

ПІДХІД ДЛЯ ПОКРАЩЕННЯ ТОЧНОСТІ ПРЕДИКТИВНИХ МОДЕЛЕЙ ЗА РАХУНОК ІНЖЕНЕРІЇ ОЗНАК З ОЦІНКОЮ НЕВИЗНАЧЕНОСТІ

Вінницький національний технічний університет, Україна

Анотація

У роботі запропоновано підхід для покращення точності передбачення давності синців з використанням оцінки невизначеності. Було застосовано байєсовий підхід для оцінки невизначеності передбачень і ці оцінки було інтегровано в модель градієнтного бустингу як додаткові ознаки. Запропонований підхід дозволив покращити модель градієнтного бустингу на 8% порівняно з моделлю, побудованою виключно на оригінальних даних. Робота демонструє, що оцінка невизначеності дозволяє не лише краще розуміти отримані результати, але й має потенціал для покращення точності предиктивних моделей.

Ключові слова: машинне навчання, аналіз даних, медичні дані, оцінка невизначеності, градієнтний бустинг.

Abstract

The paper proposes an approach to improve the accuracy of bruise age prediction using uncertainty estimation. A Bayesian approach was applied to estimate the uncertainty of the predictions, and these estimates were integrated into the gradient boosting model as additional features. The proposed approach allowed us to improve the gradient boosting model by 8% compared to the model built solely on the original data. The work demonstrates that uncertainty estimation not only allows for a better understanding of the results, but also has the potential to improve the accuracy of predictive models.

Keywords: machine learning, data analysis, medical data, uncertainty estimation, gradient boosting.

Вступ

Зі збільшенням популярності використання методів машинного навчання, параметрів та обсягів навчальних даних для моделей, зростає актуальність потреби у інтерпретованій прозорості та оцінки невизначеностей передбачень моделей. Найбільш критичною ця потреба стосується галузей де помилки моделей мають за собою критичні наслідки, як приклад - медицина [1]. Традиційно оцінка невизначеності використовується для інформування лікарів про межі надійності алгоритмічних висновків (наприклад, у діагностиці раку за допомогою даних МРТ [2]).

Медичні дані, що використовуються для навчання моделі часто можуть бути фрагментарні (наприклад, електронні медичні картки з пропущеними значеннями [3]) або ж мати присутність шуму (зчитування біомаркерів із низькою точністю чи помилковою класифікацією даних), в таких випадках розуміння невизначеності моделі може бути індикатором наявності проблем у даних або архітектурі.

У даній роботі запропонований підхід використання оцінки невизначеності не лише для інтерпретації, але й для покращення якості передбачення моделі градієнтного бустингу методами інженерії ознак.

Метою роботи є покращення якості передбачення давності синців за рахунок використання оцінки невизначеності, що отримана за допомогою байєсового підходу з подальшим використанням в моделі бустингу методами інженерії ознак.

Результати дослідження

Для роботи було використано набір даних з [4], який був наданий авторами публікації. В наборі даних описується 2290 синців за такими характеристиками, як вік, стать, розташування, площа, тип кольору, набряк та давність.

Байєсовий підхід до машинного навчання у якості передбачення надає розподіл можливих значень, що дозволяє оцінити невизначеність, як дисперсію цього розподілу [5]. Тому першим кроком дослідження є отримання зазначених оцінок. Спочатку формується апіорний розподіл для невідомого параметра форми розподілу та вектор невідомих коефіцієнтів ознак із використанням нормального розподілу, що відображає початкову невизначеність. Далі цільова змінна (давність синця) моделюється

за допомогою розподілу Вейбулла, де параметри цього розподілу залежать від вхідних ознак через лінійну комбінацію коефіцієнтів. Оцінювання параметрів здійснюється за допомогою методів Монте-Карло марковських ланцюгів, що реалізовані у бібліотеці PyMC [6]. Це дозволяє отримати повний апостеріорний розподіл передбачень для кожного спостереження. З цього розподілу обчислюються статистичні характеристики, такі як середнє значення (mean) та середньоквадратичне відхилення (std).

Наступний етап включав в себе навчання предиктивних моделей на оригінальному наборі даних та підбір їх оптимальних параметрів оцінюючи якість моделей за кросс-валідацією та пошуком по сітці. Найкращі результати продемонструвала модель градієнтного бустингу. Наступний етап роботи - додання отриманих ознак mean та std, обрахованих на попередньому етапі, та побудова і навчання нової моделі градієнтного бустингу на доповнених даних та знайдених оптимальних параметрах. В результаті впровадження оцінки невизначеності у запропоновану модель було зафіксовано покращення якості передбачення за метрикою коефіцієнту детермінації на 8.43% відносно результатів першої моделі.

В табл. 1 наведені обраховані значення важливості ознак для двох моделей. За отриманими результатами, найбільша різниця полягає в тому, що у моделі з оцінкою невизначеності важливими стали ознаки середнього значення та середньоквадратичного відхилення, отримані з першого етапу дослідження, тоді як в моделі на оригінальних даних найважливішим предиктором була ознака кольору. Таким чином була розроблена комбінована система з обрахунку невизначеності та моделі градієнтного бустингу, що надає можливість не лише для пояснення передбачення та оцінки впевненості в результаті, а й для покращення моделі бустингу, що може використовуватись у тих випадках, коли пояснення рішення моделі не є необхідним.

Таблиця 1 - Важливість ознак в моделі градієнтного бустингу.

Модель	age	sex	localization	square	color	edema	std	mean
GBR оригінальний	0.306	0.017	0.026	0.069	0.581	0.002	-	-
GBR з невизначеністю	0.271	0.013	0.02	0.05	0.162	0.002	0.102	0.379

Висновки

У роботі було запропоновано підхід для покращення якості передбачення моделі з використанням оцінки невизначеності. Було застосовано байесовий підхід та бібліотеку PyMC для обрахунку оцінки невизначеності передбачень та створені ознаки mean та std для розширення набору даних. В результаті модель градієнтного бустингу навчена на розширених даних показала приріст в точності передбачення на 8.43% в порівнянні з моделлю що навчалась лише на оригінальних даних.

Проведене дослідження показує що оцінка невизначеності є не лише важливим етапом для інтерпретації отриманих результатів але й мають потенціал для покращення точності предиктивних моделей. Подальші роботи можуть детальніше дослідити позитивний зв'язок між методами оцінки невизначеності та їх використання для покращення якості передбачень моделей у інших галузях.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Claudio F. A Bayesian approach to predictive uncertainty in chemotherapy patients at risk of acute care utilization / Claudio Fanconi, Anne de Hond, Dylan Peterson, Angelo Capodici, Tina Hernandez-Boussard// EBioMedicine. – 2023. – Vol. 92. – P. 104632.
2. Cindy X. Computational analysis of variability and uncertainty in the clinical reference on magnetic resonance imaging radiomics: modelling and performance / Cindy Xue, Jing Yuan, Gladys G. Lo [et al.] // Visual computing for industry, biomedicine, and art. – 2024. – Vol. 7, no. 1.
3. Emily J. G. Mining for equitable health: assessing the impact of missing data in electronic health records / Emily J. Getzen, Pallavi V. Kulkarni, Danielle L. Mowery [et al.] // Journal of biomedical informatics. – 2023. – P. 104269.
4. Mokaniuk O. Complex assessment of the relevance of criteria of the forensic determination of the recentness of bruises and bruises with abrasion formation for the development of expert diagnostic programm / Olexandr Mokaniuk, Hanna Bondarchuk, Roman Kartelian // Congres internațional al medicilor legiști din Republica Moldova. – 2021. – Ediția a III-a. – P. 34–39.

5. Matthias B. Accounting for uncertainty: an application of Bayesian methods to accruals models / Matthias Breuer, Harm H. Schütt // Review of accounting studies. – 2021.
6. Oriol A. PyMC: a modern, and comprehensive probabilistic programming framework in Python [Electronic resource] / Oriol Abril-Pla, Virgile Andreani, Colin Carroll [et al.] // PeerJ computer science. – 2023. – Vol. 9. – P. e1516.

Гладіголов Сергій Сергійович — аспірант кафедри системного аналізу та інформаційних технологій, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: hladiholov.s@gmail.com

Козачко Олексій Миколайович — к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: lekoz80@gmail.com

Hladiholov Serhii — postgraduate student of Department of Systems Analysis and Information Technologies, Faculty of Intelligent Information Technologies and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: hladiholov.s@gmail.com

Kozachko Oleksii — Ph.D., associate professor of the System Analysis and Information Technologies Department, Vinnytsia National Technical University, Vinnytsia, e-mail: lekoz80@gmail.com