УДК 004.8:001.895

Єрмаков В.А., здобувач,
Косенко С.І., к.т.н., доцент
Національний університет «Одеська політехніка»

ОГЛЯД СУЧАСНИХ МЕТОДІВ СТАТИСТИЧНОГО АНАЛІЗУ МЕДИЧНИХ ДАНИХ

У контексті розвитку цифрової медицини та біоінформатики значення статистичного аналізу медичних даних постійно зростає. Результати медичних досліджень, електронні медичні записи, дані проспективних спостережень, а також великі обсяги геномних та протеомних даних створюють значні обсяги інформації, що вимагає застосування ефективних методів опрацювання та інтерпретації [1]. Статистичні методи відіграють ключову роль у виділенні значущих закономірностей, оцінюванні взаємозв'язків між факторами ризику та клінічними результатами, а також у побудові прогнозних моделей. Актуальність огляду сучасних підходів до статистичного аналізу медичних даних обумовлена не лише зростанням обсягів даних, але й потребою у виборі адекватних методів для підвищення точності висновків та клінічних рекомендацій [2].

Одним із базових підходів до обробки медичних даних є регресійний аналіз. Лінійна регресія традиційно застосовується для оцінювання зв'язку між кількісними предикторами та безперервними результатами. При цьому важливо перевіряти відповідність даних припущенням моделі щодо нормальності розподілу залишків і гомоскедастичності.

Логістична регресія є стандартним методом для аналізу бінарних результатів. Вона дозволяє моделювати ймовірність настання певної події залежно від множини факторів ризику. Однією з переваг цього методу є інтерпретованість отриманих коефіцієнтів у вигляді відношення шансів. Однак для великих наборів предикторів з високою мультиколінеарністю застосовують регуляризацію, зокрема LASSO та Ridge-регресії, що дозволяють зменшувати надмірну варіацію оцінок та підвищувати стабільність моделей.

У випадку високовимірних даних, наприклад даних генетичних асоціацій, застосовують методи багатовимірного аналізу, зокрема аналіз головних компонент, факторний аналіз та множинний кластерний аналіз. Ці підходи дозволяють виявити латентні структури в даних, зменшити їхню розмірність і здійснювати візуалізацію складних взаємозв'язків між ознаками [1]. Аналіз головних компонент

особливо корисний при роботі з даними високого виміру, де пряма інтерпретація частинної кореляційної структури є складною.

Методи машинного навчання, такі як випадкові ліси, метод опорних векторів та градієнтний бустинг, набули поширення у медичному аналізі даних. У клінічних задачах вони ефективні для класифікації станів пацієнтів, прогнозування ускладнень та оцінювання ризику. Важливою перевагою є їх здатність автоматично обробляти великі набори ознак, однак інтерпретованість таких моделей часто нижча порівняно з класичними статистичними підходами.

Для оцінювання якості класифікаційних моделей у медицині використовують різні метрики, включно з площею під кривою ROC, показниками чутливості та специфічності, що дозволяє повноцінно оцінювати прогностичну здатність моделі на тестових даних [2].

Методи багатовимірного узагальнення параметрів, наприклад байєсівські підходи, активно використовуються в тих випадках, коли дані мають складну структуру або коли необхідно включити апріорні знання експертів у модель. Серед непараметричних підходів варто зазначити методи, що не роблять жорстких припущень про форму розподілу даних [1]. Наприклад, методи ядерної оцінки щільності та сплайн-регресії широко застосовуються для виявлення нелінійних трендів у наборах медичних даних. Непараметричні тести, такі як критерії Манна–Уїтні та Краскела–Уолліса, використовуються для порівняння розподілів між групами у випадках, коли дані не відповідають припущенням нормальності [2].

В результаті проведених досліджень слід зазначити, що лінійні та логістичні моделі залишаються основою для багатьох медичних досліджень завдяки своїй простоті та інтерпретованості, тоді як сучасні алгоритми машинного навчання дозволяють працювати зі складнішими даними та виявляти нелінійні залежності. Методи обробки пропущених даних, непараметричні підходи та часові моделі доповнюють арсенал інструментів, що дозволяють ефективно аналізувати великі та гетерогенні набори медичних даних.

Список використаних джерел:

1. Harrell F.E. Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis. Statistics in Medicine. 2015. Vol. 34, No 7. P. 1151–1152.

2. Obermeyer Z., Emanuel E.J. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. New England Journal of Medicine. 2016. Vol. 375, No 13. P. 1216–1219.