Вісник Національного технічного університету "ХПІ", 2021, № 2 (6)
ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

УДК 004.6:004.932 **DOI: 10.20998/2411-0558.2021.02.04**

Т. О. БІЛОБОРОДОВА*,* канд. техн. наук, доц., ІПМЕ, Київ,
І. С. СКАРГА-БАНДУРОВА, докт. техн. наук, проф., ІПМЕ, Київ,
М. О. КОВЕРГА, асп., СНУ ім. В. Даля, Сєвєродонецьк

МЕТОДОЛОГІЯ УСУНЕННЯ ДИСБАЛАНСУ КЛАСІВ НАБОРІВ ДАНИХ ЗОБРАЖЕНЬ

Представлено методологію вирішення задачі усунення дисбалансу класів в наборах даних зображень. Запропонована методологія включає етапи вилучення фрагментів зображень, аугментація фрагментів, вилучення ознак, дублювання об'єктів міноритарного класу та заснована на технології навчання з підкріпленням. В якості міри визначення незбалансованості набору даних використано показник ступеня дисбалансу. Проведено експеримент з використанням набору даних зображень облич пацієнтів з висипаннями на шкірі, анотовані у відповідності до ступеня тяжкості акне. Акне́ (от др.-греч. ἀκμή – остриё, разгар, расцвет) или угри́ – это длительное воспалительное заболевание кожи, возникающее в ситуации, когда мертвые клетки кожи и кожное сало забивают волосяной фолликул Розглянуто основні кроки реалізації методології. Результати класифікації показали доцільність застосування запропонованої методології. Точність класифікації на тестових даних склала 85 %, що на 5 % вище ніж результат, отримаий без застосування запропонованої методології. Іл.: 1. Табл.: 2. Бібліогр.: 26 назв.

Ключові слова: дисбаланс класів; незбалансований набір даних; вилучення фрагментів зображень; аугментація.

Постановка проблеми. Сучасний розвиток технологій машинного і глибокого навчання дозволяє досягти високого рівня точності при використанні інтелектуального аналізу даних і, зокрема, класифікації зображень в галузі комп'ютерного зору. У цьому випадку для отримання якісної моделі, алгоритми вимагають великої кількості розмічених даних для кожного класу на етапі навчання моделі.

Виявлення і прогнозування подій часто включає в себе класифікацію рідкісних випадків [1]. Рідкісні випадки – це події, які трапляються з невеликою частотою, але можуть мати далекосяжні наслідки [2].

Рідкісні випадки, представлені в наборах даних зображень, можуть мати різні форми, включаючи об'єкти дикої природи [3], ураження агрикультурних рослин, [4], випадки, що пов'язані з діяльністю людини, такі як кібервторгнення з використанням зображень [5], дефекти фасадів будівель [6] тощо. Дані медичних зображень можуть демонструвати нерівномірний розподіл класів в разі рідкісних клінічних випадків, захворювань, що обумовлює труднощі з формуванням збалансованого

набору даних для навчання, оскільки деякі захворювання є досить рідкісними. Як, наприклад, це представлено в роботі [7] в якій автори досліджують набір даних зображень мікроскопії гістологічних препаратів злоякісних та доброякісних захворювань молочної залози.

Рідкісні випадки призводять до проблеми незбалансованості даних, а саме незбалансованості кількості об'єктів в різних класах. Незбалансовані дані відносяться до набору даних, в якому один або декілька класів містять набагато більшу кількість прикладів, ніж інші. Превалюючий клас називають мажоритарним класом, а нечисленний по об'єктах клас – міноритарний клас [8]. Використання незбалансованих наборів даних для навчання моделі класифікації має негативний вплив на точність моделей та може привести до отримання некоректних або помилкових результатів класифікації.

Аналіз останніх досліджень і публікацій. Серед методів усунення проблеми незбалансованості даних виділяють дублювання об'єктів міноритарного класу і видалення об'єктів мажоритарного класу (oversampling та undersampling). Ці методи є універсальними, оскільки вони не залежать від обраного класифікатора, але, в той же час, можуть бути застосовані лише до структурованих даних. Також, використання дублювання об'єктів міноритарного класу і видалення об'єктів мажоритарного класу має деякі недоліки, які потенційно можуть перешкоджати навчанню моделі. У разі видалення об'єктів мажоритарного класу проблема пов'язана з тим, що видалення об'єктів з класу більшості може призвести до того, що модель пропустить важливі концепції, що відносяться до класу більшості. При використанні дублювання об'єктів міноритарного класу, проблема пов'язана з тим, що дублювання об'єктів міноритарного класу відтворює наявні об'єкти до початкового набору даних, що призводить до перенавчання [9].

Інша група методів представлена методами класифікації з урахуванням витрат (cost-sensitive classification). Зокрема, автори дослідження [10] запропонували застосування перехресної ентропії надаючи перевагу результатам класифікації з вищою ймовірністю та нейтралізуючи результати з низькою ймовірністю класу. Методи класифікації з урахуванням витрат зосереджені на модифікації існуючих алгоритмів класифікації для посилення їх здатності вчитися на об'єктах міноритарного класу [11] та мають залежність від використовуваних алгоритмів.

Зображення є неструктурованими даними. Концепція усунення незбалансованості даних зображень за рахунок їх збільшення називається аугментацією даних, при якій зображення зазнає різні трансформації, але при цьому зберігає інформацію. Трансформації зображення з метою

аугментації включають в себе масштабування, обрізання, відображення, заповнення, поворот, зміну яскравості, контрастності, рівня насиченості тощо.

Одним з напрямків усунення незбалансованості навчального набору зображень є приведення зображень до їх структурованого представлення за допомогою навчання з підкріпленням. Навчання з підкріпленням полягає у використанні навченої моделі для вилучення ознак об'єктів зображення, отримуючи таким чином структуровані дані, до яких можна застосувати метод дублювання об'єктів міноритарного класу, збалансувавши навчальний набір даних, який в якості збалансованого навчального набору даних передається далі на вхід результуючої моделі навчання. Метод дублювання об'єктів міноритарного класу полягає в збільшенні кількості вибірок класу меншини шляхом випадкового копіювання вибірок меншини, щоб збалансувати кількість об'єктів класу меншини і класу більшості [12]. Одним з широко уживаних методів є дублювання об'єктів міноритарного класу для вирішення проблем дисбалансу є генерація навчальних даних шляхом лінійної інтерполяції для класів меншини.

Використання навчання з підкріпленням для вилучення ознак і подальшого їх дублювання об'єктів міноритарного класу з метою усунення незбалансованості даних досліджується в роботах [13, 14].

Проведений аналіз досліджень показав, що при класифікації в умовах даних, що містять рідкісні випадки, некоректні результати класифікації можуть призвести до серйозних проблем [15]. Нейронні мережі, зазвичай, менше схильні до впливу дисбалансу класів і шуму, ніж алгоритми машинного навчання, але й вони не захищені від некоректних результатів класифікації при навчанні моделі на незбалансованому наборі даних [16].

Аналіз досліджень в області класифікації незбалансованих даних показав, що метод дублювання об'єктів міноритарного класу є універсальним методом збалансування набору даних. Обмеження щодо застосування цього методу може бути усунене за рахунок використання технології навчання з підкріпленням. Також, для попередження перенавчання моделі показано збільшення набору даних зі збереженням інформативності. Ці факти обумовлюють актуальність поставленої мети та вибір методів її реалізації.

Мета статті. Метою дослідження є розроблення методології усунення дисбалансу класів даних зображень.

Класифікація незбалансованих даних. Для реалізації завдань класифікації на етапі навчання моделі, зображення набору даних повинні

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

бути анотовані. Нехай S – набір анотованих зображень набору навчальних даних, що складається з N об'єктів, тобто $|S| = N$, який можна визначити наступним чином $S = \{(a_i, b_i)\}$, $i = 1, ..., N$, де $a_i \in A$ – об'єкт в n-мірному просторі ознак $A=\{a_1, a_2, ..., a_n\}$, та $b_i \in B = \{1, \ldots, C\}$ – мітка ідентифікатора класу, пов'язана з об'єктом a_i.

В залежності від розподілу класів в наборі даних S завдання навчання моделі класифікації з кількістю класів $C \geq 2$ може бути представлене як: 1) завдання збалансованої класифікації, 2) завдання незбалансованої класифікації з більшістю мажоритарних класів або більшістю міноритарних класів [19].

Збалансована класифікація має місце при класифікації набору даних S з рівномірним розподілом об'єктів C класів. В іншому випадку, завдання розглядається як незбалансована класифікація.

Якщо ми говоримо про незбалансований набір даних, то, також, визначаємо підмножини $S_{min} \subset S$ та $S_{maj} \subset S$, де S_{min} – множина меншини об'єктів класів в S, а S_{maj} – це множина об'єктів переважаючого класу в S, так що $S_{min} \cap S_{maj} = \{\Phi\}$ та $S_{min} \cup S_{maj} = \{S\}$.

Незбалансована класифікація має місце при класифікації набору даних S з нерівномірним розподілом об'єктів C класів в разі більшості S_{maj}, або меншості S_{min}.

Визначення ступеня дисбалансу класів. Для вимірювання нерівномірності розподілу об'єктів класів набору навчальних даних використовується ступінь незбалансованості (Imbalance-Degree – ID), запропонована в дослідженні [19] і розглянута авторами [20]. *ID* – міра незбалансованості класів, розроблена для мультіклассових завдань класифікації, яка розраховується наступним чином

$$ID = \frac{d(J,H)}{d(I,H)} + (C_{\min} - 1), \tag{1}$$

де $d(\cdot, H)$ – статистична відстань між розподілами J або I та H. позначає вектор дійсних розподілів J_C класів C в наборі даних S, де

$$J_C = \frac{N_C}{N} \tag{2}$$

та N_C – кількість об'єктів в кожному класі C, а N – загальна кількість об'єктів, що використовуються для навчання. Вектор $H = [h_1, h_2, ..., h_C]$ являє собою окремий випадок рівноймовірності розподілу класів, тобто

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

розподіл об'єктів класів збалансованого набору даних з розподілом об'єктів кожного класу рівним $H_C = \frac{1}{C}$. $C_{\min}$ – кількість міноритарних класів, дійсний розподіл об'єктів кожного з яких $J_C < \frac{1}{C}$. Нарешті, I – розподіл міноритарного класу, представлений у вигляді вектору $[i_1, i_2, ..., i_C]$, найвіддаленішого $C_{\min}$ від H, або вектор класів декількох $C_{\min}$ з $I_c = 0$ в кількості C – $C_{\min}$ –1 при $I_C = \frac{1}{C}$. Виходячи з цього, отримуємо розподіл одного міноритарного класу, що дорівнює $I_C = 1 - \frac{C - C_{\min} - 1}{C}$. Збалансований набір даних буде мати $ID = 0$, а незбалансовані набори даних матимуть вищі значення ID. В якості міри відстані $d(\cdot, H)$ можуть бути використані різні критерії, наприклад, такі як Евклідова відстань, відстань Чебишева, Хелінгера тощо. Таким чином, для наборів даних, що мають $ID > 0$ робиться висновок про незбалансованість набору даних і рекомендовано використання методів для збалансування класів набору даних.

Формально, це може бути представлено в наступному вигляді.

$$S = \begin{cases} \text{збалансований, якщо } ID = 0, \\ \text{незбалансований, в іншому випадку} \end{cases} \quad (3)$$

Визначення простору ознак зображення. Нехай зображення f є відображенням з просторової області E в діапазон V, тобто f: $E \to V$. Тоді, кожен елемент $x \in E$ відображається на значення $f(x) \in V$, тобто $x \in E \to f(x) \in V$. Область зображення – це безперервний Евклідовий простір $E = \mathrm{R}^n$, де n – розмірність простору ознак [17]. У випадку двовимірного зображення $E = \mathrm{R}^2$.

Таким чином, з урахуванням вилучених ознак, зображення f може бути представлене в наступним чином $f : E \subset \mathrm{Z}^l \to T \subset \mathrm{R}^n$, де E – двовимірна площина, що представляє просторову область зображення, l розподільна здатність зображення, n розмірність простору ознак, і T непорожній набір багатовимірних векторів зображення. Просторова область відноситься до самої площини зображення, а методи в просторовій області засновані на прямій зміні значення пікселів. Кожному пікселю $x_i \in E$ зображення відповідає вектор $v_i = f(x_i)$.

Простір ознак зображення може бути отриманий з використанням коефіцієнтів інтенсивності колірних моделей, координат об'єкту зображення [18].

При використанні коефіцієнтів інтенсивності колірних моделей для кольорового зображення в кожному пікселі $x \in \mathrm{R}^2$ вимірюється три значення: значення інтенсивності червоного, зеленого і синього кольорів. В цьому випадку, зображення можна представити як векторну функцію:

$$x \in \mathrm{R}^2 \to f(x) = r(x)g(x)b(x) \in \mathrm{R}^3 .$$

Чорно-біле зображення є скалярним зображенням з варіацією інтенсивності значень сірого кольору. У кожному $x \in \mathrm{R}^2$ вимірюється інтенсивність випромінювання. Чорно-біле зображення може бути представлене функцією

$$x \in \mathrm{R}^2 \to f(x) \in \mathrm{R} .$$

Вилучення фрагментів. Для уникнення перенавчання моделі на етапі лінійної інтерполяції, запропоновано використання вилучення фрагментів зображення для збільшення різноманітності даних.

Для зображення *f* набору навчальних даних *S* фрагмент можна представити наступним чином. Позначивши векторний простір *P* фрагментів шириною *w* відповідних пікселям зображення *f*, фрагмент можна описати наступним чином $F_w : E \subset \mathrm{Z}^l \to P \subset \mathrm{R}^{nw^2}$, з чого отримуємо $p_i^w = F_w(x_i) = \left(f(x_i + t), \forall t \in \left[-w/2, w/2\right]^2\right)^T$.

Таким чином, при вилученні фрагментів зображення, отримуємо на виході векторний простір *P*, представлений як набір даних фрагментів $p^w(v_i)$ вилучених з *f*. Отже, на цьому етапі новий набір даних представлений таким чином (4)

$$S' :\ S \to P \subset P_{\min} \cup P_{\mathrm{maj}}, \tag{4}$$

де $P_{\min}$ – фрагменти, отримані із зображень міноритарного класу, а P_{maj} – фрагменти, отримані із зображень мажоритарного класу.

Аугментація даних. Аугментація даних [20] може бути представлена як генерація нових даних x′ з використанням певної функції трансформації $\varphi(\cdot)$ як це представлено в (5)

$$(x') \leftarrow \varphi(x_i), \tag{5}$$

де x_i – об'єкт навчального набору даних *S'*. Кожен об'єкт x_i може бути одновимірним або багатовимірним, тобто може бути представлений у вигляді вектору або матриці.

В якості функції трансформації запропоновано використання трансляції [21], яка являє собою процес переміщення об'єкта з одного положення в інше на зображенні.

Таким чином, на основі *P* при використанні трансляції отримуємо аугментований набір $U = \left\{x_1^{'}, x_2^{'}, ..., x_m^{'}\right\} \subset P$, що складається з *m* аугментованих фрагментів.

Отже, припускаючи, що вихідним набором даних є новий набір даних з урахуванням простору ознак фрагментів *P*, аугментація даних може бути представлена наступним чином

$$P \to U \subset U_{\min} \cup U_{\text{maj}}, \tag{6}$$

де *U* – аугментований набір *P*, $U_{\min}$ – фрагменти, отримані шляхом трансляції фрагментів зображень міноритарного класу, а U_{maj} – фрагменти, отримані шляхом трансляції фрагментів зображень мажоритарного класу. Потім новий набір даних збільшується за рахунок об'єднання вихідного набору і аугментованого набору даних

$$S'\text{:}\ P \cup U. \tag{7}$$

Лінійна інтерполяція класів меншини. В якості методу усунення незбалансованості класів навчального набору даних використовується лінійна інтерполяція класів меншості [12]. Ці дані генеруються шляхом випадкового вибору одного або декількох *k*-найближчих сусідів для об'єктів класу меншини в такий спосіб.

У множині аугментованих об'єктів міноритарного класу U_{min}, для кожного $x \in U_{\min}$ *k*-найближчі сусіди *x* отримуються шляхом обчислення міри відстані між *x* і кожним іншим об'єктом в множині об'єктів міноритарного класу $U_{\min}$. Кількість згенерованих об'єктів встановлюється відповідно до ступеня дисбалансу. Для кожного $x \in U_{\min}$, *q* об'єктів $x_1, x_2, ..., x_q$ $(q \le k)$ випадковим чином вибираються з *k*-найближчих сусідів, та вони становлять множину G_{min}. Далі для кожного об'єкту $x_k \in G_{\min}$ $(k = 1,2,...,q)$, наступна формула використовується для створення нового об'єкту (8)

$$x' = x + r(0,1) * \left\| x - x_k \right\|, \tag{8}$$

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

де $r(0, 1)$ є випадковим числом від 0 до 1.

Таким чином, припускаючи, що вихідним набором даних є новий набір даних з урахуванням простору аугментованих ознак A, генерування даних може бути представлене в такий спосіб

$$U \to G \subset G_{\min} \cup G_{\text{maj}}, \tag{9}$$

де U – аугментований набір P, G – об'єкти, згенеровані на основі набору даних U, можуть бути представлені як G, з підмножинами $G_{\min}$ та G_{maj}, такими, що $G_{\min} \cap G_{\text{maj}}$, які представляють підмножину $G_{\min}$ – згенеровані об'єкти міноритарного класу, і підмножину G_{maj} – згенеровані об'єкти мажоритарного класу. Далі новий набір даних збільшується за рахунок об'єднання вихідного набору фрагментів, аугментованого набору і набору згенерованих об'єктів

$$S' : P \cup U \cup G. \tag{10}$$

Отриманий набір даних далі використовується для класифікації в якості навчального набору даних для моделі.

Результати експерименту. Для оцінки ефективності запропонованої методології, проведено експеримент з відкритим набором даних медичних зображень ACNE04 [22]. Набір даних містить 1457 зображень обличь і експертні анотації відповідно до японської шкали оцінки кожних висипів. Навчальний набір складається з 1165 зображень з чотирма класами тяжкості висипів: 0 легкий ступінь – 410 об'єктів, 1 середній ступінь – 506 об'єктів, 2 важкий ступінь – 146 об'єктів та 3 дуже важкий ступінь – 103 об'єкта. Тестовий набір містить 291 зображення, який складається з 103 об'єктів класу 0 легкий ступінь, 127 об'єктів – клас 1 середній ступінь, 362 об'єкта – клас 2 важкий ступінь і 6 об'єктів – клас 3 дуже важкий ступінь.

Проведено визначення ступеня дисбалансу класів навчального набору даних з використанням ступеня дисбалансу *ID*

$$J_{Class0} = \frac{410}{1165} = 0.3519,\ J_{Class1} = \frac{506}{1165} = 0.4343,$$
$$J_{Class2} = \frac{146}{1165} = 0.1253,\ J_{Class3} = \frac{103}{1165} = 0.0884.$$

За умов рівноймовірності класів H = 1/4 = 0.25. Кількість міноритарних класів, дійсний розподіл об'єктів кожного з яких $J_C < \frac{1}{C}$

дорівнює 3. Найбільш віддаленим від H міноритарним класом є клас 3. Таким чином, розподіл I міноритарного класу 3 визначається наступним чином.

$$I = 1 - \frac{4-3-1}{4} = 1.$$

На підставі отриманих значень визначення ступеня дисбалансу досліджуваного набору даних за допомогою евклидової відстані представлене наступним чином.

$$ID = \frac{d(0.0884, 0.25)}{d(1, 0.25)} + (3-1) = 2.216.$$

Таким чином, ID ≠ 0, що є підставою для висновку про незбалансованість набору даних.

Для вилучення фрагментів використані дві попередньо навчені моделі: shape_predictor_68_face_landmarks [23] та One Eye model [24]. В результаті реалізації етапу вилучення фрагментів з 1165 зображень для навчання моделі вилучено 3806 фрагментів зображень. Розподіл вилучених фрагментів по класах представлено наступним чином: клас 0 – 1367 зображень, клас 1 – 1716 зображень, клас 2 – 476 зображень, клас 3 – 247 зображень.

Після проведення аугментации для реалізації якої використана ковзна трансляція фрагментів розподіл патчів по класах представлено наступним чином: 0 – 3556, класу 1 – 4333, класу 2 – 1843 класу 3 – 1514 зображень.

Етап вилучення ознак реалізований з використанням ResNet-152 [25].

Лінійна інтерполяція класів меншини реалізована з використанням Synthetic Minority Oversampling Technique (SMOTE) [26]. На цьому етапі згенеровані ознаки об'єктів класів меншини. Кількість об'єктів в кожному класі визначено за кількістю об'єктів мажоритарного класу – класу 1. Таким чином, збалансовано кількість об'єктів кожного з класів, яка становить 4333 об'єктів в кожному класі. Інформація про розподіл даних на кожному етапі представлені в табл.1.

ISSN 2079-0031 (Print) ISSN 2411-0558 (Online)

Таблиця 1

Розподіл об’єктів в класах на кожному етапі запропонованої методології

Клас	Зображення	Вилучення фрагментів	Аугментація	Дублювання
Клас 0	410	1367	3556	4333
Клас 1	506	1716	4333	4333
Клас 2	146	476	1843	4333
Клас 3	103	247	1514	4333
Всього	1165	3806	11246	17332

Діаграми розподілу кількості об'єктів навчального набору даних по класах представлені на рис. 1, де (a) розподіл класів вихідних зображень, (b) розподіл класів вилучених фрагментів, (c) розподіл класів вилучених ознак з аугментованих фрагментів, (d) розподіл класів вилучених ознак в результаті реалізації дублювання для класів меншини.

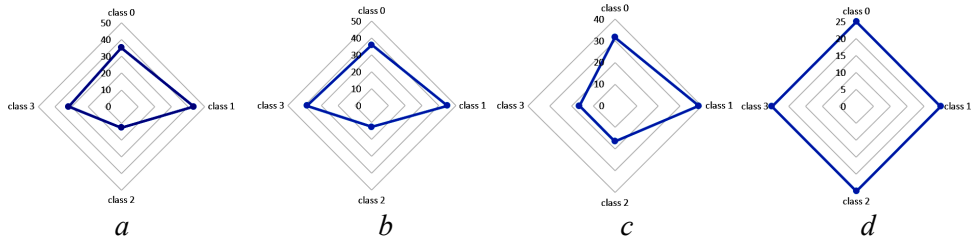

Рис. 1. Діаграми розподілу класів об'єктів

Далі використано технології навчання з підкріпленням. Отримані на етапі реалізації дублювання методом лінійної інтерполяції дані використані для навчання моделі згорткової нейронної мережі для отримання оцінки ступеня тяжкості акне по зображеннях особи. Оцінка якості моделі проведена та реалізована на даних тестового набору з використанням критеріїв *Precision, Recall, F1-score, Accuracy*. Отримані результати представлені в табл. 2.

Таблиця 2

Отримані результати класифікації

Критерій	Клас	Вихідний набір даних	Набір даних отриманий з використанням запропонованої методології
Precision	0	0.76	0.78
	1	0.76	0.73
	2	0.82	0.91
	3	0.95	0.99
	Середнє	0.8225	0.8225
Recall	0	0.75	0.79
	1	0.77	0.77
	2	0.81	0.86
	3	0.96	0.97
	Середнє	0.8225	0.8475
F1-score	0	0.76	0.79
	1	0.77	0.75
	2	0.81	0.88
	3	0.96	0.98
	Середнє	0.825	0.85
Accuracy		0.8	0.85

Судячи з результатів, представлених в табл. 1, використання запропонованого підходу дозволило підвищити критерії точності Accuracy, Recall, F1-score.

Таким чином, можна говорити про доцільність використання запропонованого підходу для усунення незбалансованості набору даних.

Висновки. Метою дослідження було розроблення методології усунення дисбалансу класів даних зображень. Запропонована методологія включає наступні етапи: вилучення фрагментів зображень, аугментація даних фрагментів, вилучення ознак та дублювання об'єктів міноритарного класу з використанням лінійної інтерполяції. Запропонована методологія заснована на технології навчання з підкріпленням та включає етап визначення незбалансованості набору даних з застосуванням ступеня дисбалансу *ID*. Проведено експеримент з застосування запропонованої методології з використанням набору даних зображень обличь пацієнтів з кожними висипами, анотовані у відповідності до ступеня тяжкості акне. Застосовані критерії якості

класифікації показали вищу точність класифікації при застосуванні запропонованої методології. Зокрема, точність класифікації на тестових даних становила 85 %, що на 5 % вище ніж без застосування запропонованої методології. Результати класифікації свідчать про перевагу та доцільність застосування запропонованої методології.

Список літератури:

1. Weiss, G.M. and Hirsh, H. Learning to predict extremely rare events. In AAAI workshop on learning from imbalanced data sets. – Austin : AAAI Press, 2000, p. 64-68.
2. King, G., & Zeng, L. (2001). Logistic Regression in Rare Events Data. Political Analysis, 9(2), p. 137-163. doi:10.1093/oxfordjournals.pan.a004868.
3. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., & Yan, J. Large-scale object detection in the wild from imbalanced multi-labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, p. 9709-9718.
4. Sambasivam, G., & Opiyo, G. D. A predictive machine learning application in agriculture: Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. Egyptian Informatics Journal, 2021, 22(1), p. 27-34.
5. Yilmaz, I., Masum, R., Siraj, A. Addressing imbalanced data problem with generative adversarial network for intrusion detection. In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI), 2020, p. 25-30.
6. Guo, J., Wang, Q., Li, Y., Liu, P. Façade defects classification from imbalanced dataset using meta learning-based convolutional neural network. Computer-Aided Civil and Infrastructure Engineering, 2020, 35(12), p.1403-1418.
7. Saini, M., Susan, S. Deep transfer with minority data augmentation for imbalanced breast cancer dataset. Applied Soft Computing, 2020, 97, 106759.
8. Yijing, L., Haixiang, G., Xiao, L., Yanan, L., Jinling, L. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. Knowledge-Based Systems, 2016, 94, p. 88-104. doi:10.1016/j.knosys.2015.11.013.
9. Liu Z., Cao H., Chen X., et al., Multi-fault classification based on wavelet SVM with PSO algorithm to analyze vibration signals from rolling element bearings, Neurocomputing, 2013, 99 (1), p. 399-410.
10. Kim, Y., Lee, Y., Jeon, M. Imbalanced image classification with complement cross entropy. Pattern Recognition Letters, 2021, 151, p. 33-40. doi:10.1016/j.patrec.2021.07.017.
11. Guo H.X., Liao X.W., Zhu K.J., Optimizing reservoir features in oil exploration man agement based on fusion of soft computing, Appl. Soft Comput, 2011, 11, p. 1144-1155.
12. Xie, W., Liang, G., Dong, Z., Tan, B., Zhang, B. An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data. Mathematical Problems in Engineering, 2019.
13. Huang Z., Dumitru C. O., Pan Z., Lei B. and Datcu M. Classification of Large-Scale High-Resolution SAR Images With Deep Transfer Learning, in IEEE Geoscience and Remote Sensing Letters, 2021, 1(18), p. 107-111. doi: 10.1109/LGRS.2020.2965558.
14. Singh R., Ahmed T., Kumar A., Singh A. K., Pandey A. K., Singh S. K. Imbalanced Breast Cancer Classification Using Transfer Learning, in IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 1(18), p. 83-93. doi: 10.1109/TCBB.2020.2980831.
15. Tang Y, Zhang Y, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 2009, 391, 281-288. DOI: 10.1109/TSMCB.2008.2002909.

16. Khoshgoftaar Taghi M, Van Hulse Jason, Napolitano Amri, Supervised Neural Network Modeling: An Empirical Investigation Into Learning From Imbalanced Data With Labeling Errors. IEEE Transactions on Neural Networks, 2010, 21(5), p. 813-830. doi:10.1109/tnn.2010.2042730.
17. Rein van den Boomgaard Lecture Notes "Image Processing and Computer Vision" URL : https//staff.fnwi.uva.nl/r.vandenboomgaard/IPCV20172018/LectureNotes/IP/Images/ImageDefinition.html (дата звернення 30.09.2021 р.).
18. Lezoray O. Patch-based mathematical morphology for image processing, segmentation and classification // International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, Cham, 2015, p. 46-57.
19. Ortigosa-Hernandez J, Inza I, Lozano J.A. Measuring the class-imbalance extent of multi-class problems. Pattern Recogn. Letters. 2017, 98, p. 32-38.
20. Iwana, B. K., Uchida, S. An empirical survey of data augmentation for time series classification with neural networks. Plos one, 2021, 16(7), p. e0254841.
21. Vyas A., Yu S., Paik J. Fundamentals of digital image processing. Multiscale Transforms with Application to Image Processing. Springer, Singapore, 2018, p. 3-11.
22. Wu, X., Wen, N., Liang, J., Lai, Y.K., She, D., Cheng, M.M. Yang, J. Joint acne image grading and counting via label distribution learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, p. 10642-10651. doi: 10.1109/ICCV.2019.01074.
23. Davisking/dlib-models. Github.com. URL: https://github.com/davisking/dlib-models. (дата звернення 30.09.2021).
24. OpenCV. Github.com. URL: https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_eye.xml. (дата звернення 30.09.2021).
25. The Microsoft Cognitive Toolkit. URL: https://www.cntk.ai/Models/Caffe_Converted/ResNet152_ImageNet_Caffe.model (дата звернення 30.09.2021).
26. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, p. 321-357.

References:

1. Weiss, G.M. and Hirsh, H. (2000), Learning to predict extremely rare events. *In AAAI workshop on learning from imbalanced data sets*. – Austin : AAAI Press, p. 64-68.
2. King, G., and Zeng, L. (2001), Logistic Regression in Rare Events Data. *Political Analysis*, 9(2), p. 137-163. doi:10.1093/oxfordjournals.pan.a004868.
3. Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., and Yan, J. (2020), Large-scale object detection in the wild from imbalanced multi-labels. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 9709-9718.
4. Sambasivam, G., & Opiyo, G. D. (2021), A predictive machine learning application in agriculture: *Cassava disease detection and classification with imbalanced dataset using convolutional neural networks. Egyptian Informatics Journal*, 22(1), p. 27-34.
5. Yilmaz, I., Masum, R., Siraj, A. (2020), Addressing imbalanced data problem with generative adversarial network for intrusion detection. *In 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, p. 25-30.
6. Guo, J., Wang, Q., Li, Y., Liu, P. (2020), Façade defects classification from imbalanced dataset using meta learning-based convolutional neural network. *Computer-Aided Civil and Infrastructure Engineering*, 35(12), p.1403-1418.
7. Saini, M., Susan, S. (2020), Deep transfer with minority data augmentation for imbalanced breast cancer dataset. *Applied Soft Computing*, 97, 106759.

8. Yijing, L., Haixiang, G., Xiao, L., Yanan, L., and Jinling, L. (2016), Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94, p. 88-104. doi:10.1016/j.knosys.2015.11.013.
9. Liu Z., Cao H., Chen X., et al., (2013), Multi-fault classification based on wavelet SVM with PSO algorithm to analyze vibration signals from rolling element bearings, *Neurocomputing*, 99 (1), p. 399-410.
10. Kim, Y., Lee, Y., Jeon, M. (2021), Imbalanced image classification with complement cross entropy. *Pattern Recognition Letters*, 151, p. 33-40. doi:10.1016/j.patrec.2021.07.017.
11. Guo H.X., Liao X.W., Zhu K.J. (2011), Optimizing reservoir features in oil exploration man agement based on fusion of soft computing, *Appl. Soft Comput*, 11, p. 1144-1155.
12. Xie, W., Liang, G., Dong, Z., Tan, B., Zhang, B. (2019), An improved oversampling algorithm based on the samples' selection strategy for classifying imbalanced data. *Mathematical Problems in Engineering*.
13. Huang Z., Dumitru C.O., Pan Z., Lei B. and Datcu M. (2021), Classification of Large-Scale High-Resolution SAR Images With Deep Transfer Learning, *in IEEE Geoscience and Remote Sensing Letters*, 1(18), p. 107-111. doi: 10.1109/LGRS.2020.2965558.
14. Singh R., Ahmed T., Kumar A., Singh A. K., Pandey A. K., Singh S. K. (2021), Imbalanced Breast Cancer Classification Using Transfer Learning, *in IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(18), p. 83-93. doi: 10.1109/TCBB.2020.2980831.
15. Tang Y, Zhang Y, Chawla NV, Krasser S. (2009), SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics. 391, 281-288. DOI: 10.1109/TSMCB.2008.2002909.
16. Khoshgoftaar Taghi M, Van Hulse Jason, Napolitano Amri (2010), Supervised Neural Network Modeling: An Empirical Investigation Into Learning From Imbalanced Data With Labeling Errors. *IEEE Transactions on Neural Networks*, 21(5), p. 813-830. doi:10.1109/tnn.2010.2042730.
17. Rein van den Boomgaard (2017), Lecture Notes "Image Processing and Computer Vision" URL : https://staff.fnwi.uva.nl/r.vandenboomgaard/IPCV20172018/LectureNotes/IP/Images/ImageDefinition.html (access 30.09.2021 p.)
18. Lezoray O. (2015), Patch-based mathematical morphology for image processing, segmentation and classification // *International Conference on Advanced Concepts for Intelligent Vision Systems. Springer*, Cham, p. 46-57.
19. Ortigosa-Hernandez J, Inza I, Lozano JA. (2017), Measuring the class-imbalance extent of multi-class problems. *Pattern Recogn. Letters*. 98, p. 32-38.
20. Iwana, B. K., and Uchida, S. (2021), An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7), p. e0254841.
21. Vyas A., Yu S., and Paik J. (2018), Fundamentals of digital image processing. *Multiscale Transforms with Application to Image Processing. Springer*, Singapore, p. 3-11.
22. Wu, X., Wen, N., Liang, J., Lai, Y.K., She, D., Cheng, M.M. Yang, J. (2019), Joint acne image grading and counting via label distribution learning. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* p. 10642-10651. doi: 10.1109/ICCV.2019.01074.
23. Davisking/dlib-models. Github.com. URL: https://github.com/davisking/dlib-models. (access 30.09.2021).
24. OpenCV. Github.com. URL: https://github.com/opencv/opencv/blob/master/data/haarcascades/haarcascade_eye.xml. (access 30.09.2021).

25. The Microsoft Cognitive Toolkit. URL: https://www.cntk.ai/Models/Caffe_Converted/ResNet152_ImageNet_Caffe.model (access 30.09.2021).
26. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, p. 321-357.

Статтю представив д.т.н., проф. Національного технічного університету "Харківський політехнічний інститут" С.Ю. Леонов

Надійшла (received) 5.10.2021

Biloborodova Tetiana, Cand.Sci.Tech, Assotiate Professor
G.E. Pukhov Institute for Modelling in Energy Engineering
15 General Naumov Street, Kyiv, 03164, Ukraine
e-mail: beloborodova.t@gmail.com
ORCID ID: 0000-0001-7561-7484

Skarga-Bandurova Inna, D.Sci.Tech., Professor
G.E. Pukhov Institute for Modelling in Energy Engineering
15 General Naumov Street, Kyiv, 03164, Ukraine
e-mail: skarga_bandurova@ukr.net
ORCID ID: 0000-0003-3458-8730

Koverha Mark, PhD student
Volodymyr Dahl East Ukrainian National University
59-a Central Avenue, Severodonetsk, Luhansk region, Ukraine, 93400
tel./phone: (064) 522-89-97, e-mail: healthunder@gmail.com

УДК 004.932.2

Методологія усунення дисбалансу класів наборів даних зображень / Білобородова Т.О., Скарга-Бандурова І.С., Коверга М.О. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2021. – № 2 (6). – С. 114 – 129.

Представлено методологію вирішення задачі усунення дисбалансу класів в наборах даних зображень, яка включає етапи вилучення фрагментів зображень, аугментацію фрагментів, вилучення ознак, та дублювання об'єктів міноритарного класу. В якості міри визначення незбалансованості набору даних використано ступінь дисбалансу. Проведено експеримент з використанням набору даних зображень облич пацієнтів з висипаннями на шкірі, що анотовані у відповідності до ступеня тяжкості акне. Розглянуто основні кроки реалізації запропонованої методології. Результати класифікації показали доцільність застосування запропонованої методології. Точність класифікації на тестових даних склала 85%, що на 5% вище ніж без застосування запропонованої методології. Іл.: 1. Табл.: 2. Бібліогр.: 26 назв.

Ключові слова: дисбаланс класів; незбалансований набір даних; виділення фрагментів; аугментація.

УДК 004.932.2

Методология устранения дисбаланса классов наборов данных изображений / Белобородова Т.А., Скарга-Бандурова И.С., Коверга М.А. // Вестник НТУ "ХПИ". Серия: Информатика и моделирование. – Харьков: НТУ "ХПИ". – 2021. – № 2 (6). – С. 114 – 129.

Предложена методология решения задачи устранения дисбаланса классов в наборах данных изображений. В качестве меры определения несбалансированности набора данных использовано степень дисбаланса. Проведен эксперимент с использованием набора данных изображений лиц пациентов с кожной сыпью, аннотированного в соответствии со степенью тяжести акне. Рассмотрены основные шаги реализации предложенной методологии. Результаты классификации показали целесообразность применения предложенной методологии. Точность на тестовых данных составила 85%, что на 5% выше результатов, полученных без применения предложенной методологии. Ил.: 1. Табл.: 2. Библиогр.: 26 назв.

Ключевые слова: дисболланс клас сов; несбалансированный набор даннях; извлечение фрагментов; аугментация.

UDC 004.932.2

Methodology for correcting class imbalance in image datasets / Biloborodova T.O., Skarga-Bandurova I.S., Koverha M.O. // Herald of the National Technical University "KhPI". Series of "Informatics and Modeling". – Kharkov: NTU "KhPI". – 2021. – № 2 (6). – P. 114 – 129.

A methodology for eliminating class imbalance in image data sets is proposed. It includes image fragment extraction, fragments augmentation, feature extraction, and duplication of minority class objects. The degree of imbalance was used as a measure to determine the imbalance of the data set. An experiment was performed using a data set of facial images of patients with facial rash, annotated according to the severity of acne. The main steps of realization of the offered methodology are considered. The results of the classification showed the feasibility of applying the proposed methodology. The accuracy of classification on test data was 85%, which is 5% higher than without the application of the proposed methodology. Figs.: 1. Tabl.: 2. Refs.: 26 titles.

Keywords: class imbalance; imbalanced data set; patch extraction; augmentation.