Применение модели XGBoost для обработки данных о пациентах

М.А. Зайнуллина
Казанский национальный исследовательский технический университет им А. Н. Туполева - КАИ
Казань, Россия
madinazay@mail.ru

В.В. Мокшин
Казанский национальный исследовательский технический университет им А. Н. Туполева - КАИ
Казань, Россия
vladimir.mokshin@mail.ru

***Аннотация*— В данной статье были рассмотрены способы нахождения определяющих признаков, рассмотрен метод прогнозирования XGBoost. Для выделения значимых признаков наличия у пациента диабета использовался корреляционный анализ. Была проведена балансировка данных SMOTE для получения более точных результатов прогноза. Метод XGBoost показал хорошие результаты для данной задачи. Разработанный алгоритм позволяет предсказывать наличие диабета у пациента с высокой точностью.**

Ключевые слова— машинное обучение, корреляционный анализ, XGBoost, прогноз

1. Введение

На сегодняшний день одним из наиболее распространенных заболеваний после онкологии и сердечно-сосудистой патологии является диабет, который нередко приводит к смерти. Помимо прочего, диабет может способствовать развитию других проблем со здоровьем. Так, по данным всемирной организации здравоохранения, диабет является одной из основных причин ампутации нижних конечностей, почечной недостаточности, слепоты, сердечных приступов и инсульта [1]. Например, у взрослых людей, у которых выявлен диабет, в 2-3 раза выше риск развития инфаркта и инсульта [2].

Очень важно вовремя выявить болезнь и начать лечение. Для своевременного обнаружения возможно использование различных методик, одна из них – использование методов прогнозирования. Благодаря прогнозированию появляется возможность поспособствовать уменьшению негативных последствий диабета вследствие своевременного обнаружения и последующего лечения.

Для прогнозирования может подойти машинное обучение. Машинное обучение (МО) — это область искусственного интеллекта, которая систематически применяет алгоритмы, чтобы синтезировать основные отношения между данными и информацией [3].

Целью исследования является изучение методов МО для прогнозирования риска развития диабета среди пациентов.

2. Сбор данных

Для построения модели в качестве набора данных использовались данные с сайта Kaggle. Ссылка на исходный датасет с данными о пациентах: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Набор данных получен из Национального института диабета, болезней органов пищеварения и почек. Все пациенты в данном датасете — женщины не моложе 21 года индейского происхождения пима.

В качестве входных были взяты следующие параметры: количество беременностей, концентрация глюкозы в плазме через 2 часа в пероральном тесте на толерантность к глюкозе, диастолическое артериальное давление (мм рт. ст.), толщина кожной складки трицепса (мм), 2-часовой инсулин в сыворотке (мЕд/мл), индекс массы тела (далее ИМТ), возраст пациента и функция заболеваемости диабетом в зависимости от родословной (это функция, которая указывает на вероятность диабета на основании семейного анамнеза).

3. Обработка данных

Перед прогнозированием сначала данные стандартизируются и нормализуются. Нормализация выполняется для объектов, данные которых отображают нормальное распределение, а стандартизация выполняется для остальных объектов, где их значения огромны или очень малы по сравнению с другими объектами.

Существуют хорошо известные методы отбора значимых факторов, например, итерационные методы прямого и обратного отбора и корреляционные методы [4]. Для отбора был выбран корреляционный анализ. Чем выше коэффициент корреляции, тем более фактор значим. Исходя из анализа из рассмотрения для получения более точных прогнозов удаляются следующие параметры: толщина кожи и давление, т.к. имеют значения коэффициента Пирсона близкие к нулю.

При классификации в условиях несбалансированности классов может быть использована балансировка классов (SMOTE), предполагающая дублирование наблюдений класса, представителей которого в наборе меньше остальных, для того чтобы не "игнорировать" класс меньшинства, которые являются наиболее важными в прогнозировании, и, как следствие, иметь более высокую производительность [5].

4. Применение модели

Для прогнозирования была использована модель МО XGBoost. XGBoost – это дерево решений, основанное на градиентном подъеме, и представляет собой действенную системную реализацию метода бустинга. Алгоритм модели XGBoost выглядит следующим образом: целевая функция разлагается по Тейлору второго порядка, а задача оптимизации целевой функции преобразуется в решение наименьшего значения квадратичной функции [6]. Целевая функция модели:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t), \qquad (1)$$

где l - функция потерь, $y_i, \hat{y}_i^t$— значение i-го элемента обучающей выборки и сумма предсказаний первых t деревьев соответственно, x_i – набор

признаков i-го элемента обучающей выборки, f_t - функция (в нашем случае дерево), которую мы хотим обучить на шаге t, $f_t(x_i)$ – предсказание на i-ом элементе обучающей выборки, $\Omega(f)$ - регуляризация функции f

Регуляризация функции вычисляется следующим образом:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2, \quad (2)$$

Где T — количество вершин в дереве, w — значения в листьях, γ и λ — параметры регуляризации.

В следующем шаге с помощью разложения Тейлора до второго члена можем приблизить оптимизируемую функцию $\mathcal{L}^{(t)}$ следующим выражением:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + g_i f_t(x_i) + 0.5 h_i f_t^2(x_i)\right) + \Omega(f_t) \quad (3)$$

В свою очередь:

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}, h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial^2 \hat{y}_i^{(t-1)}} \quad (4)$$

Поскольку мы хотим минимизировать ошибку модели на обучающей выборке, нам нужно найти минимум $\mathcal{L}^{(t)}$ для каждого *t*.

Минимум этого выражения относительно $f_t(x_i)$ находится в точке:

$$f_t(x_i) = \frac{-g_i}{h_i} \quad (5)$$

Каждое отдельное дерево ансамбля $f_t(x_i)$ обучается стандартным алгоритмом.

После принимается приобретенная выборка признаков, для разработки эффективной модели, основанной на методе МО с применением модели XGBoost. Выборка разделяется на обучающую (80%) и тестовую (20%) и производится обучение.

После обучения необходимо оценить ее качество. Площадь под кривой ошибок (ROC AUC) дает нам соотношение между истинно положительными и ложноположительными показателями. Она является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок. Чем больше значение AUC, тем «лучше» модель классификации. Данный показатель часто используется для сравнительного анализа нескольких моделей классификации [7].

На рис. 1 изображена ROC-кривая для взвешенного набора данных. Значение AUC в данном случае: 0.9.

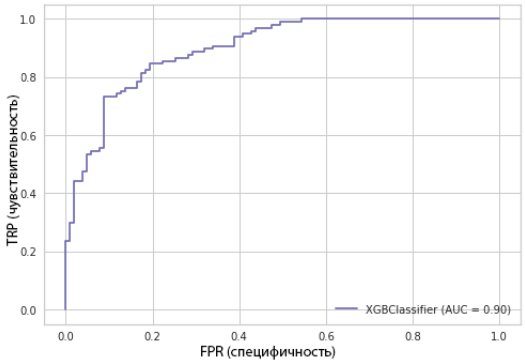

Рис. 1. ROC-кривая для взвешенного набора метода Xboost

Так как решается задача бинарной классификации (есть ли у пациента диабет), разумное значение AUC должно быть больше 0,5, а у хорошей модели классификации показатель AUC больше 0,9 (значение колеблется от сферы применения), наше значение – 0,9, а значит, можно готовить о качественности выбранного метода.

5. Заключение

В работе был проведен обзор модели машинного обучения XGBoost с целью последующего прогноза. Проведен анализ способов разделения значимых признаков с использованием корелляционного анализа, перебалансировка классов SMOTE. Качественность данного решения для этой конкретной задачи оценивается в 0,9 AUC, что является отличным результатом.

Литература

[1] Диабет [Электронный ресурс]. – Режим доступа: https://www.who.int/ru/news-room/fact-sheets/detail/diabetes (10.11.2022).

[2] Sarwar, N. Diabetes Mellitus, Fasting Blood Glucose Concentration, and Risk of Vascular Disease: A Collaborative Meta-Analysis of 102 Prospective Studies / N. Sarwar, P. Gao, S. R. Seshasai, R. Gobin, S. Kaptoge, E. Di Angelantonio, E. Ingelsson, D. Lawlor, E. Selvin, M. Stampfer, C. Stehouwer, S. Lewington, L. Pennells, A. Thompson, N. Sattar, I. White, K. Ray, J. Danesh // Lancet. – 2010. – Vol. 375 . – P. 2215-2222. DOI: 10.1016/S0140-6736(10)60484-9

[3] Awad, M. Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers/ M. Awad, R. Khanna // Apress. – 2015. – P. 268.

[4] Mokshin, A.V. Adaptive genetic algorithms used to analyze behavior of complex system / A.V. Mokshin, V.V. Mokshin, L.M. Sharnin // Communications in Nonlinear Science and Numerical Simulation. – 2019. – Vol. 71. – P. 174–186. DOI: 10.1016/j.cnsns.2018.11.014

[5] Сэмплинг в условиях несбалансированности классов [Электронный ресурс]. – Режим доступа: https://loginom.ru/blog/imbalance-class (12.11.2022).

[6] Zhang, B. Feature selection for global tropospheric ozone prediction based on the BO-XGBoost-RFE algorithm / B. Zhang, Y. Zhang, X. Jiang // Sci Rep. – 2022. – Vol. 12. – P. 9244. DOI: 10.1038/s41598-022-13498-2

[7] Кривая ошибок [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Кривая_ошибок (15.11.2022).