Секция «Теория вероятностей и математическая статистика»

Применение машинного обучения в эпидемиологических исследованиях

Научный руководитель – Яровая Елена Борисовна

Перевердиева К.Г.[1], *Куценко В.А.*[2]

1 - Московский государственный университет имени М.В.Ломоносова, Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия, *E-mail: pereverdieva@gmail.com*; 2 - Московский государственный университет имени М.В.Ломоносова, Механико-математический факультет, Кафедра теории вероятностей, Москва, Россия, *E-mail: vlakutsenko@ya.ru*

Методы машинного обучения широко используются для решения различных задач. В последние годы появились работы, в которых сравниваются подходы машинного обучения со статистическими методами, отражающие преимущество первых [1-4]. Цель данной работы определить, может ли применение алгоритмов машинного обучения улучшить качество прогноза наличия социально значимых заболеваний по сравнению со стандартными регрессионными моделями в эпидемиологических исследованиях. В работе используются данные исследования «Эпидемиология Сердечно-Сосудистых заболЕваний в регионах Российской Федерации» (ЭССЕ-РФ) [5]. На основе ЭССЕ-РФ было опубликовано более ста работ, но ни в одной из них не рассматривалось применение методов машинного обучения.

Нами были применены различные подходы для предсказания артериальной гипертонии (бинарная переменная) на данных из ЭССЕ-РФ (N = 13912). Клинические, демографические и социальные факторы риска для рассматриваемых моделей были отобраны в соответствии с [6,7]. Проводилось сравнение статистических методов предсказания, таких как логистическая регрессия без регуляризации с непрерывными и бинарными переменными, обобщенные аддитивные модели, и методов машинного обучения: lasso- и ridge-регрессии, случайный лес. Для сравнения качества моделей использовался ROC-анализ и оценивалась площадь под ROC-кривой (AUC). Данные были разделены на обучающую и тестовую выборку в отношении 70% и 30%, соответственно. Чтобы результаты не были искажены особенностями этого конкретного разбиения, AUC была вычислена для 1000 аналогичных испытаний каждого алгоритма, отличающихся лишь разбиением на обучающую и тестовую выборку.

На основе выборки из ЭССЕ-РФ были получены следующие результаты: логистическая регрессия с непрерывными переменными без регуляризации показала значимо более высокое качество предсказания на тестовой выборке ($p < 0.001$, AUC = 82.13% ± 0.52%), чем lasso- и ridge-регрессии (AUC = 81.74% ± 0.54%), регрессия с бинарными переменными без регуляризации (AUC = 80.98% ± 0.53%) и случайный лес (AUC = 81.54% ± 0.53%). Обобщенные аддитивные модели показали качество еще выше ($p = 0.003$, AUC = 82.19% ± 0.51%), чем регрессия с непрерывными переменными, однако увеличили уровень переобучения в два раза. Для объяснения полученных результатов мы использовали методы, не зависящие от структуры предсказывающего алгоритма, такие как важность признаков [8] и интеракции Фридмана [9]. В частности, важность признаков случайного леса оказалась более согласованной с результатами исследований [6,7], чем важность признаков логистической регрессии. Используя полученную информацию, мы модифицировали модель логистической регрессии. Построенная модель показала качество выше ($p < 0.001$), чем каждая из рассмотренных ранее регрессий, случайный лес и обобщенные аддитивные модели.

Таким образом, в эпидемиологии неинфекционных заболеваний целесообразно придерживаться стандартных статистических подходов. Однако более сложные методы могут

предоставить дополнительную информацию, которая позволяет модифицировать модель и значимо улучшить качество предсказания.

Данные были предоставлены отделом эпидемиологии хронических неинфекционных заболеваний ФГБУ НМИЦ ТПМ Минздрава РФ.
Работа поддерживается РФФИ, грантом № 20-01-00487.

Источники и литература

1) Raphael Couronné, Philipp Probst, Anne-Laure Boulesteix. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19 Springer Science and Business Media LLC, 2018.
2) Hyung-Chul Lee, Hyun-Kyu Yoon, Karam Nam, Youn Cho, Tae Kim, Won Kim, Jae-Hyon Bahk. Derivation and Validation of Machine Learning Approaches to Predict Acute Kidney Injury after Cardiac Surgery. Journal of Clinical Medicine 7, 322 MDPI AG, 2018.
3) David Muchlinski, David Siroky, Jingrui He, Matthew Kocher. Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data. Political Analysis 24, 87–103 Cambridge University Press (CUP), 2016.
4) Haoyuan Hong, Paraskevas Tsangaratos, Ioanna Ilia, Wei Chen, Chong Xu. Comparing the Performance of a Logistic Regression and a Random Forest Model in Landslide Susceptibility Assessments. the Case of Wuyaun Area China. 1043–1050 In Advancing Culture of Living with Landslides. Springer International Publishing, 2017.
5) Бойцов С.А., Чазов Е.И., Шляхто Е.В., Шальнова С.А., и др. Эпидемиология сердечно-сосудистых заболеваний в различных регионах России (ЭССЕ-РФ). Обоснование и дизайн исследования. Профилактическая медицина 6, 25–34 (2013).
6) Paul K. Whelton, Robert M. Carey, Wilbert S. Aronow, Donald E. Casey, Karen J. Collins, Cheryl Dennison Himmelfarb, Sondra M. DePalma, Samuel Gidding, Kenneth A. Jamerson, Daniel W. Jones, Eric J. MacLaughlin, Paul Muntner, Bruce Ovbiagele, Sidney C. Smith, Crystal C. Spencer, Randall S. Stafford, Sandra J. Taler, Randal J. Thomas, Kim A. Williams, Jeff D. Williamson, Jackson T. Wright. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension 71, 1269–1324 Ovid Technologies (Wolters Kluwer Health), 2018.
7) Bryan Williams, Giuseppe Mancia, Wilko Spiering, Enrico Agabiti Rosei, Michel Azizi, Michel Burnier, Denis L Clement, Antonio Coca, Giovanni De Simone, Anna Dominiczak, others. 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). European heart journal 39, 3021–3104 Oxford University Press, 2018.
8) Marvin N Wright, Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409 (2015).
9) Jerome H Friedman, Bogdan E Popescu, others. Predictive learning via rule ensembles. Annals of Applied Statistics 2, 916–954 Institute of Mathematical Statistics, 2008.