avgHumudity	Середня вологість повітря	29	100
6	minHumudity	Мінімальна вологість повітря	16	100
7	avgHumudityDef	Середній дефіцит вологості повітря	0	23,8
8	maxHumudityDef	Максимальний дефіцит вологості повітря	0	47,7
9	cloud	Оцінка загальної хмарності по шкалі від 0 до 10	0	10
10	lowCloud	Оцінка хмарності нижнього ярусу по шкалі від 0 до 10	0	10
11	maxWind	Максимальна швидкість вітру, м/с	3	28
12	precipitation	Кількість опадів за добу, мм	0	76,7

Перед застосуванням моделей штучного інтелекту необхідно виконати попередній аналіз вхідних даних, який будемо здійснювати за таким алгоритмом:

1. Розрахувати коефіцієнт варіації для кожної ознаки та виявити неінформативні ознаки, в яких значення варіації набуває досить мале значення.
2. Виявити та видалити пропуски в даних.
3. Розрахувати кореляційну матрицю з метою визначення наявності лінійного зв'язку між ознаками.
4. Здійснити масштабування всіх ознак таким чином, щоб їхні значення знаходились в діапазоні [–1, +1].

Згідно з коефіцієнтом варіації, є такі кількісні ознаки, варіація яких досить мала. Про це свідчить наявність неінформативності в даних, що можна перевірити за допомогою побудованої діаграми коефіцієнтів варіації ознак (рис. 1). Як видно з цієї діаграми існує неінформативність в даних для таких ознак як: середня вологість повітря; мінімальна вологість повітря; оцінка загальної хмарності за шкалою від 0 до 10; максимальна швидкість вітру (м/с).

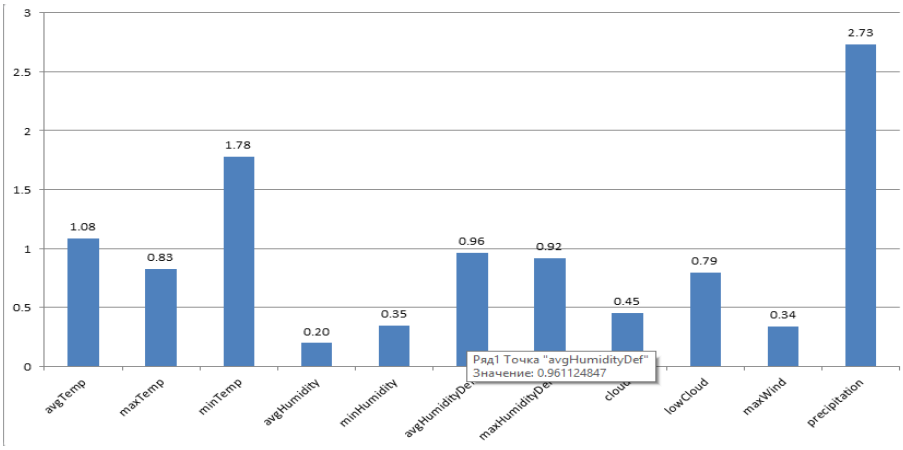

Рис. 1. Коефіцієнт варіації кількісних ознак

На другому кроці, видалення об’єктів, в яких існує хоча б одна пропущена ознака, може призвести до втрати інформації та до зниження якості класифікації. На рис. 2 показано відсоток пропусків в даних для кожної ознаки і, як видно, відсоток є незначним і вони присутні тільки для ознак, що оцінюють хмарність [5].

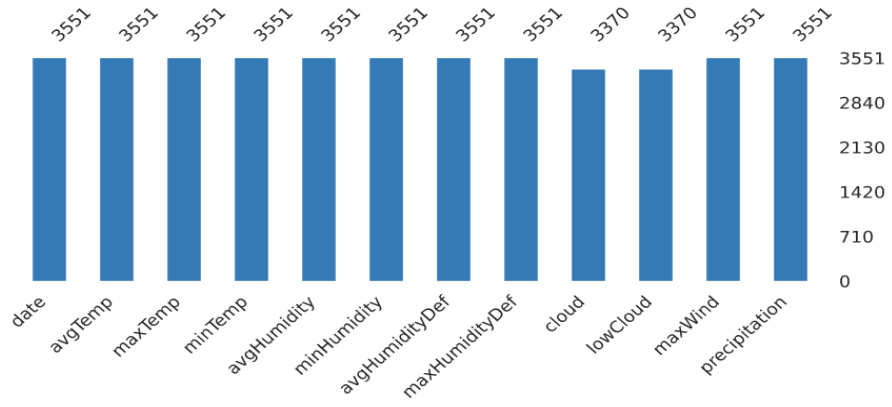

Рис. 2. Відсоток пропусків для кожної ознаки

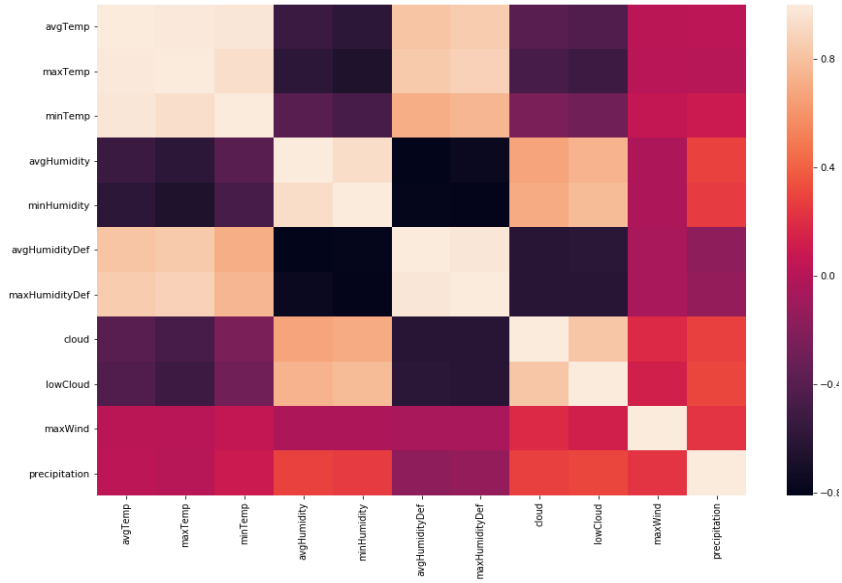

Рис. 3. Візуалізація кореляційної матриці ознак у вигляді теплової карти

Для визначення ступеня лінійного зв’язку між ознаками використовувався коефіцієнт кореляції Спірмена [6], [7]. Повна матриця кореляції візуалізована у вигляді теплової карти, показаної на рис. 3. По ній видно, що багато ознак не корелюють між собою, але існують такі ознаки, коефіцієнт кореляції яких по модулю більше 0,7, що вказує на високий лінійний статистичний зв’язок. Такими ознаками є середня, мінімальна та максимальна температури повітря. Тому доцільно використовувати в моделях штучного інтелекту одну з цих ознак.

Розробка моделі прогнозування наявності опадів

В рамках цієї роботи використовуються такі методи класифікації як градієнтний бустинг [3] та логістична регресія [4].

Задача прогнозування наявності опадів зводиться до задачі бінарної класифікації, яку сформулюємо таким чином. Дано множину об’єктів X і їх класи $Y = \{0, 1\}$, а також цільова функція y^*, значення якої $y_i = y^*(x)$ відомі на деякій підмножині класів $\{xi\}$ множини X, де xi — вектор розмірності n. Необхідно побудувати таку функцію obj, яка за відомими даними (xi, yi) наближає цільову функцію y на всій вибірці даних, $i = 1,\ldots m$.

Для побудови бінарного класифікатора розглянемо градієнтний бустинг. Ідея градієнтного бустингу полягає в побудові ансамблю дерев рішень таким чином, щоб кожне наступне дерево старалося покращити якість всієї комбінації дерев.

Класифікація здійснюється за такою формулою [8]:

$$obj\_boost(X) = \arg\max_{y \in Y} \sum_{k: b_k(x)=y}^{n} \alpha_k, \quad (1)$$

де $b_k(x)$ — відповідь k-го дерева на об'єкті x; α_k — вклад k-го дерева в композицію.

В процесі навчання послідовно будуються K дерев рішень на всіх m об'єктах та s випадково обраних ознак з n. Після навчання чергового дерева, ваги невірно класифікованих об'єктів зростають, тим самим наступне дерево здійснює фокусування в основному на них.

Градієнтний бустинг має декілька різновидностей. Найпопулярнішими моделями градієнтного бустингу є lgb та xgb, про які детально можна познайомитися в документації [3], [4].

Іншим класифікатором, який розглядається в роботі, є логістична регресія. Вона є статистичною лінійною моделлю класифікації, що дозволяє спрогнозувати апостеріорні ймовірності класів за допомогою логістичної кривої. Об'єкт відноситься до класу з найбільшою ймовірністю, що визначається за такою формулою [8]:

$$obj\_reg(X) = \arg\max_{y \in Y} P\left(y^*(x) = y\right); \tag{2}$$

$$P\left(y^*(x) = y\right) = \frac{e^{<x,\alpha_y>}}{\sum_{k=1}^{K} e^{<x,\alpha_k>}}, \tag{3}$$

тобто об'єкту x присвоюється клас з найбільшою ймовірністю, яка обчислюється згідно з softmax функцією, α_k — вектор регресійних коефіцієнтів, які пов'язані з класом k, а $x = (x1, \ldots xn)$ — вектор ознак.

Прогнозування наявності опадів пропонується здійснювати за такою формулою:

$$obj(X) = w_1 \cdot obj\_xbg(X) + w_2 \cdot obj\_\lg b(X) + w_3\, obj\_reg(X); \tag{4}$$

$$w_1 + w_2 + w_3 = 1, \tag{5}$$

де $obj\_xbg(X), obj\_\lg b(X)$ *та* $obj\_reg(X)$ — результат класифікації за методом градієнтного бустінгу (1) для типів *xbg*, *lgb* та логістичної моделі, відповідно; w_1, w_2, w_3 — ваги моделей класифікацій.

Модель класифікації (4) — це ансамбль моделей штучного інтелекту, які зважені вагами. Для забезпечення максимальної точності бінарної класифікації моделі (4) необхідно визначити оптимальні значення ваг та вектор оптимальних ознак, за якими здійснюється прогнозування опадів. Розглянемо оптимізацію цих параметрів.

Оптимізація параметрів ансамблю моделей штучного інтелекту

Для вибору оптимальних ознак класифікації використаємо такий алгоритм: на початковому етапі послідовно розглядаються одноелементні набори і вибирається такий одноелементний набір, який дає найкращі результати класифікації. Потім послідовно розглядаються вже двоелементний набори, які містять отриманий на попередньому кроці оптимальний одноелементний набір, і вибирається оптимальний двоелементний набір. Далі повторюється те ж саме, але вже для триелементних наборів. Алгоритм послідовно додає ознаки до тих пір, поки вдається поліпшити якість класифікації. Якщо поліпшити якість класифікації не вдається, то алгоритм аварійно завершує свою роботу.

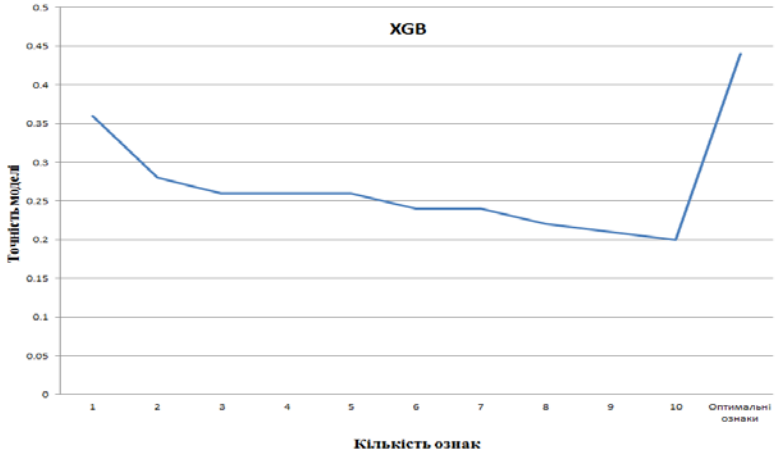

Рис. 4. Вплив ознак на точність класифікації за моделлю *xgb*

На рис. 4, 5 зображено криві точності класифікації за послідовного додавання ознак для моделей класифікацій *xbg, lgbm* [9]—[11] та логістичної моделі, відповідно. Як видно з цих рисунків, оптимальним набором ознак (ОНО) є «Середня температура повітря», «Максимальний дефіцит вологості повітря», «Максимальна швидкість вітру», «Середня вологість повітря» та «Оцінка загальної хмарності»

Точність класифікації визначено за частотою помилки:

$$F(X) = \sum_{j=1}^{m} \frac{\Delta_j}{m}; \tag{6}$$

$$F(X) = \begin{cases} 1, \text{ якщо } obj_j(X) = Y_j, \\ 0, \text{ якщо } obj_j(X) \neq Y_j. \end{cases} \tag{7}$$

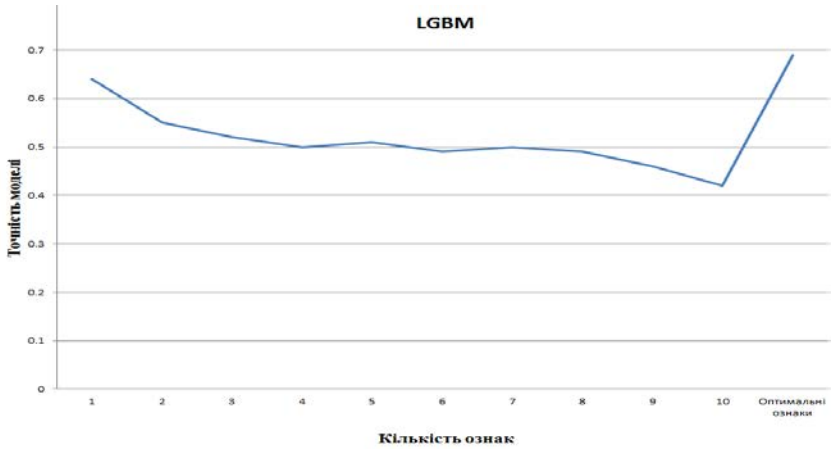

Рис. 5. Вплив ознак на точність класифікації за моделлю *lgbm*

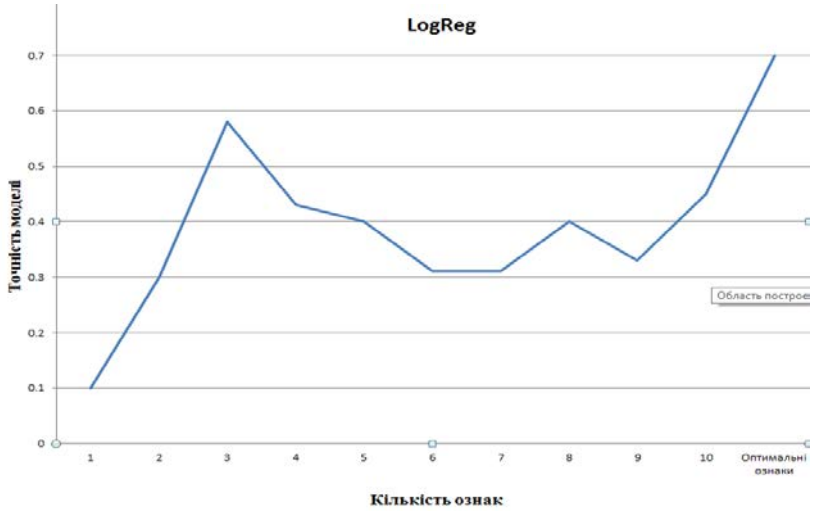

Рис. 6. Вплив ознак на точність класифікації за логістичною моделлю

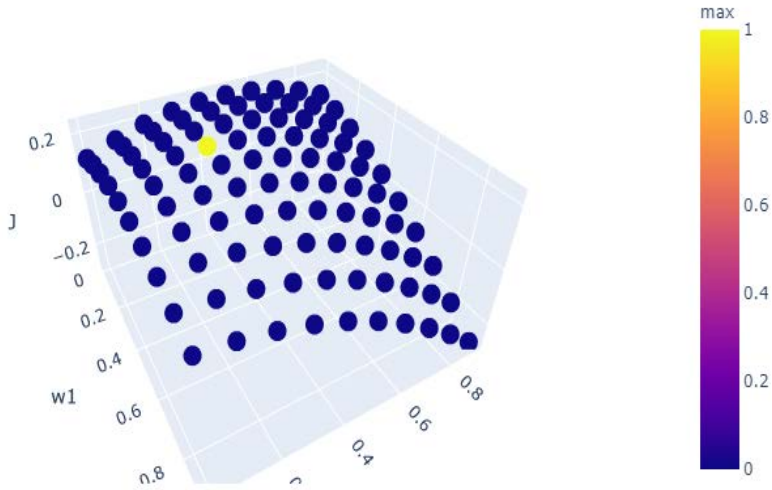

Рис. 7. Поверхня точності класифікації при різних значеннях ваг

Для визначення оптимальних значень ваг у формулі (4) побудовано поверхню, показану на рис. 7. Ця поверхня відображає точність класифікації за формулою (4), яка здійснювалася за різних значень ваг w_1 та. w_2 Значення ваг генерувалися таким чином, щоб забезпечити умову (5). Зауважимо, що достатньо було визначити оптимальні значення двох ваг. Оптимальне значення третьої ваги легко визначити за формулою (5).

Як видно з рис. 7, оптимальним значенням ваг w_1 та w_2 відповідають значення 0,3 та 0,3. Отже, формулу (5) можна переписати

таким чином:

$$obj(X) = 0,3 \cdot obj\_xbg(X) + 0,3 \cdot obj\_lgb(X) + 0,4 \cdot obj\_reg(X). \quad (8)$$

На рис. 8 показаний результат прогнозу наявності опадів за формулою (6). Точність прогнозу складає 80 %.

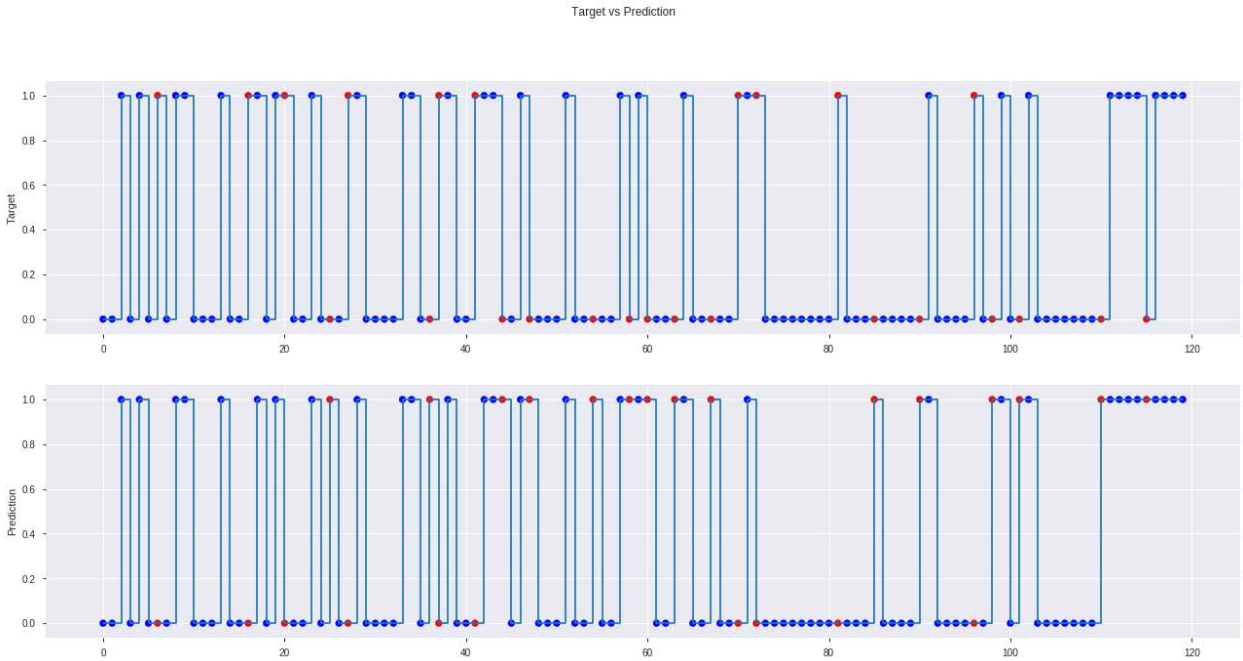

Рис. 8. Графік порівняння прогнозу за формулою (6) з реальними даними

Висновки

В роботі запропоновано інформаційну технологію оптимізації параметрів ансамблю таких моделей штучного інтелекту як моделі бустингу та логістичної регресії. Запропонована інформаційна технологія містить аналіз вхідних даних, вибір інформативного простору ознак, а також експериментальне визначення оптимальних значень ваг в ансамблі моделей класифікацій, що розглядаються. Точність прогнозу запропонованої інформаційної технології визначено за частотою помилки. Ця точність складає 80 %.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] A. Bezerra, I. Silva, L. A. Guedes, D. Silva, G. Leitão, and K. Saito, “Extracting Value from Industrial Alarms and Events: A Data-Driven Approach Based on Exploratory Data Analysis,” *Sensors*, 2019, no. 19, issue 12, pp. 11-32.

[2] *Як роблять прогнози погоди і чому вони іноді не збуваються? Прогноз.* [Електронний ресурс]. Режим доступу: https://www.bbc.com/ukrainian/features-51545290 . Назва з екрана.

[3] *Прогнози погоди і клімату та притаманні їм обмеження.* [Електронний ресурс]. Режим доступу: http://prima.franko.lviv.ua/faculty/geology/phis_geo/fourman/E-books-FVV/Intera ctive%20books/Meteorology/Weather%20Forecasting/Weather%20Ukraine/Meteo-forecasting/Analyze-forecast%20of%20limits%20climate.htm . Назва з екрана.

[4] *Прогнозування погоди.* [Електронний ресурс]. Режим доступу: http://prima.franko.lviv.ua/faculty/geology/phis_geo/fourman/E-books-FVV/Intera ctive%20books/Meteorology/ForecaseM.htm. Назва з екрана.

[5] Guolin Ke, et. al.,“LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Advances in Neural Information Processing Systems 30* (NIPS 2017), pp. 3149-3157.

[6] E. Bauer, and R. Kohavi, “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants,” *Machine Learning*, 1999, pp. 105-139.

[7] *Module pandas_profiling.* [Electronic resource]. Available: https://pandas-profiling.github.io/pandas-profiling/docs/ .

[8] *Matplotlib API Overview*. [Electronic resource]. Available: https://matplotlib.org/api/index.html .

[9] *A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics.* [Electronic resource]. Available: https://arxiv.org/abs/1811.11440 .

[10] *XGBoost Documentation.* [Electronic resource]. Available: https://xgboost.readthedocs.io/en/latest/ .

[11] *LightGBM Documentation.* [Electronic resource]. Available:https://lightgbm.readthedocs.io/en/latest/ .

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 28.01.2021

Дратований Михайло Володимирович — асистент кафедри системного аналізу та інформаційних технологій, e-mail: mishadratovany@gmail.com ;

Козачко Олексій Миколайович — канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: lekoz80@gmail.com ;

Мельник Олена Леонідівна — студентка факультету комп'ютерних систем і автоматики;

Варчук Ілона В'ячеславівна — канд. техн. наук, доцент, доцент кафедри системного аналізу та інформаційних технологій, e-mail: ilona.varchuk@gmail.com .

Вінницький національний технічний університет, Вінниця

M. V. Dratovanyi[1]
O. M. Kozachko[1]
O. L. Melnyk[1]
I. V. Varchuk[1]

Information Technology of Optimization Parameters of the Assembly Models of Artificial Intelligence for Forecasting the Presence Precipitations by Meteorological Monitoring

[1]Vinnytsia National Technical University

Data forecasting is a trivial task of systems analysis, there are different types of forecasts and predictions. One of them is a binary forecast that answers the question of whether an event will occur or not. One of the issues of meteorology is the issue of forecasting precipitation, as well as what accuracy will be in such a forecast.

The paper considers the problem of forecasting the presence of precipitation according to meteorological monitoring and proposes information technology to optimize the parameters of the ensemble of such models of machine learning as models of gradient boosting and logistic regression, built on a set of informative features. The proposed information technology allows you to perform intelligence analysis of input data and determine the optimal set of informative features, and due to the algorithm, which at each step determines the optimal one, two, three,… -element sets of features that maximize forecasting accuracy. Graphs of influence of signs on accuracy of the used models of machine learning are constructed. Each type of model has its own set of features. To provide information technology, the data provided by the Vinnytsia Center for Hydrometeorology were selected. These are the data of atmospheric monitoring of Vinnytsia for the last 10 years, which include: air temperature, humidity, dew point, cloudiness and wind speed.

To increase the accuracy of forecasting, a mathematical model is proposed, which is based on the optimal determination of the weights of the ensemble of models of gradient boosting and logistic regression. Experiments were performed that showed a fairly accurate result. The accuracy of the proposed information technology showed 80%. This confirmed the reliability of the proposed technology.

Keywords: information technology, artificial intelligence models, precipitation forecasting, informative features.

Dratovanyi Mykhailo V. — Assistant of the Chair of System Analysis and Information Technologies, e-mail: mishadratovany@gmail.com ;

Kozachko Oleksii M. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: lekoz80@gmail.com ;

Melnyk Olena L. — Student of the Department of Computer Systems and Automation;

Varchuk Ilona V. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of System Analysis and Information Technologies, e-mail: lekoz80@gmail.com

М. В. Дратованый[1]
А. Н. Козачко[1]
А. Л. Мельник[1]
И. В. Варчук[1]

Информационная технология оптимизации параметров ансамбля моделей искусственного интеллекта для прогнозирования наличия осадков по данным метеомониторинга

[1]Винницький национальный технический университет

Прогнозирование данных — это тривиальная задача системного анализа, существуют различные виды прогнозов и предсказаний. Одним из них является бинарный прогноз, который отвечает на вопрос: " Состоится событие или нет?» Один из вопросов метеорологии — это вопрос прогнозирования наличия осадков, а также какая точность будет у такого прогноза.

Рассмотрена задача прогнозирования наличия осадков по данным метеорологического мониторинга и предложена информационная технология оптимизации параметров ансамбля таких моделей машинного обучения, как модели градиентного бустинга и логистической регрессии. Они построены на основе набора информативных признаков. Предложенная информационная технология позволяет выполнить разведывательный анализ входных данных и определить оптимальный набор информативных признаков, а за счет алгоритма, который на каждом шагу определяет оптимальные одно-, двух-, трех-, ..элементные наборы признаков, максимизировать точность прогнозирования. Построены графики влияния признаков на точность использованных моделей машинного обучения. Для каждого типа моделей определен свой набор признаков. Для построения информационной технологии взяты данные, предоставленные Винницким центром по гидрометеорологии. Это данные мониторинга атмосферы г. Винницы за последние 10 лет, которые включают: температуру воздуха, влажность воздуха, точку росы, облачность и скорость ветра.

Для повышения точности прогнозирования предложена математическая модель, основанная на оптимальном определении весов ансамбля моделей градиентного бустинга и логистической регрессии. Проведены эксперименты, которые показали достаточно точный результат. Точность предложенной информационной технологии показала 80 %. Это подтвердило достоверность предложенной технологии.

Ключевые слова: информационная технология, модели искусственного интеллекта, прогнозирование наличия осадков, информативные признаки.

Дратованый Михаил Владимирович — ассистент кафедры системного анализа и информационных технологий, e-mail: mishadratovany@gmail.com ;

Козачко Алексей Николаевич — канд. техн. наук, доцент, доцент кафедры системного анализа и информационных технологий, e-mail: lekoz80@gmail.com ;

Мельник Елена Леонидовна — студент факультета компьютерных систем и автоматики;

Варчук Илона Вячеславовна — канд. техн. наук, доцент, доцент кафедры системного анализа и информационных технологий, e-mail: lekoz80@gmail.com