DOI https://doi.org/10.15407/csc.2023.03.024
УДК 004.852

О.В. РАДЧЕНКО, студент, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
ORCID: https://orcid.org/0009-0002-5810-4526,
03056, м. Київ, Берестейський просп., 37, Україна,
radchenko.oleh@lll.kpi.ua

В.А. ПАВЛОВ, канд. техн. наук, доцент, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
ORCID: https://orcid.org/0000-0002-3293-5308,
03056, м. Київ, Берестейський просп., 37, Україна,
pavlov.volodymyr@lll.kpi.ua

О.К. ГОРОДЕЦЬКА, канд. техн. наук, доцент, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
ORCID: https://orcid.org/0000-0003-1288-3528,
03056, м. Київ, Берестейський просп., 37, Україна,
o.nosovets@gmail.com

Г.А. КОРНІЄНКО, старший викладач Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
ORCID: https://orcid.org/0009-0003-2104-5745,
03056, м. Київ, Берестейський просп., 37, Україна,
galinakor5555@gmail.com

БАГАТОКЛАСОВИЙ КЛАСИФІКАТОР НА ОСНОВІ БІНАРНИХ ЛОГІСТИЧНИХ РЕГРЕСІЙ, ОДЕРЖАНИХ ЗА ПРИНЦИПАМИ МГУА

В роботі розглядається задача підвищення точності мультикласової класифікації моделей множинної логістичної регресії. Формувати класифікатор пропонується шляхом поєднання бінарних логістичних регресій оптимізованої структури. Бінарні моделі одержуються кроковим алгоритмом логістичної регресії Stepwise, удосконаленим за принципами методу групового врахування аргументів. Переваги пропонованого підходу продемонстровано на медичних даних публічно-доступного інтернет-ресурсу Kaggle.

***Ключові слова**: мультикласовий класифікатор, кроковий алгоритм, Stepwise, логістична регресія, оптимізація моделі, зовнішній критерій, метод групового врахування аргументів.*

Вступ

Точність моделей класифікації є звичайною метою алгоритмів структурно-параметричного синтезу. Розвиток методів, які забезпечують високу точність класифікації, є одним із пріоритетів задач машинного навчання [1]. Розроблено різні підходи, які застосовуються відповідно до особливостей постановки

задачі та властивостей простору ознак. Серед статистичних підходів, що формують аналітичні моделі класифікації широко застосовується метод опорних векторів (*SVM*). Завдяки використанню ядерних функцій він є достатньо конкурентним, здатним одержувати нелінійні границі між класами. Забезпечує високі результати на невеликих наборах даних. Проте він є чутливим до підбору гіперпараметрів та вимагає значних обчислювальних ресурсів [2]. Одним з ефективних підходів до класифікації є група алгоритмів, що засновані на згортці дерев прийняття рішень. Деревоподібні класифікатори мають велику гнучкість і легкість в інтерпретації результатів, проте й схильність до перенавчання [3]. Проте останні розробки в класі алгоритмів *Random Forest* [4] та самоорганізованих лісів [5] швидко розвивають підхід, долають цей недолік та мають усі перспективи для одержання стійких високих результатів.

Одним із поширених алгоритмів класифікації є логістична регресія, цей алгоритм використовується для розв'язання бінарних задач класифікації. Перевагами логістичної регресії є простота та прозора інтерпретація результату. Модель може бути використана для виявлення взаємозв'язків між ознаками та впливу кожної ознаки на результат класифікації. Проте логістична регресія має обмеження при розв'язанні задач багатокласової класифікації та втрачає точність при складних сценаріях [6].

Багатокласові завдання класифікації відрізняються певною специфікою. У контексті машинного навчання багатокласова класифікація є викликом, оскільки вимагає розподілу даних на багато класів з високою точністю та стабільністю [7]. Мультикласові класифікатори можуть формуватися безпосередньо як єдина функція, що визначає відношення об'єкта до класу, чи як сполучення бінарних класифікаторів, сформованих за принципом один клас проти всіх інших класів. Серед ефективних методів для вирішення таких завдань класифікація на основі бінарних логістичних регресій виявляється одним з універсальних і застосовуваних підходів [8].

У роботі пропонується підхід до класифікації – *мультикласовий класифікатор* на основі бінарних логістичних регресій, отриманих за принципами методу групового врахування аргументів (МГУА). Цей підхід комбінує переваги бінарних логістичних регресій та групового врахування аргументів для досягнення високої точності та стабільності при класифікації багатьох класів. Підхід може бути ефективним у великому спектрі задач класифікації, включно зі складними сценаріями та наборами даних із різноманітними характеристиками [9, 10]. Пропонується розробити модифікацію крокового алгоритму багатовимірної бінарної логістичної регресії *Stepwise*, де за принципами МГУА оптимізуються параметри алгоритму: рівні значимості за тестом відношення логарифмічних правдоподібностей для включення та виключення аргументів моделі. Алгоритм формує бінарні логістичні регресії для кожного класу за принципом «один проти всіх» та комбінує їх у єдиний мультикласовий класифікатор, що повертає належність об'єкта до класу за максимумом значення, досягнутого окремими бінарними регресіями.

Мета роботи полягає у розробці алгоритму синтезу моделей логістичної регресії оптимальної складності та формування мультикласового класифікатора на основі одержаних моделей бінарних логістичних регресій.

Постановка задачі

Вхідними параметрами задачі є матриця спостережень (вхідні дані) $x \in R^M$ та вектор залежної змінної $Y \in (N)$, що є індексом класу 0, 1, 2, ...

$$X = \begin{Bmatrix} x_{11} & x_{12} & ... & x_{1M} \\ x_{21} & x_{22} & ... & x_{2M} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ,,, & x_{nM} \end{Bmatrix}, \; Y = \begin{Bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{Bmatrix}. \qquad (1)$$

Тут n – кількість спостережень, M – кількість змінних (факторів).

Метою завдання є побудова моделі логістичної регресії виду [11] оптимальної структури:

$$Y=\frac{1}{1+e^{-z}}, z=b_0+\sum_{j=1}^{m} b_j x_{i_j}, \qquad (2)$$

де $m < M$ та bj – коефіцієнти регресії.

Параметр m повинен відповідати оптимальній складності моделі, базис $(x_{i_1}, x_{i_2}, ..., x_{i_m})$ – оптимальній структурі. Для всіх варіантів сполучень класів сформованих за принципом «один проти всіх» одержуємо набір бінарних логістичних регресій, що надалі поєднуємо у мультикласовий класифікатор.

Модифікація крокового алгоритму *Stepwise*

У роботі пропонується модифікувати розширену версію крокового алгоритму *Stepwise*, відомого, як *Bidirectional Elimination,* додавши в процедуру включення/виключення предикторів цикл для визначення оптимальних порогів відбору, та зовнішній критерій, що обмежує складність моделі. На кожній ітерації алгоритму зберігається модель та значення зовнішнього критерію. Вибір моделі завершується результатом за максимальним значенням зовнішнього критерію.

Алгоритм *Stepwise* було обрано базовим алгоритмом для модифікацій за таких обставин:

1. Алгоритм має зручну структуру для модифікації за допомогою заміни критеріїв включення та виключення ознак із моделі. Це дає змогу досліджувати різні критерії, такі як значущість *p*-значення, коефіцієнти *t*-статистики, коефіцієнт детермінації *R*-квадрат, інформаційні критерії, тощо [12].

2. Алгоритм може бути у природний спосіб розширений етапом регуляризації [13], що дає змогу контролювати складність моделі та зменшує ризик перенавчання завдяки введенню штрафних параметрів для коефіцієнтів регресії. Етап уможливлює адаптацію алгоритму до конкретних вимог і обрання оптимальної модифікації для досягнення найкращих результатів.

3. Кроковий алгоритм легко може бути модифіковано завдяки застосуванню різних гібридних стратегії, що у різний спосіб поєднують процеси включення та виключення аргументів при формуванні оптимальної структури моделі. Зміна стратегії пошуку дає змогу обрати найефективніший підхід для побудови моделі з мінімальними затратами ресурсів [14].

4. В умовах завдань великої розмірності можуть бути використані додаткові обмеження до процесу включення та виключення ознак [15]. Такі обмеження є корисними в ситуаціях, коли має значення економія ресурсів.

Алгоритм складається з таких етапів:

1. Матриця спостережень X розширюється узагальненими змінними.

2. Матриця спостережень x ділиться на тренувальну (A), валідаційну (B) та тестову (C) вибірки за критерієм подібності відстані Махаланобіса [16] у заданому процентному співвідношенні *a/b/c*.

3. Створюється сітка пар порогових значень включення та виключення ($\alpha_{вкл}$ та $\alpha_{викл}$) предикторів, спираючись на базові рекомендовані $\alpha_{вкл} = 0,05$ та $\alpha_{викл} = 0,1$ [17].

4. Для кожної пари порогових значень із сітки виконується модифікована процедура крокового алгоритму *Stepwise*, під час якої включаються або виключаються предиктори з моделі за статистичним тестом відношення логарифмічних правдоподібностей на тренувальній вибірці (A).

4.1. Включення предикторів: початкова ітерація починається з моделі без предикторів до якої додається предиктор, що має найбільший коефіцієнт кореляції Пірсона з вихідною змінною.

Розраховується тест відношення логарифмічних правдоподібностей моделі, отриманої на попередній ітерації q (без предиктора) з моделлю із кожним новим предиктором, ще не включеним у модель [18]:

$$G=-2\left(L_{q+1}-L_q\right), \qquad (3)$$

де L_{q+1} – функція логарифмічної правдоподібності моделі з включеним предиктором на ітерації (q+1), L_q функція логарифмічної правдоподібності моделі, отриманої на попередній ітерації.

Функція логарифмічної правдоподібності визначається як [19]:

$$\log L\left(\theta|x\right)=\sum_{i=1}^{n}\log f\left(x_i|\theta\right), \qquad (4)$$

де: θ – вектор параметрів моделі, який потрібно оцінити, $x = \{x_1, x_2, ..., x_n\}$ – вибірка даних, $f\left(x_i|\theta\right)$ – функція щільності для кожного спостереження x_i залежно від параметрів θ.

Значення значущості p отримується за функцією розподілу χ^2 відповідно до (3) зі ступенем свободи, рівним 1, оскільки моделі відрізняються на один предиктор. З усіх перебраних на цій ітерації обирається предиктор із мінімальним значенням значущості p: при $p < \alpha_{вкл}$, приймається рішення про включення предиктора у модель. За умови включення предиктора в модель виконується процедура виключення предикторів із моделі.

4.2 Виключення предикторів: проводиться тест відношення правдоподібностей моделі на поточній ітерації з моделлю, що не включає один із предикторів, який уже було включено в модель раніше. Тоді формула логарифмічної правдоподібності G (3) матиме вигляд (5):

$$G=-2\left(L_q - L_{q+1}\right). \qquad (5)$$

Обирається предиктор із найбільшим значенням p: якщо $p > \alpha_{викл}$, то приймається рішення про виключення предиктора з моделі.

5. На кожній ітерації включення або виключення предиктора для моделі розраховується та запам'ятовується *значення зовнішнього критерію.*

Зовнішній критерій $C_{зовн}$ визначається на основі матриці розбіжностей класифікації, є складовим, що враховує значення на навчальній вибірці A та валідаційній B, див. (6).

Кожна частина критерію є також складовою, що враховує точність класифікації у класах та баланс точності класифікації між класами див. (7)–(10).

$$C_{зовн} = \alpha\Delta^A + \left(1-\alpha\right)\Delta^B, \qquad (6)$$

де Δ^A – значення критерію класифікації на навчальній вибірці, Δ^B – значення критерію класифікації на валідаційній вибірці, α – коефіцієнт балансу критерію класифікації на навчальній та валідаційній вибірках. Складова критерію розраховується у такий спосіб:

$$\Delta^A = \left(\gamma-1\right)\cdot {}^{1}\!/_{k}\cdot\sum_{i=1}^{k}\Delta^{A_i} + \gamma\frac{1}{1+{}^{1}\!/_{C_k^2}\cdot\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left|\Delta^{A_i}-\Delta^{A_j}\right|}, \qquad (7)$$

де γ – ваговий коефіцієнт точності класифікації у класах та балансу точності класифікації між класами, k – кількість класів, $\Delta^{A_i}, i=\overline{1,k}$ – точність класифікації у класі i на вибірці A, що визначається за формулою:

$$\Delta^{A_i} = \left(\frac{n^{A_i*}}{n^{A_i}}\right), \qquad (8)$$

де n^{A_i*} – кількість вірно прогнозованих об'єктів для i-го класу у вибірці A, n^{A_i} – кількість спостережень i-того класу у вибірці A.

Точність Δ^B розраховується на точках тестової вибірки B:

$$\Delta^B = \left(\gamma-1\right)\cdot {}^{1}\!/_{k}\cdot\sum_{i=1}^{k}\Delta^{B_i} + \gamma\frac{1}{1+{}^{1}\!/_{C_k^2}\cdot\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left|\Delta^{B_i}-\Delta^{B_j}\right|}, \qquad (9)$$

де Δ^{B_i}, $i=\overline{1,k}$ – точність класифікації у класі і на вибірці В, що визначається за формулою:

$$\Delta^{B_i} = \left(\frac{n^{B_i*}}{n^{B_i}}\right), \qquad (10)$$

де n^{B_i*} – кількість вірно прогнозованих об'єктів для i-го класу у вибірці B, n^{B_i} – кількість спостережень i-того класу у вибірці B.

6. Коли моделі для всіх пар значень порогів перебрано, обирається оптимальна модель за максимальним значенням критерію $C_{зовн}$.

7. Оцінка одержаної бінарної регресії на тестовій вибірці C здійснюється за трьома метриками:

7.1. Точність класифікації у класах, яка вираховується, як відношення правильно прогнозованих об'єктів класів до кількості всіх об'єктів даного класу.

7.2. F-міра:

$$F_1 = \frac{TP}{TP+\frac{1}{2}\left(FP+FN\right)}, \qquad (11)$$

де TP (*True Positive*) – сума вірно спрогнозованих об'єктів першого класу, FP (*False Positive*) –

Таблиця 1. **Оцінки точності класифікації за кожним класом окремо**

Клас	Оригінальна модель (*ROC AUC*-міра)	Оригінальна модель (*F*1-міра)	Оптимізована модель (*ROC AUC*-міра)	Оптимізована модель (*F*1-міра)
1	0,77	0,93	0,83	0,948
2	0,71	0,60	0,76	0,651
3	0,70	0,80	0,89	0,846

Таблиця 2. **Порівняльна таблиця середньої за класами точності класифікації**

Вид оцінки	Оригінальна модель	Оптимізована модель
F-міра	0,77	0,815
Точність	0,8531	0,90
ROC AUC-міра	0,727	0,827

сума хибно спрогнозованих об'єктів першого класу, *FN* (*False Negative*) – сума хибно спрогнозованих значень другого класу. Для випадку одержання регресій за принципом «один проти всіх» під «усіма» розуміють другий клас.

7.3. *Area Under Curve Receiver Operating Characteristic* (*ROC AUC*) площа під *ROC*-кривою, яка використовується для оцінки якості моделей класифікації. *ROC*-крива є графічним зображенням залежності між чутливістю (*true positive rate – TPR*) та специфічністю (*true negative rate – TNR*) класифікатора при зміні порогового значення для визначення класів.

Порівняльна оцінка роботи алгоритмів

Оцінку роботи класифікаторів, побудованих згідно базової версії алгоритму *Stepwise* та вдосконаленої версії, пропонується провести на медичних даних з публічно-доступного інтернет-ресурсу *Kaggle* [20]. Розглядається завдання класифікації здоров'я плоду жінок. Набір даних містить змінні, що характеризують стан плоду та вихідну класифікацію здоров'я плоду. Стан плоду характеризується такими ознаками: базова частота серцевих скорочень плоду, кількість прискорень рухів плоду на секунду, кількість рухів плоду за секунду, кількість скорочень матки за секунду, кількість різких сповільнень на секунду, кількість тривалих уповільнень на секунду, відсоток часу з аномальною короткостроковою мінливістю, середнє значення короткострокової мінливості, відсоток часу з аномальною довгостроковою мінливістю та інші статистичні ознаки (сумарно – 22 ознаки). Набір даних містить 2126 спостережень, отриманих із кардіотокограм, які були класифіковані на 3 класи здоров'я плоду: норма, підозра на патологію, патологія.

Порівняння роботи алгоритмів здійснювалося на двох серіях моделей:

1. Оригінальні бінарні моделі, створені на множині вхідних аргументів, що складається з початкових ознак та їх розширень узагальнених аргументів.

2. Оптимізовані моделі, створені на тій самій множині аргументів із застосуванням оптимізаційного алгоритму.

Для побудови оригінальних бінарних логістичних моделей було використано програмний модуль *scikit-learn* із відкритим вихідним кодом мовою програмування *Python, sklearn.linear_model.LogisticRegression*. У модуль включено засоби регуляризації параметрів, перевірку оригінальної та оптимізованої моделі здійснюють на розбитті вибірок даних по Махалонобісу [16], що вже початково дає приріст точності стосовно використання моделі

логістичної регресії для класифікації без збалансованого розбиття та регуляризації параметрів.

Оптимізовані моделі бінарної логістичної регресії, одержані за принципом один проти всіх для класів 1–3, наведено у (12–14). За результатами роботи алгоритма, було обрано 14 ознак з 22, які найкраще описують вихідну зміну.

$$z = 1{,}818 - 2.27x7 / x16 - 8{,}69x13 / x8 + 1{,}57 * x17 + \\ +1{,}41 * x2 * x16 - (1{,}82 * x19) / x12 - 8{,}50 * x9 * x10 + \\ +1{,}349 * x4 * x17 + 2{,}050 * x10 * x14 + \\ + (10{,}67 * x9) / x7 + 1{,}279 * x2 * x13 - \\ - (2{,}79 * x6) / x5 - 1{,}60 * x6 * x18; \quad (12)$$

$$z = -12{,}9177 - 3{,}0564x7 / x16 + 5{,}8045x13 / x8 + \\ +7{,}9805 * x17 - 1{,}4079 * x2 * x16 + \\ + (1{,}359 * x19) / x12 + 1{,}1862 * x9 * x10 - \\ -5{,}8067 * x4 * x17 - 1{,}2523 * x10 * x14 - \\ - (3{,}1442 * x9) / x7 - 1{,}867 * x2 * x13 + \\ + (6{,}1256 * x6) / x5 - 1{,}4946 * x6 * x18; \quad (13)$$

$$z = 2{,}7598 + 3{,}7331x7 / x16 + 2{,}1095x13 / x8 - \\ -5{,}4517 * x17 - 1{,}3439 * x2 * x16 + \\ + (2{,}4267 * x19) / x12 - 2{,}6786 * x9 * x10 - \\ -1{,}8959 * x4 * x17 - 3{,}2582 * x10 * x14 + \\ + (1{,}8278 * x9) / x7 - 1{,}5192 * x2 * x13 + \\ + (9{,}6849 * x6) / x5 + 1{,}3632 * x6 * x18, \quad (14)$$

де:

'accelerations' ($x2$) – кількість прискорень рухів плоду на секунду;

'uterine_contractions' ($x4$) – кількість скорочень матки за секунду;

'light_decelerations' ($x5$) – кількість різких сповільнень на секунду;

'prolongued_decelerations' ($x6$) – кількість тривалих уповільнень на секунду;

'abnormal_short_term_variability' – ($x7$) відсоток часу з аномальною короткостроковою мінливістю;

'percentage_of_time_with_abnormal_long_term_variability' ($x9$) – відсоток часу з аномальною довгостроковою мінливістю;

'mean_value_of_long_term_variability' ($x10$) середнє значення довгострокової мінливості;

'histogram_min' ($x12$) – мінімальне значення гістограми ознак спостереження;

'histogram_max' ($x13$) – максимальне значення гістограми ознак спостереження;

'histogram_number_of_peaks' ($x14$) – кількість піків гістограми значень ознак спостереження;

'histogram_mode' ($x16$) – модус гістограми ознак спостереження;

'histogram_mean' ($x17$) – середнє значення гістограми ознак спостереження;

'histogram_median' ($x18$) – медіанне значення гістограми ознак спостереження;

'histogram_variance' ($x19$) – дисперсія гістограми ознак спостереження;

У табл. 1 і 2 наведено показники якості класифікації на тестовій вибірці для кожного класу окремо та для мультикласового класифікатора на основі оригінальних та оптимізованих бінарних логістичних моделей.

Пропонований алгоритм дав змогу одержати суттєво спрощені (14 ознак проти 22) моделі бінарних логістичних регресій. Мультикласовий класифікатор на основі оптимізованих моделей, має приріст точності та *F*-міри на 0,045, та середньої *ROC AUC*-міри на 0,1.

Висновки

У роботі запропоновано вдосконалення алгоритму багатокласової класифікації на основі бінарних логістичних регресій за принципами методу групового врахування аргументів. Здійснено апробацію алгоритму на медичних даних із публічно-доступного інтернет-ресурсу. Одержані показники демонструють суттєвий приріст якості класифікації системи та відносну ефективність пропонованої версії алгоритму синтезу мультикласових систем класифікації. Підвищення якості класифікації досягнуто внаслідок оптимізації структури бінарних класифікаторів.

ЛІТЕРАТУРА

1. Schmidhuber, J., 2015. “Deep learning in neural networks: An overview”. Neural Networks. Vol. 61, pp. 85–117.
2. Ben-Hur, A., Horn, D., Siegelmann, H. T., Vapnik, V., 2001. "Support vector clustering". Journal of Ma-chine Learning Research, 2, pp.125–137.
3. Von Winterfeldt, D., Edwards, W., 1986. “Decision trees”. Decision Analysis and Behavioral Research. Cambridge University Press. pp. 63–89. ISBN 0-521-27304-8.
4. Babenko, V., Nastenko, I., Pavlov, V., Horodetska, O., Dykan, I., Tarasiuk, B., Lazoryshinets, V., 2023. “Classification of Pathologies on Medical Images Using the Algorithm of Random Forest of Optimal-Complexity Trees”. Cybernetics and Systems Analysis, 59 (2), pp. 346–358. DOI: 10.1007/s10559-023-00569-z.
5. Davydko, O., Hladkyi, Y., Linnik, M., Nosovets, O., Pavlov, V., Nastenko, Ie., 2021. “Hybrid Classifiers Based on CNN, LSOF, GMDH in COVID-19 Pneumonic Lesions Types Classification Task”. Proceedings of the XVI IEEE International Conference CSIT-21& International Workshop on Inductive Modeling. Lviv, Ukraine, 23-26 September, pp. 380-384. DOI: 10.1109/CSIT52700.2021.9648752.
6. Bisong, E., Bisong, E., 2019. “Logistic regression”. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, pp. 243-250.
7. Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H., Shahmoradi, L., 2017. “Accuracy improvement for diabe-tes disease classification: a case on a public medical dataset”. Fuzzy Information and Engineering, 9(3), pp. 345-357.
8. Kirasich, K., Smith, T., Sadler, B., 2018. “Random forest vs logistic regression: binary classification for heterogeneous datasets”. SMU Data Science Review, 1(3), Article 9. [online]. Available at: <https://scholar.smu.edu/datasciencereview/vol1/iss3/9> [Accessed: 1 Jan. 2023].
9. Ivakhnenko, A.G., Stepashko, V.S., 1985. Noise-immunity of modeling. Kiev: Naukova dumka, 216 p. (In Russian).
10. Ivakhnenko, A.G.; Ivakhnenko, G.A., 1995. “The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH)”. Pattern Recognition and Image Analysis. 5 (4), pp. 527–535.
11. Strano, M., Colosimo, B.M., 2005. "Logistic regression analysis for experimental determination of forming limit diagrams". International Journal of Machine Tools and Manufacture. 46 (6), pp. 673–682. DOI:10.1016/j.ijmachtools.2005.07.005.
12. Zhang, Zh., 2016. “Variable selection with stepwise and best subset approaches”. Annals of translational medicine, 4 (7). 136. DOI: 10.21037/atm.2016.03.35.
13. El-Koka, A., Cha, K.H., Kang, D.K., 2013. “Regularization parameter tuning optimization approach in lo-gistic regression”. In 2013 15th International Conference on Advanced Communications Technology (ICACT), pp. 13-18.
14. Derksen, S., Keselman, H.J., 1992. “Backward, forward and stepwise automated subset selection algo-rithms: Frequency of obtaining authentic and noise variables”. British Journal of Mathematical and Statistical Psychology, 45(2), pp. 265-282.
15. Zhou, J., Foster, D.P., Stine, R.A., Ungar, L.H., Guyon, I., 2006. “Streamwise feature selection”. Journal of Machine Learning Research, 7(9). pp. 1861-1885.
16. Hupalo, M., Pavlov, V., Nastenko Ye., Kornienko G., 2023. “Modeling results optimization based on data splitting by Mahalanobis distance similarity criterion”. Biomedical Engineering and Technology, 11, pp. 21–30 (In Ukrainian).
17. In Lee, K., Koval, J.J., 1997. “Determination of the best significance level in forward stepwise logistic re-gression”. Communications in Statistics-Simulation and Computation, 26 (2), pp. 559-575.
18. Woolf, B., 1957. “The log likelihood ratio test (the G-test)”. Annals of human genetics, 21(4), pp. 397-409.
19. Buse, A., 1982. “The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note”. The American Statistician. 36 (3a), pp. 153-157.
20. Fetal Health Classification. [online]. Available at: <https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification> [Accessed: 17 Dec. 2022].

Received 19.09.2023

ЛІТЕРАТУРА

1. Schmidhuber J. Deep learning in neural networks: An overview. Neural Networks. 2015. T. 61. C. 85–117.
2. Ben-Hur A., Horn D., Siegelmann H.T., Vapnik, V. Support vector clustering. (2001)”. Journal of Machine Learning Research. 2. P. 125–137.
3. Von Winterfeldt, D., Edwards, W. “Decision trees”. Decision Analysis and Behavioral Research. Cambridge University Press. 1986. C. 63–89.
4. Бабенко, В., Настенко, Е., Павлов, В., Городецька, О., Дикан, І., Тарасюк, Б., Лазоршннець, В. Класифікація патології за медичними зображеннями алгоритмом випадкового лісу дерев оптимальної складності. Кібернетика та системний аналіз, 2023. 59 (2). С. 190-202.

5. Davydko O., Hladkyi Y., Linnik M., Nosovets O., Pavlov V., Nastenko Ie. Hybrid Classifiers Based on CNN, LSOF, GMDH in COVID-19 Pneumonic Lesions Types Classification Task. Proceedings of the XVI IEEE International Conference CSIT-21& International Workshop on Inductive Modeling. Lviv, Ukraine, 23-26 September, 2021 P. 380–384. DOI: 10.1109/CSIT52700.2021.9648752.
6. Bisong E., Bisong E. Logistic regression. Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, 2019, P. 243–250.
7. Nilashi M., Ibrahim O., Dalvi M., Ahmadi H., Shahmoradi L. Accuracy improvement for diabetes disease classification: a case on a public medical dataset. Fuzzy Information and Engineering, 2017, 9 (3). C. 345–357.
8. Kirasich, K., Smith, T., Sadler, B. "Random Forest vs Logistic Regression: Binary Classification for Heter-ogeneous Datasets," SMU Data Science Review, 2018. Розділ 1: номер 3, стаття 9. Режим доступа: https://scholar.smu.edu/datasciencereview/vol1/iss3/9.
9. Ивахненко А.Г., Степашко В.С. Помехоустойчивость моделирования. Киев: Наук. думка, 1985. 216 с.
10. Ivakhnenko O.G.; Ivakhnenko G.A. The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH). Pattern Recognition and Image Analysis. 1995. 5 (4). P. 527–535.
11. Strano M., Colosimo B.M. Logistic regression analysis for experimental determination of forming limit di-agrams". International Journal of Machine Tools and Manufacture. 2005. 46 (6). P. 673–682. doi:10.1016/j.ijmachtools.2005.07.005.
12. Zhang, Zh. Variable selection with stepwise and best subset approaches. Annals of translational medicine, 2016, 4 (7), 136. DOI: 10.21037/atm.2016.03.35.
13. El-Koka, A., Cha, K.H., Kang, D.K. Regularization parameter tuning optimization approach in logistic re-gression. In: 2013 15th International Conference on Advanced Communications Technology (ICACT). IEEE, 2013. C. 13–18.
14. Derksen, S., Keselman, H.J. Backward, forward and stepwise automated subset selection algorithms: Fre-quency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 1992, 45 (2). C. 265–282.
15. Zhou, J., Foster, D.P., Stine, R.A., Ungar, L.H., Guyon, I. Streamwise feature selection. Journal of Machine Learning Research, 2006, 7 (9). P. 1861-1885.
16. Гупало М., Павлов В., Настенко Є., Корнієнко Г. Оптимізація результатів моделювання шляхом розбиття вибірок за критерієм подібності відстані Махаланобіса. Biomedical Engineering and Tech-nology, 2023, 11, С. 21–30.
17. In Lee K., Koval J.J. Determination of the best significance level in forward stepwise logistic regression. Communications in Statistics-Simulation and Computation, 1997, 26 (2). C. 559–575.
18. Woolf B. The log likelihood ratio test (the G test). Annals of human genetics, 1957, 21.4: 397–409.
19. Buse A. The Likelihood Ratio, Wald, and Lagrange Multiplier Tests: An Expository Note. The American Statistician. 1982. 36 (3a). C. 153–157.
20. Fetal Health Classification [Електронний ресурс] Режим доступу: https://www.kaggle.com/datasets/andrewmvd/fetal-health-classification.

Надійшла 19.09.2023

O.V. Radchenko, student, National Technical University of Ukraine
"Ihor Sikorsky Kyiv Polytechnic Institute",
ORCID: https://orcid.org/0009-0002-5810-4526,
37 Beresteyskyi Avenue, Kyiv 03056, Ukraine,
Radchenko.oleh@lll.kpi.ua
V.A. Pavlov, PhD, As. Prof, National Technical University of Ukraine
"Ihor Sikorsky Kyiv Polytechnic Institute",
ORCID: https://orcid.org/0000-0002-3293-5308,
37 Beresteyskyi Avenue, Kyiv 03056, Ukraine,
pavlov.volodymyr@lll.kpi.ua
O.K. Horodetska, PhD, As. Prof., National Technical University of Ukraine
"Ihor Sikorsky Kyiv Polytechnic Institute",
ORCID: https://orcid.org/0000-0003-1288-3528,
37 Beresteyskyi Avenue, Kyiv 03056, Ukraine,
o.nosovets@gmail.com

G.A. Korniienko, Senior Lecturer, National Technical University of Ukraine
"Ihor Sikorsky Kyiv Polytechnic Institute",
ORCID: https://orcid.org/0000-0003-2104-5745,
37 Beresteyskyi Avenue, Kyiv 03056, Ukraine,
galinakor5555@gmail.com

MULTICLASS CLASSIFIER BASED ON BINARY LOGISTIC REGRESSIONS OBTAINED ACCORDING TO THE PRINCIPLES OF GMDH

Introduction. The issue of accuracy improvement in classification tasks is always topical, and various approaches have been developed, applied in accordance with the peculiarities of the problem formulation and properties of the feature space. Among the most effective models, classifiers based on multiple logistic regressions have proved themselves.

Purpose. The aim of the paper is to develop an algorithm for solving multiclassification problems on the basis of binary logistic models built by the stepwise multiple logistic regression algorithm of the Stepwise type, improved according to the principles of the method of group accounting of arguments.

Methods. The paper proposes a modification of the stepwise algorithm for creating binary multivariate logistic regressions Stepwise, where it is proposed to optimize the algorithm parameters in accordance with the principles of the method of group consideration of arguments: significance levels by the logarithmic likelihood ratio test for inclusion and exclusion of model arguments. The choice of optimal parameters is realized in accordance with an external criterion that takes into account the balance of classification accuracy of training and test samples and the balance of class classification accuracy. Subsequently, the binary class models obtained by the one-versus-all principle are combined into a multiclass classifier that returns the answer according to the maximum likelihood of the class. The comparison of classification models obtained by the classical Stepwise algorithm and the one proposed in the robot is carried out on the medical data of the publicly available Internet resource Kaggle.

Conclusion. The paper substantiates and demonstrates the advantages of classifiers based on logistic multivariate regressions optimized according to the principles of the method of group consideration of arguments relative to the classical version of the Stepwise algorithm. The effective application of the algorithm in solving multiclass classification problem is shown.

Keywords: *multiclass classifier, stepwise algorithm, Stepwise, logistic regression, model optimization, external criterion, Group Method of Data Handling.*