УДК 004.67:618.3

ПІДХОДИ ДО КЛАСИФІКАЦІЇ НЕЗБАЛАНСОВАНИХ ТА ЗСУНУТИХ НАБОРІВ ДАНИХ

Білобородова Т.О., Скарга-Бандурова І.С.

APPROACHES FOR CLASSIFICATION OF IMBALANCED AND SKEWED DATASETS

Biloborodova T., Skarga-Bandurova I.

Розглянуто проблему вибору підходу, алгоритму для бінарної класифікації незбалансованих та зсунутих наборів даних. Проаналізовано математичні моделі, виділені властивості кожного з підходів. Проведена класифікація з використанням алгоритмів, що представляють кожен з підходів. Виконана перевірка якості та ефективності моделей з використанням таких показників, як помилки першого та другого роду, частка істинно позитивних та істинно негативних спостережень, чутливість та специфічність побудованих моделей для тестового набору даних. За допомогою Biased Minimax Probability Machine (BMPM), отримані характеристики класифікації даних, що враховують властивості даних. Підтверджено якість мінімаксного підходу до класифікації незбалансованих даних.

Ключові слова: *лінійна бінарна класифікація, розрізнювальний, породжувальний, мінімаксний підходи, незбалансовані дані*

Опис проблеми та аналіз літературних даних. На даний момент існує безліч підходів до класифікації даних, що використовують різноманітні алгоритми. Якість, точність та ефективність класифікації даних безпосередньо залежить від відповідності обраного алгоритму властивостям досліджуваних даних. Вибір підходу – це пошук компромісу між якістю класифікації та інтерпретованістю отриманих результатів. У цьому контексті, особливу увагу потребує класифікація наборів даних, що містять різноманітні дефекти, такі як пропуски значень, незбалансованість, нерівноміність розподілу класів в навчальних вибірках та ін. Бінарна класифікація для вирішення реальних проблем, що виникають в промисловості, наприклад, при виявленні дефектів продукту або технологічного обладнання, в медицині - при діагностиці деяких важких захворювань, часто проводиться на зсуненому та незбалансованому наборі даних, де спостережень, що належать одному з класів, набагато менше, ніж тих, що належать іншому класу. Класифікація навчального набору з великим зсувом зазвичай видає моделі не схильні до перенавчання, які не можуть перелаштуватися до тестових даних і не в змозі врахувати їх важливі закономірності. Як наслідок, алгоритми правильно класифікують спостереження, що належать більшості, але показують дуже низьку якість класифікації для спостережень з класом, що є в меншості.

Одним з методів подолання проблеми зсунених та незбалансованих даних є розробка та випробування нових підходів та алгоритмів, або, у більшості випадків, їх комбінації. Так, проблема бінарної класифікації незбалансованого набору даних була розглянута у [1], де автори спробували збалансувати вихідний набір даних і підвищити точність класифікації завдяки поєднанню процедур передискретизації та відміни дискретизації. Удосконалення розрізнювального підходу з використанням алгоритму SVM до класифікації незбалансованих даних виконано у роботі [2]. Запропонований алгоритм поєднує бінарний SVM та однокласний SVM з Гаусовою радиальною базисною функцією. Автори [3] запропонували для вирішення проблеми незбалансованих даних ансамбль класифікаторів. В дослідженні відстежено залежність якості класифікації від комбінації алгоритмів в ансамблі та характеристики даних. Автори дослідження [4] набору незбалансованих даних електрокардіограм пацієнтів для виявлення апное довели переваги використання методу передискретизації даних з використанням Гаусового розподілу в якості функції визначення розподілу. В роботі [5] запропонований різновид мінімаксного підходу до класифікації незбалансованих даних - Biased Minimax Probability Machine, який є першим кількісним методом контролю за тим, як рішення гіперплоскості змінюється на користь класифікації

більш важливого класу, для вирішення упереджених задач класифікації.

Зважаючи на суттєві здобутки, питання розробки якісної моделі для класифікації незбалансованих даних залишається відкритим. Складність отримання точної моделі полягає в тому, що коректний вибір підходу повністю залежить від даних. Те, що працює в одному випадку, може виявитися абсолютно некоректним з іншим набором даних. Евристичний підхід, з великою ймовірністю, дасть низьку прогностичну ефективність. Цілеспрямований вибір підходу, алгоритму допоможе забезпечити при класифікації максимально можливу для досліджуваних даних точність.

Постановка задачі. Таким чином, для визначення найкращої моделі в роботі ставиться задача провести аналіз відповідності підходів до класифікації даних з урахуванням їх властивостей та виконати оцінку ефективності моделей класифікації для досліджуваного набору даних.

Підходи до класифікації даних. Для лінійної бінарної класифікації даних використовуються два широко відомих підходи - класифікація за допомогою породжувальних моделей (generative model) та класифікація з використанням розрізнювальних (умовних) моделей (discriminative models). З метою оцінювання відповідності підходів до класифікації даних з урахуванням їх властивостей, таких як тип даних, величина досліджуваного набору, кількість вхідних параметрів, кількість класів вихідної змінної, наявність зсуву, незбалансованості, відсутність даних, тощо, проведений аналіз останніх досліджень [6, 7, 8, 9], за результатами якого, можна зробити висновок, що в деяких випадках [7, 8] проаналізовані підходи однаково підходять до різних типів даних та для класифікації наборів даних з бінарною вихідною змінною [6] (табл. 1).

Алгоритми породжувального підходу показують кращу якість моделі при використанні для великих наборів даних з великою кількістю вхідних параметрів [8, 10, 11]. Це є необхідною умовою при їх використанні для отримання високих показників якості отриманої моделі. Алгоритми розрізнювального підходу, навпаки, показують кращі результати при класифікації невеликих наборів даних з обмеженою кількістю вхідних змінних [10, 11]. Дослідження в порівнянні цих підходів і визначення найбільш якісного методу, залишаються актуальним завданням обробки даних [12]. Також, невирішене питання якості використання породжувального підходу для невеликих наборів даних. Крім того, для реальних даних досі не знайдений теоретично правильний загальний критерій вибору між розрізнювальним та породжувальним підходами до класифікації даних.

MPM являє собою розрізнювальний класифікатор, заснований на породжувальних попередніх знаннях. Він дозволяє безпосередньо оцінити вірогідну точність, мінімізуючи максимальну ймовірність помилкової класифікації. MPM створює класифікатор, який забезпечує найгіршу оцінку ймовірності помилкової класифікації майбутніх тестових даних на підставі надійних оцінок середнього і коваріаційних матриць класів, отриманих при класифікації навчального набору даних. MPM знаходить межу між середніми значеннями класів. Використання мінімаксного підходу для мультикласової класифікації [9] будується на використанні стратегії "один проти всіх", генетичного алгоритму та параметричної редукції. Автори визначили стандартне відхилення та середню помилку для різних модифікацій мінімаксного підходу та на підставі порівняння цих параметрів зробили висновок про перевагу запропонованого ними підходу. Дослідження по порівнянню запропонованого мінімаксного підходу до мультикласової класифікації даних з розрізнювальним та породжувальним підходами авторами не було проведене і тому не може свідчити про його перевагу. Загалом, для мінімаксного підходу кількість вхідних змінних не має значення, тому в цьому сенсі він є універсальним. Перевагами мінімаксного підходу є те, що він може працювати з урахуванням розподілу, припускаючи, що середнє і коваріація, які безпосередньо оцінюються по даним, достовірно представляють реальне середнє коваріації даних, або без будь-якого припущення про розподіл даних, будуючи класифікатор безпосередньо з даних.

Таблиця 1

Відповідність підходів класифікації властивостям даних

Підходи	Різний тип даних	Велика кількість вхідних змінних	Кількість вхідних змінних обмежена	Бінарна вихідна змінна	Мультикласова вихідна змінна	Великі дані	Невеликі набори даних	Зсунені, незбалансовані дані	Відсутні значення
Породжувальний	+	+		+			+		
Розрізнювальний	+		+	+	+	+			+
Мінімаксний	+	+	+	+			+	+	+

За наявності відсутніх значень в проаналізованих роботах автори використовують розрізнювальний [7] або мінімаксний підходи [5]. Вид мінімаксного підходу BMPM [6, 13] показує високу якість отриманої моделі при використанні для класифікації зсунених або незбалансованих даних. Разом з тим, доцільність використання мінімаксного підходу для великих даних також не доведена.

Припущення. На підставі вищевказаного, висунене теоретичне припущення про важливість урахування властивостей даних при виборі підходу до лінійної бінарної класифікації даних, за якими мінімаксний підхід має перевагу перед розрізнювальним та генеративним підходами при класифікації даних, що характеризуються зсувами, незбалансованністю та мають відсутні значення.

Опис досліджуваних даних. Прикладом задачі класифікації зсунених та незбалансованих даних з відсутніми значеннями є задача прогнозування гіпоксії новонародженого. Цей стан виникає при відхиленні певних показників перебігу вагітності, що нотуються лікарем протягом всього періоду. Даний стан зустрічається досить нечасто, але може призводити до таких тяжких наслідків як дитячий церебральний параліч або, навіть, смерть новонародженого. Прогнозування гіпоксії новонародженого по поєднанню певних значень показників перебігу вагітності, що сигналізують про можливий розвиток гіпоксії у майбутньої дитини допоможе лікарям своєчасно прийняти міри по наданню спеціалізованої допомоги та уникнути або зменшити вплив можливих негативних наслідків для новонародженого.

В якості залежних змінних, які є найбільш специфічними для досліджуваного стану новонародженого (наявність або відсутність гіпоксії), використані показники перебігу вагітності, відібрані в результаті розвідувального аналізу досліджуваних даних [14], який дозволив виділити статистично значущі характеристики - показники перебігу вагітності, і, за рахунок цього, зменшити початкову кількість показників з 29 до 6. Решта показників: ступінь зрілості плаценти при ультразвуковому дослідженні на 30-38 тижнях вагітності, товщина плаценти при ультразвуковому дослідженні на 30-38 тижнях вагітності, протромбіновий індекс крові, вертикальний розмір амніотичної рідини при ультразвуковому дослідженні на 30-38 тижнях вагітності, швидкість осідання еритроцитів крові на 21 тижні вагітності, швидкість осідання еритроцитів крові на 30 тижні вагітності, вказують лише на їх релевантність до діагнозу новонародженого, не надаючи інформації про значення цих показників, що нададуть змогу з достатнім ступенем впевненості, діагностувати наявність гіпоксії у новонародженого. Саме значення релевантних показників перебігу вагітності є предикторами виникнення патології.

Особливості вихідних даних. Вихідна залежна змінна, що містить інформацію про діагноз новонародженого, є номінальною. Вихідна змінна є незбалансованою. Випадків новонароджених з патологією зазвичай набагато менше ніж новонароджених з відсутністю патології. Вхідні змінні – швидкість осідання еритроцитів крові на 21, 30 тижнях вагітності, вертикальний розмір амніотичної рідини і товщина плаценти при ультразвукове дослідження на 30-38 тижнях вагітності є кількісними. Змінні протромбіновий індекс і ступінь зрілості плаценти - порядкові. У даній роботі досліджуються безперервні вхідні змінні, тому змінні протромбіновий індекс і ступінь зрілості плаценти виключені.

Задача дослідження – оцінка ефективності досліджуваних моделей класифікації для визначення станів новонародженого, визначення найкращої в сенсі якості та точності моделі для досліджуваних даних. Для вирішення цієї задачі надалі розглядається характеристика досліджуваних підходів.

Математична модель лінійної бінарної класифікації. При вирішенні лінійних задач бінарної класифікації [6], дані представляються у вигляді елементів векторного простору. Завдання класифікації зводиться до зіставлення кожного вектора значень вхідних змінних *x* значенню вихідної змінної *y*, визначальною поділ об'єктів на класи.

Для досліджуваних даних, бінарна класифікація - це класифікація бінарної вихідної змінної - діагноз новонародженого. Клас $y_1 = 1$ відповідає новонародженим, яким при народженні поставлений діагноз гіпоксія, а $y_2 = -1$ - новонароджені, в яких при народженні відсутня гіпоксія.

Досліджувані дані перебігу вагітності та діагнозу новонародженого представляють собою набір даних, що містить вхідні змінні $(x_1, x_2, \dots x_m) \in R^m$ та відповідні їм значення $y_m = \{1, -1\}$, які є міткою класу для значень показників об'єкта x_m.

Задача класифікації полягає в знаходженні вирішального правила, яке по заданому вектору значень вхідних змінних *x* вказує до якого класу *y* належить відповідне спостереження *i*. Побудова такого правила еквівалентна розбиттю простору ознак на множину областей, що не перетинаються. Графічне відображення лінійної бінарної класифікації представлене на рис. 1.

Гіперплоскість рішення $w^T z = b$ ($w \neq 0, z \in \mathbb{R}, b \in \mathbb{R}$, індекс T означає транспонування) розділяє дані на два класи. Рішення про відповідність певного спостереження до певного класу приймається якщо відповідний спостереженню вектор значень вхідних змінних належить до певної області. Вирішальне правило, яке ще називають класифікатором, вказує до якої

області простору ознак належить вектор значень змінних *x*.

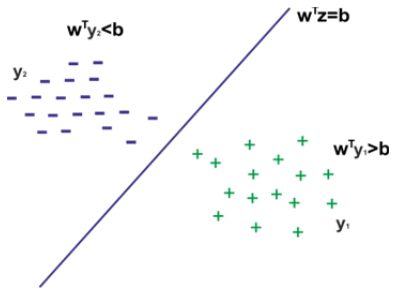


Рис. 1. Лінійна бінарна класифікація

Розрізнювальний та породжувальний підходи до класифікації даних. Фундаментальна відмінність між розрізнювальними та породжувальними моделями полягає в наступному. Розрізнювальні моделі вивчають межу між класами, а породжувальні моделі моделюють розподіл окремих класів. Візуальне представлення підходів продемонстровано на рис. 2.

Адаптуючи [11] до наших даних, різницю між цими підходами можна описати наступним чином. Вихідна змінна досліджуваних даних має два класи стану новонародженого: з патологією ($y_1 = 1$) та без досліджуваної патології ($y_2 = -1$) та вхідні змінні x – показники перебігу вагітності. Беручи до уваги модель, отриману за допомогою навчального набору даних, алгоритм намагається знайти границю рішення, яка відокремлює новонароджених с патологією та новонароджених без досліджуваної патології. Для того, щоб класифікувати нові дані перебігу вагітності для виявлення майбутнього стану новонародженого, алгоритм перевіряє до якої сторони рішення належить випадок, та пропонує прогнозування у відповідності до цього. Така стратегія застосовується при класифікації з використанням алгоритмів розрізнювального підходу.

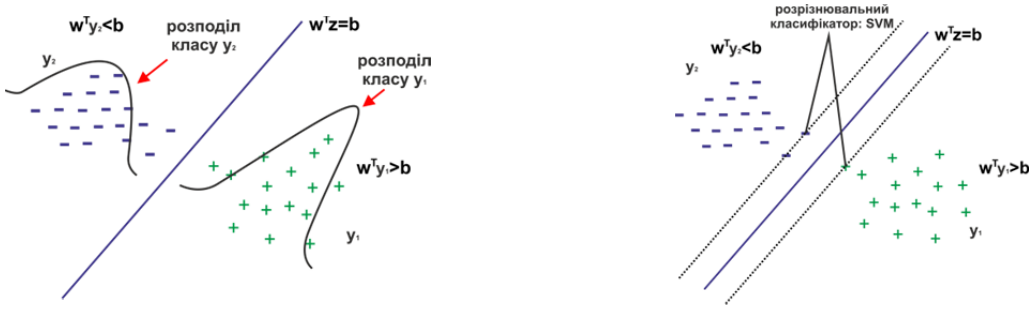
Класифікація за допомогою алгоритмів породжувального підходу відбувається наступним чином. По-перше, маючи дані спостереження перебігу вагітності з наявністю патології у новонародженого, будується модель показників перебігу вагітності з патологією новонародженого. Потім, маючи дані спостереження перебігу вагітності без патології у новонародженого, можна побудувати окрему модель показників перебігу вагітності для новонародженого без патології. Для класифікації нового спостереження перебігу вагітності, необхідно зіставити нові показники перебігу вагітності з моделлю перебігу вагітності з патологією новонародженого, щоб побачити чи схожі нові показники на показники перебігу вагітності з патологією чи без патології, які використовувались в навчальному наборі.

Породжувальні класифікатори [11] моделюють спільний розподіл $p(x, y)$ вхідних змінних x і вихідних змінних y, факторизовані у вигляді $p(x \mid y)p(y)$ та вивчають параметри моделі шляхом максимизації ймовірності, що задана $p(x \mid y)p(y)$.

При побудуванні моделі з використанням породжувального підходу алгоритм використовує приховані параметри, оцінюючи припущення і розподіл моделі. Ця інформація використовується для прогнозування невідомих даних, оскільки передбачається, що модель отримана за допомогою навчального набору реальних даних. Прикладом породжувальної моделі є байєсовський класифікатор.

Недоліком породжувального підходу є те, що розподіл навчального та тестового наборів може відрізнятися, що ставить під сумнів якість класифікації.

Розрізнювальні класифікатори [15] моделюють умовний розподіл $p(x \mid y)$ класів вихідної змінної з урахуванням їх особливостей і вивчають параметри моделі шляхом максимізації умовної правдоподібності на підставі $p(x \mid y)$.



а б

Рис. 2. Візуальне представлення: а – породжувального та б – розрізнювального підходів

Розрізнювальні класифікатори визначають параметр α дискримінантної функції $\lambda(a)$ через максимізацію умовної ймовірності $\hat{\alpha} = \arg\max_{\alpha} p(\underline{y}_{tr} \mid \underline{x}_{tr}, \alpha)$. Тобто, процедура оцінки мінімального емпіричного ризику може виступати як оцінка максимальної правдоподібності з функцією правдоподібності $p(\underline{y}_{tr} \mid \underline{x}_{tr}, \alpha)$. Відомо, що якщо використовується бінарна функція втрат, то коефіцієнт помилки класифікації є загальним ризиком.

Розрізнювальні моделі не пропонують чітких уявлень про зв'язки між особливостями і класами в наборі даних. Замість того, щоб використовувати ресурси для повного моделювання кожного класу, вони зосереджені на моделюванні кордону між класами. Прикладом розрізнювального класифікатора є SVM алгоритм, нейронні мережі, гаусові процеси та ін. Недоліком розрізнювального підходу є те, що розподіл даних при класифікації все ж таки необхідно брати до уваги.

Виходячи з вищевказаного, можна зробити висновок, що для класифікації досліджуваного набору даних, у відповідності до його характеристики, більш доцільним є використання алгоритмів породжувального або комбінованого підходів.

Мінімаксний підхід для класифікації даних. Lanckriet et al. [16] запропонували метод мінімізації максимальної ймовірності помилкової класифікації – Minimax Probability Machine (MPM).

Припустимо, що два випадкових n-мірних вектора y_1 і y_2 є два класи даних, що належать сімейству розподілів із заданими середніми і коваріаційними матрицями, що позначаються як $\{\overline{y_1}, \Sigma_{y_1}\}$, $\{\overline{y_2}, \Sigma_{y_2}\}$ відповідно до задачі класифікації по двом класам, де $y_1, y_2, \overline{y_1}, \overline{y_2} \mathbb{R} \in^n$ та $\Sigma_{y_1}, \Sigma_{y_2} \mathbb{R} \in^{n \times n}$. Клас y_1 представляє переважаючий клас, а клас y_2 відповідно менш переважаючий клас.

З надійною оцінкою $\{\overline{y_1}, \Sigma_{y_1}\}$, $\{\overline{y_2}, \Sigma_{y_2}\}$ для двох класів даних, мінімаксний підхід намагається визначити гіперплоскість $w^T z = b$ ($w \neq 0, z \in \mathbb{R}, b \in \mathbb{R}$, індекс T означає транспонування), яка розділяє дані на два класи з максимальною ймовірністю.

Передбачається, що ця гіперплоскість рішення мінімізує найгіршу (максимальну) ймовірність помилкової класифікації або частоту помилок для класифікації майбутніх точок даних. Мінімаксна оптимізація класифікації полягає в приведенні задачі лінійної класифікації до проблеми опуклою оптимізації, а точніше до вигляду конічного програмування другого порядку (SOCP) з глобальним оптимальним рішенням. Для її вирішення використовується ітеративний підхід найменших квадратів. Припущення про розподіл призводить до оптимізації, в результаті якої отримано рівняння Чернова. Таким чином, ми вирішуємо ту ж задачу оптимізації, але з монотонно зростаючою нижньою межею точності α. Тому гіперплоскість матиме вищу прогнозовану ймовірність правильної класифікації майбутніх даних.

Для більш ефективної роботи MPM також може бути розширений до вирішення нелінійної задачі з використанням параметричної редукції. Ефективність досягається шляхом препроцесорного кроку, в якому вхід алгоритму замінюється на менший вхід, що називається «ядром».

Класифікація з використанням мінімаксного підходу по даним перебігу вагітності для прогнозування виникнення патології у новонародженого включає вивчення всіх можливих варіантів значень показників. Проводиться оцінка кожного спостереження. Для того, щоб визначити, до якого класу приведе спостереження, необхідно зробити припущення, що для прогнозування стану новонародженого кожне подальше спостереження буде характеризувати перебіг вагітності з певним станом новонародженого найкращим чином. Алгоритм вибере найбільш характерні значення показників, що характеризують певний стан новонародженого, припустивши при цьому, що кожен наступний вибір буде відповідати найбільш характерним значенням показників і т.д.

Ефективність цього метода доведена в дослідженні [6] на синтетичних і реальних наборах даних.

При мінімаксному підході та при використанні більшості алгоритмів, модель показує прийнятні результати, коли кількість екземплярів кожного класу приблизно однаково. Коли число екземплярів одного класу набагато перевищує число екземплярів другого класу, виникають проблеми. Класифікація навчального набору з великим зсувом зазвичай видає моделі не схильні до перенавчання, які не можуть перелаштуватися до тестових даних і не в змозі врахувати їх важливі закономірності.

Проблемою використання мінімаксного підходу для незбалансованих даних є те, що в реальних випадках значення для двох класів не завжди одне і те ж саме, що означає, що нижня межа α для двох класів не обов'язково однакова. Ця проблема вирішується за допомогою різновида мінімаксного підхода для класифікації зсунених даних, який називається Biased Minimax Probability Machine (BMPM) [17]. Метою застосування цього підходу є підвищення точності переважаючого класу, а не загальної точності, наскільки це можливо, при збереженні точності менш переважаючого класу на прийнятному рівні. На відміну від традиційних збалансованих (упереджених) методів навчання, BMPM може кількісно включати в себе упередженість для одного класу і, отже, підкреслювати більш значущі класи. Дослідження [5] показали перспективність цього методу в незбалансованому навчанні та медичній діагностиці. Він підходить для діагностики

патологій новонародженого, тому що класи досліджуваних даних незбалансовані, випадків новонароджених з патологією зазвичай набагато менше ніж новонароджених з відсутністю патології.

Проведення класифікації з використанням алгоритмів розрізнювального, породжувального та мінімаксного підходів. Для визначення найкращого підходу до класифікації реальних медичних даних, проведена класифікація за допомогою класифікатора породжувального підходу з використанням байєсового алгоритму, класифікація за допомогою розрізнювального підходу з використанням алгоритму SVM та класифікація з використанням BMPM.

В якості алгоритму класифікації об'єктів для розрізнювального підходу використано алгоритм SVM. Цей алгоритм підходить для вирішення завдань бінарної класифікації, працює з різними типами даних і, також, з відсутніми значеннями. В якості алгоритма породжувального підходу до класифікації об’єктів, використано байєсовий алгоритм. Для мінімаксного підходу використовується BMPM [17]

Нижче приведені результати оцінки ефективності обраних методів, що випробувались на реальних даних перебігу вагітності. В контексті діагностики захворювань [7] частку істинно позитивних спостережень зазвичай називається чутливістю, а частку істинно негативних спостережень називається специфічністю. Мета класифікації станів новонароджених по даним перебігу вагітності - максимально висока чутливість при прийнятному значенні специфічності. Нижче наведені експериментальні результати з використанням SVM, байєсового алгоритмів та BMPM. Дані розділені випадковим чином на навчальний набір, що містить 80% даних, та тестовий набір, що містить відповідно 20% даних. За допомогою навчальних наборів даних отримані моделі. Проведена за їх допомогою класифікація тестових наборів даних, та виконана оцінка ефективності класифікації тестового набору даних.

Для перевірки якості моделі, по-перше, визначена кількість помилок класифікації першого (позитивний результат прийнятий за негативний) і другого роду, коли в результаті класифікації негативний результат прийнятий за позитивний для тестових наборів даних для кожного підходу. Результат представлений в табл.2.

З табл.2 виходить, що байєсовський алгоритм та SVM при проведенні класифікації для тестового набору показали стовідсоткові помилки першого роду для досліджуваних даних. Алгоритм BMPM показав найбільшу помилку другого роду. По результатах цих показників важко зробити висновки про ефективність моделей, тому далі представлені визначені частки істинно позитивних та істинно негативних спостережень для досліджуваних моделей.

Таблиця 2

Помилки першого та другого роду для тестового набору даних для порівнюваних підходів.

Метод	Помилки першого роду		Помилки другого роду	
	кількість	%	кількість	%
SVM	15	100	1	4.34
Байєс	15	100	0	0
BMPM	12	80	3	14.29

Таблиця 3

Частка істинно позитивних та істинно негативних спостережень

Метод	Істинно позитивні спостереження		Істинно негативні спостереження	
	кількість	%	кількість	%
SVM	0	0	22	95.66
Байєс	0	0	23	100
BMPM	3	20	20	85.71

По показникам істинно позитивних та істинно негативних спостережень, представлених в табл. 2, найкращі результати якості побудованої моделі показав алгоритм BMPM.

Чутливість та специфічність побудованих моделей, визначена на підставі показників, представлених в табл.2, 3. Отримані результати наведені в табл. 4.

Таблиця 4

Результати показників ефективності моделей

Метод	Специфічність	Чутливість
SVM	0	0.59
Байєс	0	0.60
BMPM	0.2	0.62

Оцінка специфічності для тестового набору даних показує найкраще прогнозування для спостережень з наявністю патології у новонародженого при використанні моделі BMPM. Оцінка чутливості показує, що найкраще прогнозування для спостережень без патології у новонародженого для дає модель BMPM.

Висновки. В роботі розглянуті підходи лінійної бінарної класифікації даних до прогнозування клінічних станів з урахуванням характеристик досліджуваних даних. В дослідженні використані алгоритми розрізнювального, породжувального та мінімаксного підходів для лінійної бінарної класифікації даних. Отриман результат оцінки ефективності тестування запропонованих моделей на реальних даних перебігу вагітності та стану новонародженого.

Отримані результати класифікації досліджуваних даних доводять наступне: 1) класифікація з використанням мінімаксного підходу є конкурентоздатнішою в порівнянні з існуючими розрізнювальних та породжувальних підходами; 2) мінімаксний підхід показав найменшу помилку першого роду та другий результат по помилкам другого роду для тестового набору даних; 3)

найкращу якість класифікації тестового набору даних на підставі оцінок специфічності та чутливості дає мінімаксна модель для незбалансованих даних BMPM. Отримані результати оцінки якості та ефективності класифікації свідчать про перевагу мінімаксного підходу до класифікації з використанням алгоритму BMPM.

Визначені параметри досліджуваних даних, урахування яких визначає вибір кращого підходу до лінійної бінарної класифікації даних. Ці параметри включають такі показники: тип даних, кількість вхідних змінних, бінарність або мультикласовість вихідної змінної, об`єм досліджуваних даних, зсуненість, незбалансованість даних, відсутні значення.

Підтверджені теоретичні припущення щодо якості мінімаксного підходу – з урахуванням властивостей даних при виборі підходу до лінійної бінарної класифікації даних алгоритм BMPM демонструє перевагу щодо алгоритмів інших підходів.

Продемонстровано та деталізовано підходи до лінійної бінарної класифікації зсунених або незбалансованих наборів даних. Оцінка моделей визначена на тестовому наборі реальних даних перебігу вагітності по наступним критеріям: помилки першого та другого роду, частка істино позитивних та істино негативних спостережень, специфічність, чутливість. По цим критеріям найкращі показники в порівняні з іншими методами, представниками яких обрані алгоритми SVM, байєсовський, BMPM. Найкращі результати якості та ефективності моделі показав метод з використанням алгоритма мінімаксного підходу для зсунених або незбалансованих даних - BMPM.

Урахування визначених в дослідженні властивостей даних дозволило обрати кращій підхід до лінійної бінарної класифікації при вирішенні конкретної прикладної задачі прогнозування стану новонародженого на підставі даних перебігу вагітності. Визначена краща стратегія при проведенні аналізу невеликих наборів з незбалансованими даними.

Література

1. Cateni S., Colla V., Vannucci M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems //Neurocomputing. – 2014. – Т. 135. – С. 32-41.
2. Li P., Chan K. L., Fang W. Hybrid kernel machine ensemble for imbalanced data sets //Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. – IEEE, 2006. – Т. 1. – С. 1108-1111.
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. – Т. 94. – С. 88-104.
4. Sanabila H. R., Kusuma I., Jatmiko W. Generative oversampling method (GenOMe) for imbalanced data on apnea detection using ECG data //Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on. – IEEE, 2016. – С. 572-579.
5. Huang K. et al. Biased Minimax Probability Machine for Medical Diagnosis //ISAIM. – 2004.
6. Gu B., Sun X., Sheng V. S. Structural minimax probability machine, IEEE Transactions on Neural Networks and Learning Systems. – 2017.
7. Huang K. Z., Yang H., Lyu M. R. Machine learning: modeling data locally and globally. – Springer Science & Business Media, 2008.
8. Huang K. Learning From Data locally and globally : дис. – Chinese University of Hong Kong, 2004.
9. Dang T. D., Nguyen H. N. Multi-class minimax probability machine //Knowledge and Systems Engineering, 2009. KSE'09. International Conference on. – IEEE, 2009. – С. 150-153.
10. Xue J. H., Titterington D. M. Comment on “On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes”, Neural processing letters. – 2008. – Т. 28. – №. 3. – С. 169-187.
11. Andrew Ng CS229 Generative Learning algorithms, Part IV of Lecture notes, режим доступу: http://cs229.stanford.edu/notes/cs229-notes2.pdf ,
12. Ng A. Y., Jordan M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Advances in neural information processing systems. – 2002. – С. 841-848.
13. Huang K. et al. Learning classifiers from imbalanced data based on biased minimax probability machine //Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. – IEEE, 2004. – Т. 2. – С. II-II.
14. Скарга-Бандурова І.С., Білобородова Т.О., Пошуковий аналіз даних для визначення релевантних факторів гіпоксичного ураження плода, Вісник НТУ "ХПІ". Збірник наукових праць. Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2016. – № 44 (1216). – 102-115 с.
15. Ng A. Y., Jordan M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Advances in neural information processing systems. – 2002. – С. 841-848.
16. Lanckriet G. et al. Minimax probability machine, Advances in neural information processing systems. – 2002. – С. 801-807.
17. Haiqin Yang, Kaizhu Huang, Irwin King, Michael R. Lyu and Laiwan Chan. Matlab Toolbox for Biased Minimax Probability Machine (BMPM-1.0), режим доступу: http://www.cse.cuhk.edu.hk/~miplab/mempm_toolbox/index.htm, 2004.

References

1. Cateni S., Colla V., Vannucci M. A method for resampling imbalanced datasets in binary classification tasks for real-world problems //Neurocomputing. – 2014. – Т. 135. – С. 32-41.
2. Li P., Chan K. L., Fang W. Hybrid kernel machine ensemble for imbalanced data sets //Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. – IEEE, 2006. – Т. 1. – С. 1108-1111.
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. – Т. 94. – С. 88-104.
4. Sanabila H. R., Kusuma I., Jatmiko W. Generative oversampling method (GenOMe) for imbalanced data on apnea detection using ECG data //Advanced Computer

Science and Information Systems (ICACSIS), 2016 International Conference on. – IEEE, 2016. – С. 572-579.
5. Huang K. et al. Biased Minimax Probability Machine for Medical Diagnosis //ISAIM. – 2004.
6. Gu B., Sun X., Sheng V. S. Structural minimax probability machine, IEEE Transactions on Neural Networks and Learning Systems. – 2017.
7. Huang K. Z., Yang H., Lyu M. R. Machine learning: modeling data locally and globally. – Springer Science & Business Media, 2008.
8. Huang K. Learning From Data locally and globally : дис. – Chinese University of Hong Kong, 2004.
9. Dang T. D., Nguyen H. N. Multi-class minimax probability machine //Knowledge and Systems Engineering, 2009. KSE'09. International Conference on. – IEEE, 2009. – С. 150-153.
10. Xue J. H., Titterington D. M. Comment on "On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes", Neural processing letters. – 2008. – Т. 28. – №. 3. – С. 169-187.
11. Andrew Ng CS229 Generative Learning algorithms, Part IV of Lecture notes, режим доступу: http://cs229.stanford.edu/notes/cs229-notes2.pdf ,
12. Ng A. Y., Jordan M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Advances in neural information processing systems. – 2002. – С. 841-848.
13. Huang K. et al. Learning classifiers from imbalanced data based on biased minimax probability machine //Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. – IEEE, 2004. – Т. 2. – С. II-II.
14. Skarga-Bandurova I.S., Biloborodova T.O., Poshukoviy analiz danyh dlya vyznachennya relevantnyh factoriv hipoxichnogo urajennya ploda, Visnyk NTU "HPI". Zbirnyk naukovyh prac. Seriya: Informatyka i modeluvannya. – Harkiv: NTU "HPI". – 2016. – № 44 (1216). – 102-115 p.
15. Ng A. Y., Jordan M. I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, Advances in neural information processing systems. – 2002. – С. 841-848**.**
16. Lanckriet G. et al. Minimax probability machine, Advances in neural information processing systems. – 2002. – С. 801-807.
17. Haiqin Yang, Kaizhu Huang, Irwin King, Michael R. Lyu and Laiwan Chan. Matlab Toolbox for Biased Minimax Probability Machine (BMPM-1.0), available at: http://www.cse.cuhk.edu.hk/~miplab/mempm_toolbox/index.htm, 2004.

Белобородова Т.А., Скарга-Бандурова И.С. Подходы к классификации несбалансированных и ассимметричных наборов данных

Рассмотрена проблема выбора подхода, алгоритма для бинарной классификации с учетом свойств входных данных. Проанализированы математические модели, выделенны свойства каждого из подходов. Проведена классификация с использованием алгоритмов, представляющих каждый из подходов. Выполнена оценка качества и эффективности моделей с использованием таких показателей, как ошибки первого и второго рода, доля истинно положительных и истинно отрицательных наблюдений, чувствительность и специфичность полученных моделей для тестового набора данных. С помощью Biased Minimax Probability Machine (BMPM), полученны характеристики классификации данных, учитывающих свойства данных. Подтверждено качество минимаксного подхода к классификации несбалансированных данных.

Ключевые слова: *линейная бинарная классификация, дискриминантный, генеративный, минимаксный подходы, несбалансированные данные*

Biloborodova T.O., Skarga-Bandurova I.S. Approaches for Classification of Imbalanced and Skewed Datasets

The problem of choosing data classification approach taking into account data features is considered. Mathematical models are analyzed, the properties of each approach are highlighted. The estimation of quality and efficiency of models with different indicators as the type 1 and type 2 errors, a share of truly positive and truly negative observations, sensitivity and specificity of the received models for a test data set is carried out. With the Biased Minimax Probability Machine, the data classification characteristics factored in data properties are obtained. The quality of the minimax approach to the classification of unbalanced data has been confirmed.

Keywords: *linear binary classification, generative and discriminative classifiers, minimax, biased data*

Білобородова Т.О. – аспірант, старший викладач кафедри комп'ютерної інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: beloborodova.t@gmail.com
Скарга-Бандурова І.С. – доктор технічних наук, доцент, професор кафедри комп'ютерної інженерії Східноукраїнського національного університету імені Володимира Даля, e-mail: skarga_bandurova@ukr.net

Рецензент: д.т.н., проф. ***Смолій В.М.***

Стаття подана 01.09.2017