DOI: https://doi.org/10.33216/2222-3428-2021-21-1

УДК 004.855.5:004.6

Білобородова Т.О., Коверга М.О., Петров П.О., Ломакін С.О. Критська Я.О.

ДОСЛІДЖЕННЯ МЕТОДІВ ВИРІШЕННЯ ПРОБЛЕМИ НЕЗБАЛАНСОВАНИХ ДАНИХ

У статті представлено дослідження методів вирішення проблеми незбалансованих даних. Рідкісні події 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. Незбалансовані дані можуть негативно вплинути на точність моделей і привести до отримання некоректних або помилкових результатів класифікації. Методи, спрямовані на вирішення проблеми незбалансованості даних, поділяють на три групи: методи рівня даних, методи рівня алгоритмів і ансамблеві методи. В статті представлена таксономія методів вирішення проблеми незбалансованості даних. До методів вирішення проблеми негативного впливу незбалансованості даних на результати класифікації на рівні даних віднесено дублювання об'єктів міноритарного класу, видалення об'єктів мажоритарного класу, гібридні методи. В якості методів на рівні алгоритмів, як найбільш широко поширені, визначають методи на основі алгоритму SVM, нейронних мереж та байєсовського алгоритму. Ансамблеві методи поділяють на методи на основі бустинг алгоритмів, ансамблеві методи на основі дублювання об'єктів міноритарного класу, ансамблеві методи на основі видалення об'єктів мажоритарного класу. Формалізовано явище незбалансованості даних. Представлено структури використання методів вирішення проблеми негативного впливу незбалансованості даних для кожного підходу. Представлено аналіз критеріїв оцінки результатів класифікації незбалансованих даних, що поділяються на критерії на основі номінальної оцінки, критерії на основі числової оцінки, критерії на основі ймовірності прогнозування. Проаналізовано переваги та недоліки розглянутих методів, спрямованих на вирішення проблеми незбалансованості даних та представлено результати цього аналізу. Визначено, що використання методів на рівні даних часто є кращим способом вирішення проблеми незбалансованих даних і, також, їх використання не виключає подальше використання інших методів на рівні алгоритмів або ансамблевих методів, для коректної оцінки результатів використання методів усунення негативного впливу незбалансованості, необхідно вибирати критерії оцінки, які дають краще розуміння того, наскільки добре метод і моделі справляються з поставленими цілями.

***Ключові слова**: незбалансовані дані, класифікація, дублювання об'єктів міноритарного класу, видалення об'єктів мажоритарного класу*

Актуальність дослідження. Сучасний розвиток технологій машинного і глибокого навчання дозволяє досягти високого рівня точності при використанні інтелектуального аналізу даних і, зокрема, класифікації даних, в різних областях застосування. Основними завданнями класифікації є виявлення справжніх або прогнозування майбутніх подій на основі історичних даних. 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 визначення рідкісних подій [1]. Рідкісні події - це події, які відбуваються з невеликою частотою, але можуть мати далекосяжні наслідки [2]. Рідкісні події можуть мати різні форми, включаючи стихійні лиха, антропогенні небезпеки, такі як фінансове шахрайство, промислові аварії, насильницькі конфлікти, і захворювання. Медичні дані 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.

Рідкісні події 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 мажоритарних класом, а самий нечисленний по об'єктах клас - міноритарний клас [3]. Незбалансовані дані можуть негативно вплинути на точність моделей і привести до отримання некоректних або помилкових результатів класифікації.

Таким чином, актуальним завданням є дослідження методів вирішення проблеми незбалансованості даних.

Метою статті є дослідження методів вирішення проблеми незбалансованості розподілу даних. Для виконання поставленої мети виділено наступні завдання:

- Визначення явища незбалансованих даних.
- Визначення таксономії, аналіз і порівняння методів, спрямованих на вирішення проблеми негативного впливу незбалансованості даних.
- Аналіз критеріїв оцінки результатів класифікації незбалансованих даних.

Формалізація проблеми незбалансованих даних. Розглядаючи незбалансований набір навчальних даних S с m об'єктів, тобто $|S| = m$, можна визначити наступним чином $S = \{(x_i, y_i), i = 1, ..., m$, де $x_i \in X$ - екземпляр в n-мірному просторі ознак $X=\{f_1, f_2, ..., f_n\}$, та $y_i \in Y = \{1, …, C\}$- мітка ідентифікатора класу,

пов'язана з екземпляром x_i. Зокрема, C = 2 являє собою задачу двокласової класифікації. Крім того, ми 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 $S_{min} \cup S_{maj} = \{S\}$.

Будь-які об'єкти, згенеровані на основі набору даних S, можуть бути позначені як E, з непересічними підмножинами E_{min} та E_{maj}, що представляють меншість і більшість об'єктів E, відповідно, щоразу, коли вони застосовуються.

Методи вирішення проблеми незбалансованості даних. Методи, спрямовані на вирішення проблеми незбалансованості даних, можна розділити на три групи: методи рівня даних, методи рівня алгоритмів і ансамблеві методи [4].

Таксономія методів, спрямованих на вирішення проблеми незбалансованих даних, яка включає найбільш часто зустрічаються рішення, представлена на рис. 1.

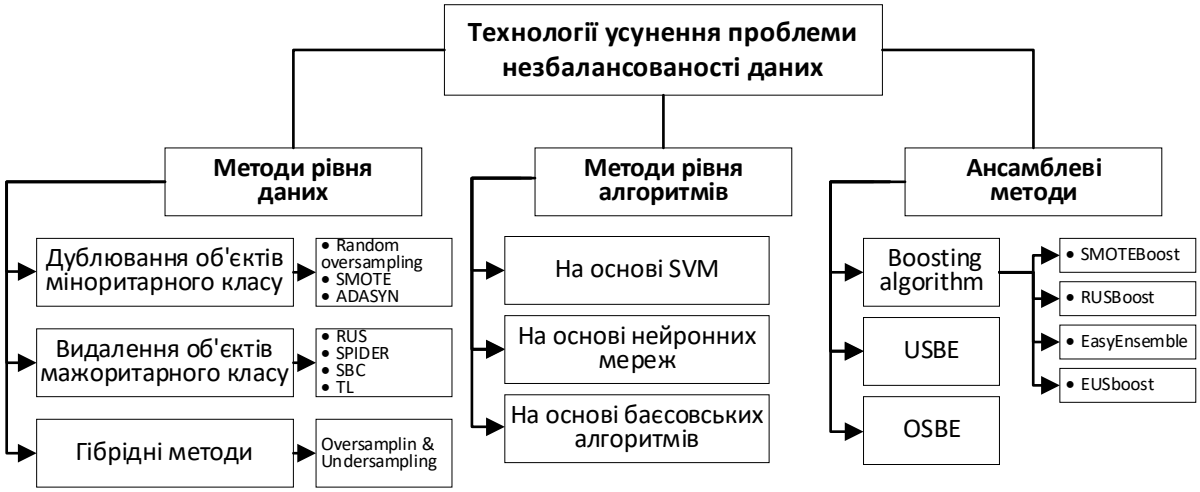


Рисунок 1 – Таксономія методів проблеми незбалансованості даних

До методів вирішення проблеми негативного впливу незбалансованості даних на результати класифікації на рівні даних відносять дублювання об’єктів міноритарного класу (англ. Oversampling), видалення об’єктів мажоритарного класу (англ. Undersampling), гібридні методи. В якості методів на рівні алгоритмів, як найбільш широко поширені, визначають методи на основі алгоритму SVM (Support Vector Machine), нейронних мереж, байєсовського алгоритму. Ансамблеві методи поділяють на методи на основі бустинг алгоритмів, ансамблеві методи на основі дублювання об’єктів міноритарного класу (англ. Over-Sampling Balanced Ensemble, OSBE), ансамблеві методи на основі видалення об’єктів мажоритарного класу (англ. Under-Sampling Balanced Ensemble, USBE).

Детальніше методи вирішення проблеми негативного впливу незбалансованості даних на результати аналізу даних розглянуто далі.

Методи рівня даних. Методи рівня даних зосереджені на зміні розмірів навчальних наборів даних, щоб збалансувати всі види класів. Методи на рівні даних використовуються для збалансування простору даних незбалансованого набору даних з метою пом'якшення ефекту незбалансованості розподілу класів в процесі навчання.

Методи, які використовуються на рівні даних більш універсальні, оскільки вони не залежать від обраного класифікатора. Їх ділять на три групи в залежності від методу, який використовується для рівномірного розподілу класів [3]: дублювання об'єктів міноритарного класу, видалення об'єктів мажоритарного класу, гібридні методи.

Схематично, використання методів на рівні даних можна подати так, як це представлено на рис. 2.

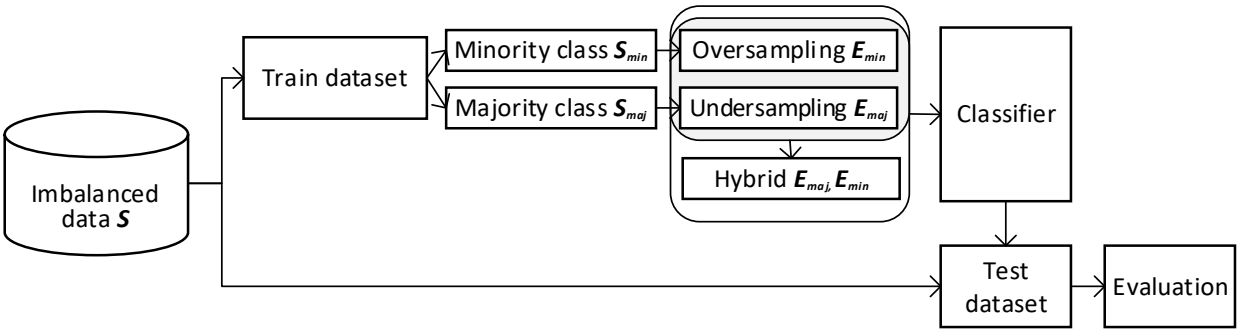


Рисунок 2 – Структура використання методів вирішення проблеми негативного впливу незбалансованості даних на рівні даних

Дублювання об'єктів міноритарного класу: вирішення проблеми негативного впливу нерівномірного розподілу шляхом створення нових об'єктів з класу меншості. Процес дублювання об'єктів міноритарного класу реалізується шляхом додавання набору E, відібраного з класу меншості: для набору випадково вибраних прикладів меншини в S_{min}, необхідно збільшити вихідний набір S, реплікуючи обрані об'єкти і додаючи їх до S. Таким чином, кількість спільних об'єктів в S_{min} збільшується на $|E|$, і баланс розподілу класів S коригується відповідним чином. Це забезпечує механізм для зміни ступеня балансу розподілу класів до будь-якого бажаного рівня.

Широко використовуються методи створення згенерованих об'єктів меншості - це випадкове дублювання зразків меншості (англ. Random oversampling), SMOTE [5], ADASYN [6].

Видалення об'єктів мажоритарного класу: вирішення проблеми негативного впливу нерівномірного розподілу шляхом відкидання об'єктів з класу більшості. У той час як дублювання об'єктів міноритарного класу додає дані до початкового набору даних, випадкове видалення об'єктів мажоритарного класу видаляє дані з вихідного набору даних. Зокрема, випадковим чином вибирається набір об'єктів більшості класу в S_{maj} і видаляється з S, таким чином, що $|S|=|S_{min}|+|S_{maj}|-|E|$. Отже, недостатня вибірка дає нам простий метод настройки балансу вихідного набору даних S.

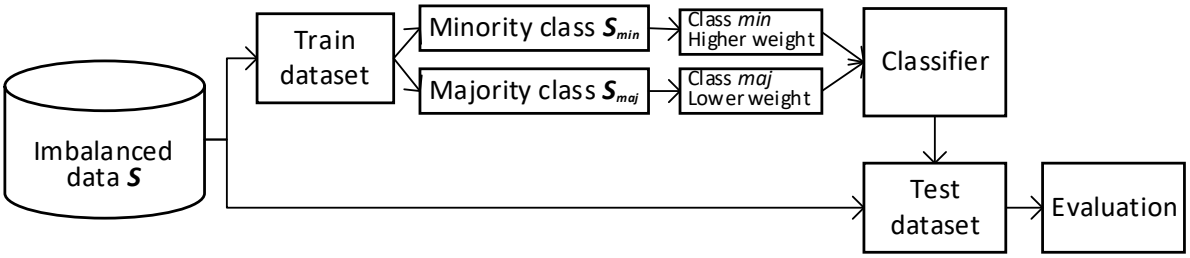
Найпростіший, але найефективніший метод - це метод випадкової вибірки з недостатньою вибіркою (англ. Random Under-Sampling, RUS), який включає випадкове виключення прикладів з класу більшості [7]. Також, до методів видалення об'єктів мажоритарного класу відносяться SPIDER [8], SBC[9], TL[10].

Гібридні методи - це методи зрівноважування даних шляхом комбінування дублювання об'єктів міноритарного класу і видалення об'єктів мажоритарного класу. Кількість даних основного класу скорочується за рахунок використання концепції видалення об'єктів мажоритарного класу, а також кількість даних другорядного класу збільшується шляхом додавання з використанням концепції дублювання об'єктів міноритарного класу. У цьому випадку використовується комбінація вищеописаних методів.

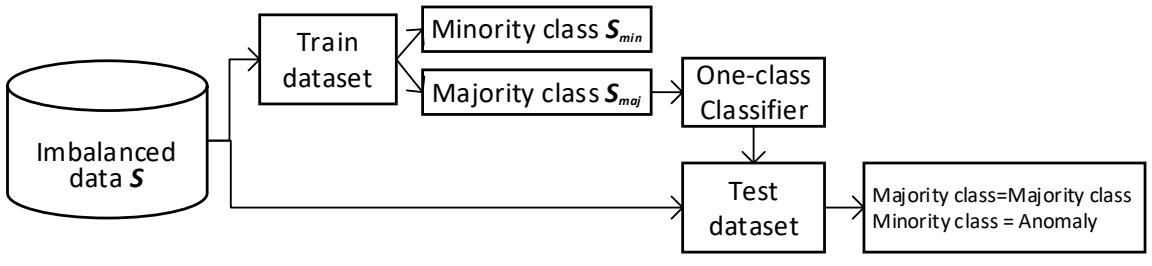
Використання дублювання об'єктів міноритарного класу і видалення об'єктів мажоритарного класу має деякі недоліки, які потенційно можуть перешкоджати навчанню моделі. У разі видалення об'єктів мажоритарного класу проблема пов'язана з тим, що видалення об'єктів з класу більшості може призвести до того, 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 [11].

Методи рівня алгоритмів. Методи рівня алгоритмів зосереджені на модифікації існуючих алгоритмів класифікації для посилення їх здатності вчитися на об'єктах міноритарного класу [12]. У той час як методи рівня даних намагаються збалансувати розподіл класів, методи рівня алгоритмів спрямовані на процес навчання алгоритмів з урахуванням ваг об'єктів, що отримуються в результаті навчання, пов'язаних з неправильною класифікацією прикладів [13]. Більшість алгоритмів цього сімейства засновані на SVM, нейронних мережах та баєсовському алгоритмі.

Виділяють дві основні концепції навчання на незбалансованість даних на рівні алгоритмів. Перша полягає в призначенні більш високих ваг результатами помилкової класифікації об'єктів міноритарного класу, як це представлено на рис. 3(а).



(а)



(б)

Рисунок 3 – Структура використання методів вирішення проблеми негативного впливу незбалансованості даних на рівні алгоритмів

Друга концепція навчання на основі незбалансованих даних - це розгляд об'єктів міноритарного класу як викиди і використання методів виявлення шумів і викидів для моделювання міноритарного класу, наприклад, однокласового класифікатора, як це представлено на рис. 3 (б).

Ансамблеві методи. Ансамбль класифікаторів вважається популярною технологією для боротьби з незбалансованим навчанням, в основному через їх здатності значно поліпшити продуктивність одного класифікатора [14]. Методи ансамблю можна розглядати як побудову системи з декількома класифікаторами, яка об'єднує безліч базових класифікаторів. Для кожного базового класифікатора підходи на рівні даних часто використовуються в якості попередньої обробки.

Схематично, використання ансамблевих методів можна представити таким чином, як це представлено на рис. 4.

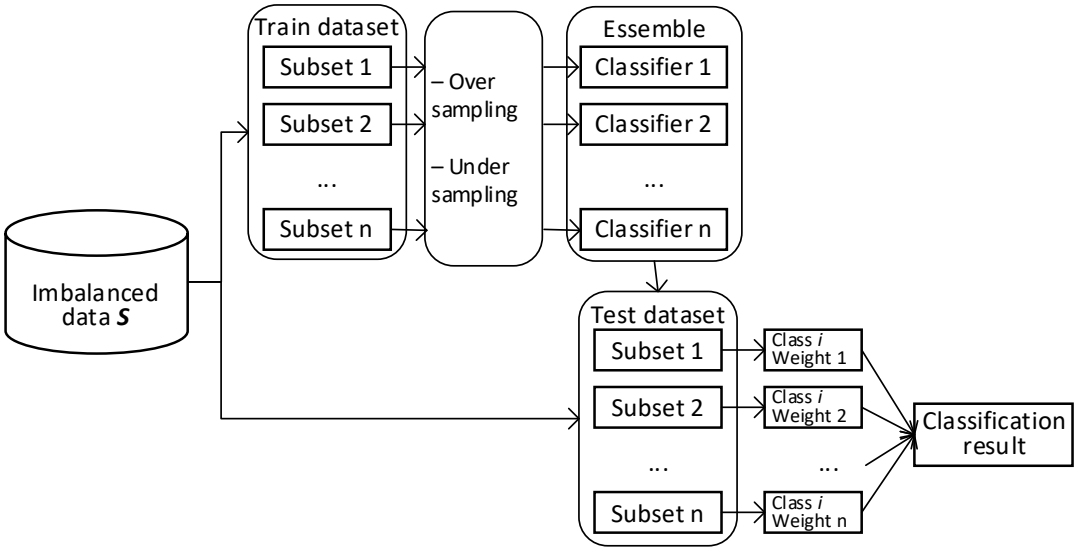


Рисунок 4 – Структура використання ансамблевих методів вирішення проблеми негативного впливу незбалансованості даних

Етап застосування методів рівня даних, таких як дублювання об'єктів міноритарного класу, видалення об'єктів мажоритарного класу є необов'язковим.

Розрізняють три основні групи ансамблевих методів: методи на основі бустінг алгоритмів, ансамблеві методи на основі дублювання об'єктів міноритарного класу (OSBE), ансамблеві методи на основі видалення об'єктів мажоритарного класу (USBE).

Після навчання ансамблю моделей для нового об'єкта з кожної моделі можна отримати кілька індивідуальних гіпотез. Це обумовлює необхідність використання правил ансамблю, які об'єднують ці гіпотези. Автори [3] представили п'ять правил ансамблю для об'єднання результатів множинної класифікації різних класифікаторів, включаючи правило максимуму, правило мінімуму, правило твори, правило голосування більшістю і правило суми.

Найбільш широко використовуваним є бустінг алгоритм, який застосовується в багатьох відомих ансамблевих алгоритмах, таких як SMOTEBoost [15], RUSBoost [16], EasyEnsemble [17], EUSboost [18]. MetaCost [19] створює єдиний чутливий до ваг класифікації класифікатор на основі навчання, що дає переваги швидкої класифікації і інтерпретованих результатів.

Оцінка результатів класифікації незбалансованих даних. Критерій оцінки є ключовим фактором як для оцінки ефективності класифікації, так і для визначення прогресу в навчанні класифікатора. Якість алгоритмів навчання зазвичай оцінюється за результатами класифікації на тестових даних з використанням різних критеріїв оцінки ефективності навчання алгоритму. Залежно від інформації, яка може бути отримана за результатами використання алгоритмів класифікації, або в залежності від її інтерпретації, розрізняють три групи критеріїв [20]: критерії на основі номінальної оцінки, критерії на основі числової оцінки, критерії на основі ймовірності прогнозування.

Критерії на основі номінальної оцінки: прогнозовані мітки класу порівнюються з фактичними істинними значеннями класу для оцінки моделі. Використовуються результати хибнопозитивних, помилково негативні, істинно позитивних, істинно негативних результатів класифікації, представлених в матриці невідповідності.

До критеріїв оцінки класифікації, які, також, дозволяють оцінити ефективність методів, спрямованих на зменшення негативного впливу незбалансованості даних, відносять точність, функцію втрат, коефіцієнт кореляції Метьюса (Matthews correlation coefficient, MCC) [21], F-measure [22], G-Measure.

Точність є найбільш часто використовуваним критерієм. Однак при класифікації незбалансованих даних точність може бути не кращим вибором, тому що часто має тенденцію до класу більшості. Критерієм ефективності, більш адаптованим для роботи з незбалансованими даними, є F-measure, який менш схильний до

негативного впливу незбалансованого розподілу класів, оскільки вимірює виконання класифікації кожного класу незалежно.

Критерії на основі числової оцінки: розглядається деяка числова оцінка, пов'язану з прогнозованими результатами, для оцінки об'єктів тестових даних відповідно до ймовірністю їх приналежності до класу.

До цієї групи відносять такі візуальні критерії, як ROC крива [23] і площа під ROC кривої (Area under curve, AUC), які часто використовуються для оцінки ефективності класифікації.

ROC крива, також, є менш схильним критерієм під негативний вплив незбалансованого розподілу класів завдяки незалежному виміру виконання класифікації кожного класу.

Критерії на основі ймовірності прогнозування: числові результати, пов'язані з прогнозом, інтерпретуються як ймовірності приналежності прикладів до класу. Оцінка імовірнісних оцінок зазвичай виконується за допомогою оцінки Брієра (Brier score) [24]. Основна ідея полягає в тому, щоб обчислити середньоквадратичну помилку між передбаченими оцінками ймовірності і істинним значенням класу, де позитивний клас визначається як 1, а негативний клас як 0. Оцінка Брієра - це середнє значення по всіх об'єктах.

Використання критеріїв на основі номінальної оцінки є найпростішим підходом, який, проте, має обмеження у вигляді неможливості розрізняти прогнози в межах одного і того ж класу, оскільки немає способу розрізняти більш-менш ймовірні номінальні оцінки з однаковим значенням класу. Використання критеріїв на основі ймовірності прогнозування дозволяє розглядати результат ймовірнісно, додаючи деяку ступінь впевненості прогнозом. Критерії на основі числової оцінки знаходиться між цими двома підходами, де оцінки, отримані за допомогою класифікаторів, використовуються для упорядкування прогнозів примірників без прив'язки до ймовірнісної інтерпретації. Однак не існує стандартного способу інтерпретації оцінок, як у випадку критеріїв на основі ймовірності.

Порівняльний аналіз методів вирішення проблеми негативного впливу незбалансованого розподілу даних. Проаналізовано переваги та недоліки різних методів, спрямованих на вирішення проблеми незбалансованості даних. Результати порівняльного аналізу представлені в табл.1. Критерії оцінювання результатів визначено наступним чином: 1 - критерії на основі номінальної оцінки, 2 - критерії на основі чисельної оцінки, 3 - критерії на основі ймовірності прогнозування.

Таблица 1 - Характеристики методов устранения негативного влияние несбалансированного распределения данных

Підхід	Метод	Переваги	Недоліки	Критерії оцінювання
Рівень даних	Дублювання об’єктів міноритарного класу	• Незалежність від алгоритмів класифікації • Використання за умови невеликого набору даних	• Перенавчання моделі • Неефективність на великих даних • Збільшення розміру навчальної вибірки і часу на побудову класифікатора	1, 2, 3
	Видалення об’єктів мажоритарного класу	• Незалежність від алгоритмів класифікації • Економія обчислювальних ресурсів	• Упущення важливих концепцій, що відносяться до класу більшості • Упередженість виборки	1, 2, 3
Рівень алгоритмів	Концепція призначення ваг	• Можливість використання апріорних знань для визначення ваг	• Залежність від алгоритмів класифікації	1, 2, 3
	Концепція використання лише мажоритарного класу	• Зручність застосування для виявлення аномалій	• Відсутність можливісті багатокласової класифікації • Залежність від алгоритмів класифікації	3
Ансамблеві методи	Бустинг, OSBE, USBE	• Можливість використання декількох алгоритмів	• Необхідність додаткових обчислень для отримання результату • Залежність від алгоритмів класифікації	1, 2, 3

Як це можна побачити з табл. 1, кожен з розглянутих методів має переваги та недоліки і не існує універсального методу для вирішення проблеми незбалансованості. Розглянуті в дослідженні методи можуть бути оцінені з використанням критеріїв на основі номінальної оцінки, чисельної оцінки та ймовірності прогнозування, окрім концепції використання лише мажоритарного класу. Результати класифікації з використанням цього методу оцінюються з використанням критерію на основі ймовірності прогнозування.

Висновки. В проведеному дослідженні методів вирішення проблеми незбалансованості розподілу даних

представлено визначення явища незбалансованих даних, визначена таксономії методів вирішення проблеми незбалансованості розподілу даних. Дослідження включило визначення набору методів вирішення проблеми негативного впливу незбалансованості даних за трьома підходами: рівень даних, рівень алгоритмів та ансамблеві методи, визначення набору критеріїв оцінки результатів класифікації незбалансованих даних в залежності від поставленого завдання та використовуваного методу. Визначено переваги та недоліки розглянух підходів та критеріїв. Використання методів на рівні даних часто є кращим способом вирішення проблеми незбалансованих даних і, також, їх використання не виключає подальше використання інших методів на рівні алгоритмів або ансамблевих методів. Для коректної оцінки результатів використання методів вирішення проблеми негативного впливу незбалансованості, необхідно вибирати критерії оцінки, які дають краще розуміння того, наскільки добре метод і моделі справляються з поставленими цілями. Ключовим рішенням при виборі того чи іншого методу є мета класифікації. Точне розуміння мети допоможе подолати проблеми з незбалансованим набором даних і забезпечить отримання найкращих можливих результатів.

Література

1. Weiss G. M., Hirsh H. Learning to predict extremely rare events //AAAI workshop on learning from imbalanced data sets. – Austin : AAAI Press, 2000. – P. 64-68.
2. King G., Zeng L. Logistic regression in rare events data //Political analysis. – 2001. – Vol. 9. – №. 2. – P. 137-163. doi:10.1093/oxfordjournals.pan.a004868.
3. Yijing L. et al. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data //Knowledge-Based Systems. – 2016. – Vol. 94. – P. 88-104. doi:10.1016/j.knosys.2015.11.013.
4. Haixiang G. et al. Learning from class-imbalanced data: Review of methods and applications //Expert Systems with Applications. – 2017. – Vol. 73. – P. 220-239. doi:10.1016/j.eswa.2016.12.035.
5. Chawla N. V. et al. SMOTE: synthetic minority over-sampling technique //Journal of artificial intelligence research. – 2002. – Vol. 16. – P. 321-357..
6. He H. et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning //2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). – IEEE, 2008. – P. 1322-1328.
7. Tahir M. A. et al. A multiple expert approach to the class imbalance problem using inverse random under sampling //International workshop on multiple classifier systems. – Springer, Berlin, Heidelberg, 2009. – P. 82-91.
8. Stefanowski J., Wilk S. Selective pre-processing of imbalanced data for improving classification performance //International Conference on Data Warehousing and Knowledge Discovery. – Springer, Berlin, Heidelberg, 2008. – P. 283-292.
9. Yen S. J., Lee Y. S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset //Intelligent Control and Automation. – Springer, Berlin, Heidelberg, 2006. – P. 731-740.
10. Tomek I., Two modifications of CNN. IEEE Trans. Syst. Man Cybernet, 1976. – Vol. 6. – P. 769–772.
11. Liu Z. et al. Multi-fault classification based on wavelet SVM with PSO algorithm to analyze vibration signals from rolling element bearings //Neurocomputing. – 2013. – Vol. 99. – P. 399-410.
12. Haixiang G. et al. Optimizing reservoir features in oil exploration management based on fusion of soft computing //Applied Soft Computing. – 2011. – Vol. 11. – №. 1. – P. 1144-1155.
13. Elkan C. The foundations of cost-sensitive learning //International joint conference on artificial intelligence. – Lawrence Erlbaum Associates Ltd, 2001. – Vol. 17. – №. 1. – P. 973-978.
14. Guo H., Viktor HL Boosting with data generation: improving the classification of hard to learn examples // International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. - Springer, Berlin, Heidelberg, 2004. - P. 1082-1091.
15. Chawla N. V. et al. SMOTEBoost: Improving prediction of the minority class in boosting //European conference on principles of data mining and knowledge discovery. – Springer, Berlin, Heidelberg, 2003. – P. 107-119.
16. Seiffert C. et al. RUSBoost: A hybrid approach to alleviating class imbalance //IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. – 2009. – Vol. 40. – №. 1. – P. 185-197.
17. Liu T. Y. Easyensemble and feature selection for imbalance data sets //2009 international joint conference on bioinformatics, systems biology and intelligent computing. – IEEE, 2009. – P. 517-520.
18. Galar M. et al. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling //Pattern recognition. – 2013. – Vol. 46. – №. 12. – P. 3460-3471.
19. Domingos P. Metacost: A general method for making classifiers cost-sensitive //Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. – 1999. – P. 155-164.
20. Fernández A. et al. Learning from imbalanced data sets. – Berlin : Springer, 2018. – Vol. 11. doi: 10.1007 / 978-3-319-98074-4.
21. Matthews B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme //Biochimica et Biophysica Acta (BBA)-Protein Structure. – 1975. – Vol. 405. – №. 2. – P. 442-451.
22. Haixiang G. et al. Optimizing reservoir features in oil exploration management based on fusion of soft computing //Applied Soft Computing. – 2011. – Vol. 11. – №. 1. – P. 1144-1155.
23. Fawcett T. An introduction to ROC analysis //Pattern recognition letters. – 2006. – Vol. 27. – №. 8. – P. 861-874.

24. Predd J. B. et al. Probabilistic coherence and proper scoring rules //IEEE Transactions on Information Theory. – 2009. – Vol. 55. – №. 10. – P. 4786-4792.

R e f e r e n c e s

1. Weiss G. M., Hirsh H. Learning to predict extremely rare events //AAAI workshop on learning from imbalanced data sets. – Austin : AAAI Press, 2000. – P. 64-68.
2. King G., Zeng L. Logistic regression in rare events data //Political analysis. – 2001. – Vol. 9. – №. 2. – P. 137-163. doi:10.1093/oxfordjournals.pan.a004868.
3. Yijing L. et al. Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data //Knowledge-Based Systems. – 2016. – Vol. 94. – P. 88-104. doi:10.1016/j.knosys.2015.11.013.
4. Haixiang G. et al. Learning from class-imbalanced data: Review of methods and applications //Expert Systems with Applications. – 2017. – Vol. 73. – P. 220-239. doi:10.1016/j.eswa.2016.12.035.
5. Chawla N. V. et al. SMOTE: synthetic minority over-sampling technique //Journal of artificial intelligence research. – 2002. – Vol. 16. – P. 321-357..
6. He H. et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning //2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). – IEEE, 2008. – P. 1322-1328.
7. Tahir M. A. et al. A multiple expert approach to the class imbalance problem using inverse random under sampling //International workshop on multiple classifier systems. – Springer, Berlin, Heidelberg, 2009. – P. 82-91.
8. Stefanowski J., Wilk S. Selective pre-processing of imbalanced data for improving classification performance //International Conference on Data Warehousing and Knowledge Discovery. – Springer, Berlin, Heidelberg, 2008. – P. 283-292.
9. Yen S. J., Lee Y. S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset //Intelligent Control and Automation. – Springer, Berlin, Heidelberg, 2006. – P. 731-740.
10. Tomek I., Two modifications of CNN. IEEE Trans. Syst. Man Cybernet, 1976. – Vol. 6. – P. 769–772.
11. Liu Z. et al. Multi-fault classification based on wavelet SVM with PSO algorithm to analyze vibration signals from rolling element bearings //Neurocomputing. – 2013. – Vol. 99. – P. 399-410.
12. Haixiang G. et al. Optimizing reservoir features in oil exploration management based on fusion of soft computing //Applied Soft Computing. – 2011. – Vol. 11. – №. 1. – P. 1144-1155.
13. Elkan C. The foundations of cost-sensitive learning //International joint conference on artificial intelligence. – Lawrence Erlbaum Associates Ltd, 2001. – Vol. 17. – №. 1. – P. 973-978.
14. Guo H., Viktor HL Boosting with data generation: improving the classification of hard to learn examples // International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. - Springer, Berlin, Heidelberg, 2004. - P. 1082-1091.
15. Chawla N. V. et al. SMOTEBoost: Improving prediction of the minority class in boosting //European conference on principles of data mining and knowledge discovery. – Springer, Berlin, Heidelberg, 2003. – P. 107-119.
16. Seiffert C. et al. RUSBoost: A hybrid approach to alleviating class imbalance //IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. – 2009. – Vol. 40. – №. 1. – P. 185-197.
17. Liu T. Y. Easyensemble and feature selection for imbalance data sets //2009 international joint conference on bioinformatics, systems biology and intelligent computing. – IEEE, 2009. – P. 517-520.
18. Galar M. et al. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling //Pattern recognition. – 2013. – Vol. 46. – №. 12. – P. 3460-3471.
19. Domingos P. Metacost: A general method for making classifiers cost-sensitive //Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. – 1999. – P. 155-164.
20. Fernández A. et al. Learning from imbalanced data sets. – Berlin : Springer, 2018. – Vol. 11. doi: 10.1007 / 978-3-319-98074-4.
21. Matthews B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme //Biochimica et Biophysica Acta (BBA)-Protein Structure. – 1975. – Vol. 405. – №. 2. – P. 442-451.
22. Haixiang G. et al. Optimizing reservoir features in oil exploration management based on fusion of soft computing //Applied Soft Computing. – 2011. – Vol. 11. – №. 1. – P. 1144-1155.
23. Fawcett T. An introduction to ROC analysis //Pattern recognition letters. – 2006. – Vol. 27. – №. 8. – P. 861-874.
24. Predd J. B. et al. Probabilistic coherence and proper scoring rules //IEEE Transactions on Information Theory. – 2009. – Vol. 55. – №. 10. – P. 4786-4792.

В статье представлено исследование методов решения проблемы несбалансированных данных. Редкие события приводят к проблеме несбалансированности данных, а именно несбалансированности количества объектов в разных классах. Несбалансированные данные относятся к набору данных, в котором один или несколько классов содержат гораздо большее количество объектов, чем другие. Несбалансированные данные могут негативно повлиять на точность моделей и привести к получению некорректных или ошибочных результатов классификации. Методы, направленные на решение проблемы несбалансированности данных, разделяют на три группы: методы уровня данных, методы уровня алгоритмов и ансамблевые методы. В статье представлена таксономия методов решения проблемы несбалансированности данных. К методам

решения проблемы негативного влияния несбалансированности данных на результаты классификации на уровне данных отнесены дублирование объектов миноритарного класса, удаление объектов мажоритарного класса и гибридные методы. В качестве методов на уровне алгоритмов, как наиболее широко распространенные, выделяют методы на основе алгоритма SVM, нейронных сетей и байесовский алгоритма. Ансамблевые методы делятся на методы на основе бустинг алгоритмов, ансамблевые методы на основе дублирования объектов миноритарного класса, ансамблевые методы на основе удаления объектов мажоритарного класса. Формализовано явление несбалансированности данных. Представлены структуры использования методов решения проблемы негативного влияния несбалансированности данных для каждого подхода. Представлен анализ критериев оценки результатов классификации несбалансированных данных, которые делятся на критерии на основе номинальной оценки, критерии на основе числовой оценки, критерии на основе вероятности прогнозирования. Проанализированы преимущества и недостатки рассмотренных методов, направленных на решение проблемы несбалансированности данных и представлены результаты этого анализа. Определено, что использование методов на уровне данных часто является лучшим способом решения проблемы несбалансированных данных и, также, их использование не исключает дальнейшее использование других методов на уровне алгоритмов или ансамблевых методов, для корректной оценки результатов использования методов устранения негативного влияния несбалансированности, необходимо использовать критерии оценки, которые дают лучшее понимание того, насколько хорошо метод и модели справляются с поставленными целями.
Ключевые слова: *несбалансированные данные, классификация, дублирование объектов миноритарного класса, удаление объектов мажоритарного класса*

The article presents a study of methods for problem solving of imbalanced data. Rare events lead to the problem of data imbalance, namely the imbalance in the number of objects in different classes. Imbalanced data refers to a dataset in which one or more classes contain many more objects than others. Imbalanced data can negatively affect the accuracy of the classification models and lead to incorrect or erroneous results. Methods aimed at solving the problem of data imbalance are divided into three groups: data-level methods, algorithm-level methods, and ensemble methods. The article presents a taxonomy of methods for solving the problem of imbalanced data. The methods for solving the imbalanced data problem on the classification results at the data–level include oversampling, undersampling, and hybrid methods. The most widespread algorithms–level methods are methods based on the SVM algorithm, neural networks and the Bayesian algorithm. Ensemble methods are divided into methods based on boosting algorithms, over-sampling balanced ensemble, under-sampling balanced ensemble. The definintion of imbalanced data is formalized. Structures of using methods for solving the problem of the negative impact of imbalanced data for each approach are presented. The analysis of classification evaluation criterias for imbalanced data is presented. These criterias are divided into criterias based on a nominal rating, criterias based on a numerical rating, criterias based on the prediction probability. The advantages and disadvantages of the considered methods aimed at solving the imbalanced data problem are analyzed and the results of this analysis are presented. It has been determined that the use of methods at the data–level is often the best way to solve the imbalanced data problem, and, also, their use does not exclude the further use of other methods at the algorithm–level or ensemble methods; which provide a better understanding of how well the method and models are performing at the stated objectives.
Keywords*: imbalanced data, classification, oversampling, undersampling*

Білобородова Т.О. – докторант ІПМЕ ім. Г.Є.Пухова, кандидат технічних наук, доцент, ORCID ID: 0000-0001-7561-7484
Коверга М.О. – аспірант кафедри комп'ютерних наук та інженерій СНУ ім. В.Даля, ORCID ID: 0000-0001-9906-4845
Петров П.О. – аспірант кафедри комп'ютерних наук та інженерій СНУ ім. В.Даля, ORCID ID: 0000-0002-5672-1072
Ломакін С.О. – студент кафедри комп'ютерних наук та інженерій СНУ ім. В.Даля
Критська Я.О. - старший викладач кафедри комп'ютерних наук та інженерій СНУ ім. В.Даля, ORCID ID: 0000-0003-4575-2559