

ОЦІНЮВАННЯ ЯКОСТІ КЛАСИФІКАЦІЇ НА ОСНОВІ БІНОМІАЛЬНОЇ МОДЕЛІ

С. В. Свириденко¹, С. А. Смирнов¹

¹ Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»,
Фізико-технічний інститут

Анотація

У роботі розглянуті та проаналізовані критерії якості оцінювання класифікації об'єктів екзаменаційної вибірки. Запропонована нова методика оцінки якості класифікації. Наведені алгоритми дозволяють швидко і з високою точністю обчислювати ймовірності моделі.

Ключові слова: класифікація, оцінка якості, біноміальна модель, об'єм вибірки

Вступ

На сьогоднішній день існує декілька підходів оцінювання якості класифікації, зокрема методи на основі нормальної та поліноміальної моделі. Класифікація проводиться на екзаменаційній вибірці, яка повинна містити об'єкти-представники із всіх заданих класів. При цьому виникає проблема визначення статистично коректного об'єму такої вибірки. Проте наведені методики мають ряд недоліків, що призводить до критичних похибок при ймовірнісних обчисленнях.

1. Методика оцінювання якості класифікації об'єктів

Процедура оцінювання на основі нормальної моделі базується на апроксимації біноміального розподілу гаусовим. При оцінюванні якості класифікації вважається відомою величина – ймовірність правильної класифікації, яку повідомляє розробник методики класифікації. Завдання допустимої ймовірності помилки перевірки якості методики класифікації дозволяє побудувати певний довірчий інтервал за таблицею квантилей стандартного нормального розподілу. Методика вважається якісною, якщо відповідна вибірка характеристика потрапляє в нього. У даному методі враховується інформація про кількість класів та представників кожного класу, але методика використовується для кожного класу окремо. Проте нормальна модель не дає можливості визначити достатній об'єм вибірки з урахуванням вимоги до точності класифікації на рівні класу [1]. Від цього недоліку міг би бути вільний підхід на основі поліноміальної моделі. Специфіка підходу полягає у наступному: в якості основної ймовірнісної моделі використовується поліноміальна схема Бернуллі. Але далі для розрахунків знову використовується нормальна апроксимація. Необхідно зауважити, що поліноміальна модель і критерій Тортора [2] ефективні

в умовах достатньо великих вибірок, коли популяція об'єктів якогось класу домінує над іншими, наближаючись до половини всього об'єму вибірки. Якщо ж всі класи представлені у вибірці приблизно однаково, то розрахунки можуть привести до некоректних результатів.

Тепер, зважаючи на недоліки вище наведених моделей визначення статистично коректного об'єму, запропонуємо нову методику оцінювання на основі суто біноміальної моделі та нових високоточних обчислювальних алгоритмів для неї, повністю виключаючи похибки апроксимації. В основі всіх моделей покладено ймовірність правильної класифікації P_0 та ймовірність помилки 1-го роду α . Нехай відома екзаменаційна вибірка, що складається із N об'єктів (без обмежень загальності розглядається два класи), а по результатах класифікації отримуємо s неправильно розпізнаних об'єктів. Позначимо: q – заявлена розробником ймовірність неправильної класифікації, α – допустима ймовірність помилки перевірки якості методики класифікації. Функція розподілу для дискретної випадкової величини s визначається за біноміальним законом таким чином:

$$B(N, s', q) = Pr(s = s' | q) = \frac{N!}{(N - s')! \cdot s'!} q^{s'} \cdot (1 - q)^{N - s'}. \quad (1)$$

Для перевірки узгодженості статистичних даних, отриманих за вказаною вибіркою розміром N , з величиною допустимої ймовірності помилки α використовуємо нерівність наступного виду:

$$B(N, s', q) \geq \alpha, \quad (2)$$

яка теоретично визначає довірчий інтервал для вибірових оцінок $s \in [s'; s'']$

Будемо вважати розмір вибірки N статистично коректним, якщо вказаний довірчий інтервал існує, тобто нерівність (2) має рішення. Тому, щоб не перебирати всі можливі значення, достатньо знайти

максимальне значення функції $B(N, s, q)$, і тоді перевірити лише одну нерівність:

$$B(N, s'_{\max}, q) \geq \alpha. \quad (3)$$

Якщо дана нерівність виконується, то об'єм N – статистично коректний, і отриманим оцінкам можна довіряти.

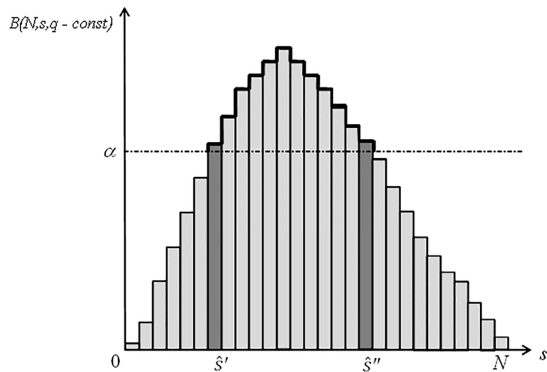


Рис. 1. Максимум функції $B(N, s, q)$ знаходиться вище рівня α

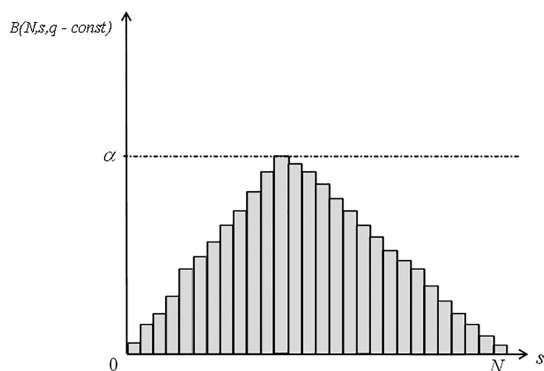


Рис. 2. Максимум функції $B(N, s, q)$ знаходиться на рівні α

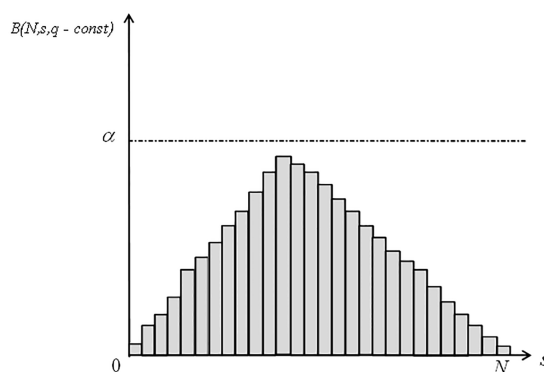


Рис. 3. Максимум функції $B(N, s, q)$ знаходиться нижче рівня α

На рис. 1 і 2 розглянуто випадки, коли N – статистично коректний, оскільки максимальне значення функції не нижче заданого рівня α , на відміну від рис. 3, де розмір вибірки не є статистично коректним, а отже отриманим оцінкам не можна довіряти.

Таким чином отримується інформація про статистичну коректність об'єму вибраної вибірки об'єктів для подальшого дослідження процедури класифікації. Далі метод буде давати підтвердження/спростування правильності вибраної процедури класифікації. Отже, методика складається з таких етапів:

- формування досліджуваної вибірки об'єктів, в якій повинні бути представники всіх класів з числа заданих, які потрібно розпізнавати;
- уточнення необхідних параметрів: q – заявлена розробником ймовірність неправильної класифікації; α – допустима ймовірність помилки перевірки якості методики класифікації задана замовником;
- класифікація об'єктів вибірки, за допомогою досліджуваної методики;
- знаходження кількості неправильно розпізнаних об'єктів, s' ;
- знаходження загальної точності класифікації, $P_0 = 1 - q$;
- побудова довірчого інтервалу для оцінки $s' \in [s'; s'']$ за α ;
- пошук такого значення s'_{\max} , при якому досягається максимум функції $B(N, s, q)$;
- встановлення статистичної коректності об'єму вибірки за критерієм, визначеним формулою (3) для заданого рівня помилки 1-го роду α ;
- перевірка потрапляння вибіркової оцінки в довірчий інтервал $s' \in [s'; s'']$, висновки щодо коректності методики.

Для реалізації отриманої методики потрібно розробити декілька алгоритмів для швидкого і високо-точного обчислення оцінок та довірчих інтервалів для біноміального розподілу. У стандартній схемі комп'ютерних обчислень за виразом (1) для великих значень N і малих значень q відбувається бітове переповнення стандартної форми представлення дійсної величини в пам'яті, що призводить до грубих помилок обчислення. Для запобігання цьому запропоновано деякі математичні перетворення, які дозволяють проводити обчислення моделі з великою точністю.

2. Допоміжні обчислювальні алгоритми

Розглянемо алгоритм розрахунку ймовірності за біноміальним розподілом, який має два варіанта:

- 1) обчислення $B(N, s, q = \frac{s}{N})$, коли на вхід подається два параметра (s, N) ;
- 2) обчислення $B(N, s, q)$, коли на вхід подається три параметра (s, N, q) .

Наведемо математичні перетворення для обчислення біноміального розподілу з контрольовано зростаючою похибкою. Розглянемо перший випадок, коли задані параметри (s, N) , тоді:

$$B(s, q = \frac{s}{N}, N) = \frac{N!}{(N-s)! \cdot s!} \left(\frac{s}{N}\right)^s \left(1 - \frac{s}{N}\right)^{N-s}.$$

Після простих перетворень отримаємо:

$$B(s, q = \frac{s}{N}, N) = \frac{N!}{N^s} \cdot \frac{s^s}{s!} \cdot \frac{(N-s)^{N-s}}{(N-s)!} = \\ = \prod_{i=1}^N \frac{i}{N} \prod_{j=1}^s \frac{s}{j} \prod_{k=1}^{N-s} \frac{N-s}{k}.$$

Перепишемо формулу так, щоб добуток був представлений через множники величиною порядку одиниці. Для цього розіб'ємо добуток на 2 частини, кожна з яких складається з N елементів, тоді функція прийматиме вигляд:

$$B(s, q = \frac{s}{N}, N) = \underbrace{\left[\frac{1}{N} \cdots \frac{N}{N} \right]}_N \times \\ \times \underbrace{\left[\left(\frac{s}{1} \cdots \frac{s}{s} \right) \left(\frac{N-s}{1} \cdots \frac{N-s}{N-s} \right) \right]}_N. \quad (4)$$

Впорядкуємо обидві частини формули (4): першу в порядку зростання, другу – спадання. Утворимо пари з елементів, які будуть взаємно компенсуючими, для цього необхідно більший елемент з однієї частини помножити на менший з іншої згідно з їх порядком. Тим самим створюємо множники не сильно відмінні від одиниці. Візьмемо натуральний логарифм від правої і лівої частини рівняння (яка складається з добутків пар), відповідно множення перетвориться в додавання. За рахунок цього можна значно знизити ресурси обчислювальної системи та зменшити накопичення помилки з кожною ітерацією, тому що при множенні похибка збільшується мультиплікативно, на відміну від додавання (адитивно).

Розглянемо другий випадок, коли потрібно розрахувати біноміальну ймовірність для заданих трьох параметрів (s, N, q) . Розпишемо добуток і приведемо формулу (1) аналогічним способом до такого виду:

$$B(s, q = \frac{s}{N}, N) = \underbrace{[1 \cdots N]}_N \times \\ \times \underbrace{\left[\left(\frac{q}{1} \cdots \frac{q}{s} \right) \left(\frac{1-q}{1} \cdots \frac{1-q}{1-q} \right) \right]}_N \quad (5)$$

Далі впорядковуємо (5) і згрупуємо множники відповідно до попередньої процедури. Візьмемо від правої та лівої частини (утвореною добутками пар) натуральний логарифм і отримаємо вираз, що значно спрощує ресурси і підвищує точність обчислення. Перевірка точності обчислення алгоритмів здійснювалася за допомогою спеціальної бібліотеки GMP та мови програмування C++. Потрібно додати, що запропоновані спрощення формул показують хороші результати для невеликих вибірок об'єктів, а коли потрібно обчислити для достатньо великих вибірок застосовуємо спеціальну арифметику в програмі, яка розроблена в бібліотеці GMP.

Висновки

- 1) Розроблено математичну модель та новий метод перевірки якості класифікації, який оснований на біноміальній моделі. Запропонована методика оцінки якості класифікації, яка підтверджує чи спростовує достовірність результатів класифікації та оцінок їх якості, при достатньому (статистично достовірному) розмірі екзменаційної вибірки. Отримані оцінки дозволяють оптимізувати витрати на формування вибірки.
- 2) Розроблені та протестовані нові ефективні алгоритми високоточного комп'ютерного обчислення біноміального розподілу, які використовуються при виконанні процедури оцінювання якості класифікації. Розроблені алгоритми дозволяють швидко і з високою точністю обчислювати параметри моделі.

Перелік використаних джерел

1. А. Попов М. Методология оценки точности классификации объектов на космических изображениях. — Проблемы управления и информатики, 2007. — С. 97–103.
2. R. Tortora. A note on sample size estimation for multinomial populations. — American Statistical, 1978. — P. 100–102.