Математические структуры и моделирование
2025. № 2 (74). С. 45–54

УДК 004.896
DOI 10.24147/2222-8772.2025.2.45-54

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ КЛАССИФИКАЦИИ ФЕНОТИПОВ ЗАБОЛЕВАНИЯ ЖЕЛУДОЧНО-КИШЕЧНОГО ТРАКТА С ПОМОЩЬЮ МЕТОДОВ ОБРАБОТКИ ДАННЫХ

С.А. Агалаков[1]
к.ф.-м.н., доцент, e-mail: agalakovsa@gsuite.omsu.ru
А.А. Березин[2]
аспирант, e-mail: andreyberezin55@gmail.com

[1]Омский государственный университет им. Ф.М. Достоевского, Омск, Россия
[2]Омский государственный технический университет, Омск, Россия

Аннотация. Статья посвящена проблеме повышения точности классификации фенотипов заболевания желудочно-кишечного тракта. В ходе предварительных исследований была выявлена модель CatBoost, которая демонстрирует наилучшие результаты в данной задаче, достигая точности 92,85 % на тренировочной выборке и 79,31 % на тестовой выборке. Главной целью данной работы является улучшение точности предсказаний модели, решающей задачу классификации заболеваний, за счёт исследования особенностей исходного набора данных. Предлагаются методы обработки данных, направленные на увеличение качества предсказаний модели, включая поиск и удаление аномальных данных в выборке. Также рассматривается вопрос выбора критерия «аномальности» для специфичных данных. Кроме того, в статье обсуждаются методы работы с проблемой дисбаланса классов, что является важным аспектом для повышения общей эффективности классификации.

Ключевые слова: машинное обучение, задача классификации, логистическая регрессия, нейронные сети, поиск аномальных данных, дисбаланс классов, аугментация данных.

Введение

Нейронные сети и модели машинного обучения набирают особую популярность в наши дни. Эти технологии способны решать множество задач, демонстрируя широкий спектр применения: от голосовых ассистентов до консультантов в промышленности и других профессиональных сферах.

Зачастую на начальном этапе освоения машинного обучения создание моделей кажется весьма простым: модель быстро строится, показывает высокие результаты, и, на первый взгляд, всё работает идеально. Например, в статье авторов [1] рассматривается очень популярная задача распознавания рукописных цифр из базы данных MNIST, где авторы определили модель свёрточной нейросети с наилучшими гиперпараметрами, показав выдающуюся точность, равную 99,41 %. Но тонкость в том,

что для этой задачи был подготовлен и создан датасет из 60 тысяч изображений, каждое из которых было обработано непосредственно для практического обучения начинающих специалистов в области распознавания изображений, что обеспечивало лёгкость в работе с моделью и её высокие показатели точности.

Однако когда дело касается прикладных задач с реальными данными в областях экономики, производства, техники и медицины, то тогда точность стандартных моделей машинного обучения получается невысокой. Это связано с тем, что входные данные зачастую нестандартные, они могут содержать различные искажения или выбросы, что и приводит к менее впечатляющим результатам. В частности, медицинские данные могут содержать аномалии по нескольким причинам, таким как ненормальное состояние пациента, ошибки приборов или записи наблюдений.

1. Постановка задачи исследования

В данной работе рассматривается задача улучшения точности классификации фенотипов заболеваний желудочно-кишечного тракта (ЖКТ) за счёт очистки и нормализации исходного набора данных.

Основная цель исследования – увеличение точности предсказаний через анализ особенностей набора данных, устранение возможных шумов и выбросов, а также корректное представление признаков, значимых для точной классификации.

В процессе предварительных экспериментов наилучшая модель продемонстрировала наилучшие результаты, достигнув точности 92,85 % на тренировочной выборке и 79,31 % на тестовой. Однако наблюдаемое расхождение между этими показателями указывает на потенциальные ограничения качества данных. В этой связи главное внимание в задаче уделено обработке данных, так как её улучшение должно способствовать более точному выявлению закономерностей в рамках задачи классификации, что, в свою очередь, увеличит обобщающую способность модели.

2. Поиск оптимальной модели задачи классификации

Исследование направлено на распределение пациентов по одному из шести классов фенотипов на основании данных анкетирования и медицинских показателей. В исходном наборе данных содержится информация о 281 пациенте, представленная в таблице, для каждого из которых определено 112 признаков. Эти признаки охватывают различные аспекты состояния здоровья человека и его образа жизни.

Цель классификации – определить принадлежность каждого пациента к одной из следующих групп фенотипов заболеваний ЖКТ: здоровые, пациенты с постинфекционным синдромом раздражённого кишечника (ПИ-СРК), пациенты с ожирением, лица с коморбидными состояниями, пациенты с эссенциальным фенотипом и те, у кого проявляется смешанный фенотип.

В процессе решения поставленной задачи изначально были проанализированы модели логистической регрессии и кластерного анализа [2]. Однако, учитывая их низкую эффективность для рассматриваемого набора данных, была проведена оценка более сложных моделей, включая деревья решений, ансамбли деревьев и искусственные нейронные сети [3]. В итоге наилучший результат был достигнут с исполь-

зованием модели градиентного бустинга CatBoost [4], который продемонстрировал точность 92,85 % на тренировочной выборке и 79,31 % на тестовой выборке (табл. 1).

Также следует рассмотреть дополнительные метрики: точность и полноту для каждого класса модели CatBoost на тренировочной и тестовой выборках (рис. 1).

Таблица 1. Результаты классификации на исходных данных

Модель	**Точность на тренировочной выборке, %**	**Точность на тестовой выборке, %**
Логистическая регрессия	75,00	68,97
Кластеризация методом k-средних	—	23,49
MLPClassifier	81,74	79,31
Нейросеть Keras	73,00	68,97
Решающее дерево	82,53	62,06
Случайный лес	77,78	72,41
XGBoost	67,86	62,07
CatBoost	92,85 %	79,31 %

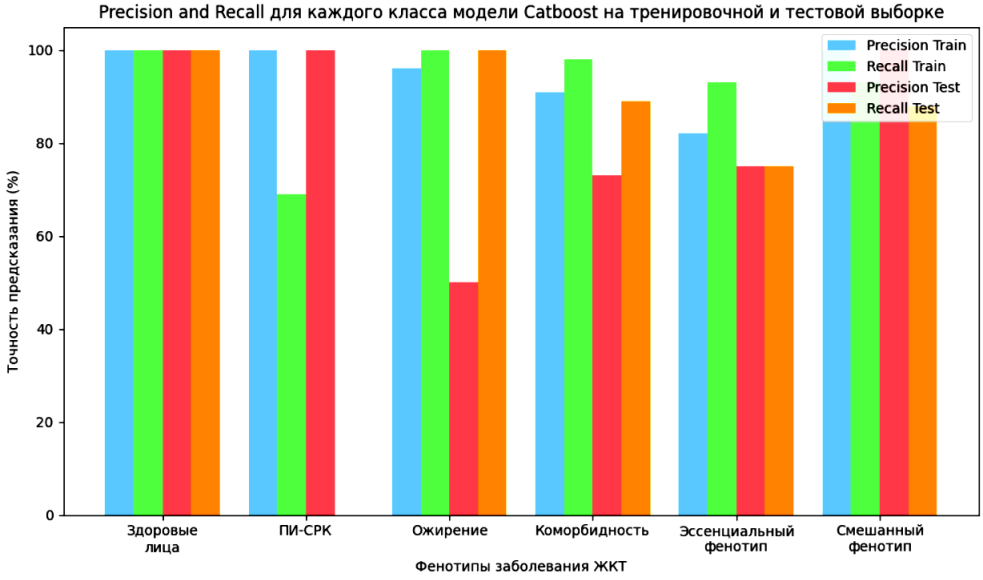

Рис. 1. Диаграмма точности и полноты CatBoost по классам

Например, при анализе полноты (recall) для каждого класса на тестовой выборке наблюдается следующее: для двух классов модель корректно определяет все наблюдения, для двух других классов полнота составляет около 89 %.

Однако для класса «1» (ПИ-СРК) полнота равна нулю. Более детальный анализ предсказанных значений показал, что в тестовой выборке представлены два наблюдения данного класса, и оба были классифицированы неверно. Это приводит к зна-

чению полноты, равному нулю, что может свидетельствовать о наличии проблемы с количеством или качеством данных для этого класса в выборке.

3. Поиск аномальных данных

В задаче классификации, когда различные модели показывают низкую точность, имеет смысл проанализировать датасет на наличие аномальных данных. Возникает вопрос о надёжности собранных наблюдений, так как данные могут содержать аномалии по разным причинам: нестабильное состояние пациентов, ошибки оборудования или погрешности при записи данных. Для улучшения качества модели будет проведена очистка данных от выбросов с использованием различных методов.

3.1. Критерий аномальности на основе графика расстояний между метками

Следующая идея основана на методах удаления выбросов, аналогичных подходу для линейной регрессии, где вычисляются остатки – разница между предсказанными и фактическими значениями (рис. 2).

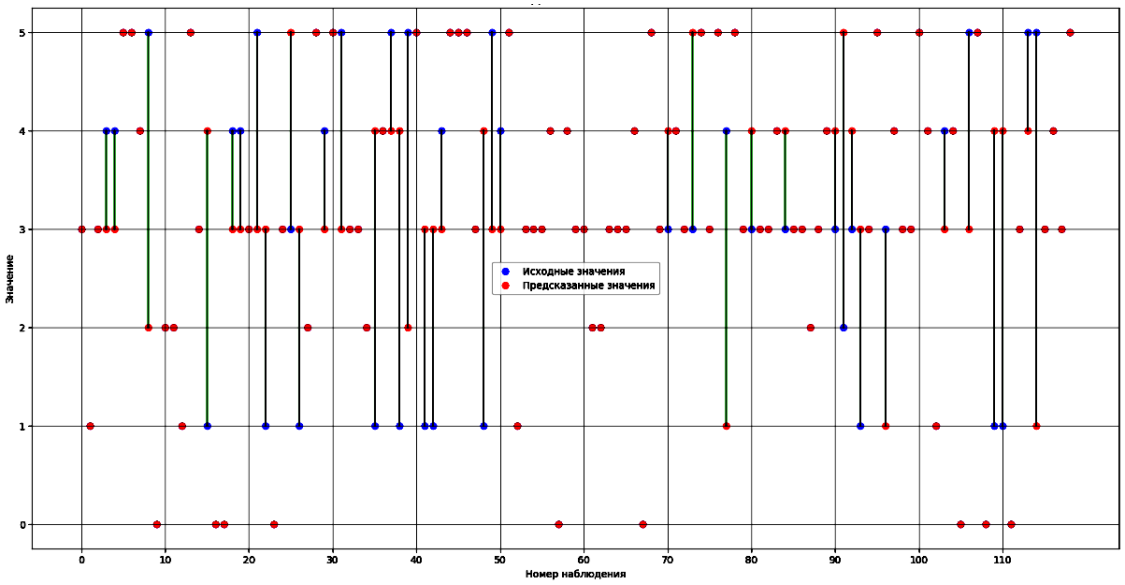

Рис. 2. График исходных значений и предсказанных

Проблемой является то, что для логистической регрессии этот метод сложнее применить, поскольку она выводит не непрерывные значения, а классы. В нашем случае результатом классификации являются метки $\{0, 1, 2, 3, 4, 5\}$. Если предположить, что эти метки имеют смысловое различие, например, 0 соответствует «здоровым», а 5 – «смешанному фенотипу», то сильное расхождение между исходной меткой и предсказанным значением можно рассматривать как выброс. На основании этого было принято решение удалить данные, где разница между исходной меткой и предсказанным классом превышает 2.

После удаления таких наблюдений точность моделей логистической регрессии и нейросетей снизилась по сравнению с исходными результатами. Это свидетельствует о том, что данный подход оказался неэффективным для улучшения качества классификации в рамках поставленной задачи. Скорее всего, это объясняется тем, что удаление наблюдений с расхождением меток больше чем на 2 могло привести к потере значимой информации, которая могла способствовать более точной дифференциации классов. Таким образом, данный метод обработки данных не подходит для решения задачи классификации фенотипов заболеваний ЖКТ, и необходимы другие подходы для выявления и устранения аномальных данных.

3.2. Критерий аномальности на основе выходных значений вероятностей логистической регрессии

С учётом того, что удаление выбросов на основе расстояний между метками, как это делается в линейной регрессии, не привело к улучшению результатов, была предложена идея: анализировать не предикторы, а вероятности принадлежности наблюдений к классам, которые также предоставляются моделью логистической регрессии [5]. В частности, мы сосредоточим внимание на ошибочных вероятностях: если предсказание оказывается неверным с вероятностью более 50 %, мы будем считать такое наблюдение выбросом.

Применение этого подхода к поиску и удалению выбросов позволило добиться увеличения точности моделей. В частности, для MLPClassifier точность на тестовой выборке повысилась до 82,60 %, а на тренировочной – 100 % (табл. 2).

Таблица 2. Модели с наилучшей точностью после удаления выбросов

Модель	**Параметры**	**Точность на тренировочной выборке, %**	**Точность на тестовой выборке, %**
MLPClassifier	{solver = “sgd”, hidden_layer_sizes: (42, 50, 6), alpha = 0.1, max_iter = 100}	100	82,60
XGBoost	{learning_rate = 0.1, max_depth = 2, n_estimators = 50}	95,65	78,26

Важно отметить, что в ходе этого процесса из исходного датасета было исключено 46 наблюдений из 281, что составляет приблизительно 16 % от общего объема данных. Таким образом, данный метод представляется более эффективным в сравнении с предыдущим подходом.

4. Проблема несбалансированности классов

После применения множества моделей, подбора параметров и анализа аномалий необходимо провести детальный анализ распределения классов в данных. На предыдущих этапах работы не уделялось внимания количеству наблюдений и их распределению по классам, однако если рассчитать общее число наблюдений, то можно увидеть следующее распределение по классам (табл. 3).

Таблица 3. Количество наблюдений по классам

Фенотип	**Количество пациентов**
Здоровые лица	18
ПИ-СРК	31
Ожирение	24
Коморбидность	91
Эссенциальный	51
Смешанный	66

Явно наблюдается неравномерное распределение пациентов по классам. Это несоответствие может быть причиной низкой точности классификации для всех моделей машинного обучения. В частности, некоторые классы могут быть недостаточно представлены в обучающем наборе, что приводит к тому, что нейросеть может не выявлять закономерности в данных, а просто определять результат в пользу мажоритарного класса.

4.1. Методы работы с несбалансированными классами

В настоящее время тема дисбаланса классов является актуальной и важной в машинном обучении. Существует несколько способов, которые могут повысить точность классификации при наличии несбалансированных классов [6]:

1. **Сбор дополнительных данных.** К сожалению, в рамках данной работы нет возможности собрать новые наблюдения.

2. **Сокращение количества наблюдений для выравнивания классов по миноритарному представленному классу.** Не подходит для данной задачи, так как выборка уже небольшая. Уменьшение количества наблюдений с 281 до 108 приведёт к значительной потере данных, что негативно скажется на способности моделей извлекать уникальные признаки, важные для дальнейшей диагностики новых пациентов.

3. **Добавление копий.** Данный метод также не подходит, поскольку он подразумевает генерацию множества одинаковых данных, что не гарантирует оптимальность модели. Это может привести к переобучению, так как модель будет обучаться на одних и тех же примерах.

4. **Взвешивание классов при обучении модели и применение стратификации при разделении данных на обучающий и тестовый наборы.** Эти методы были реализованы в данной работе, однако не привели к улучшению результатов.

5. **Аугментация данных.** Данный подход представляет собой интересное решение, заключающееся в генерации новых образцов на основе имеющихся данных с помощью изменения некоторых признаков определёнными алгоритма-

ми. Аугментация способствует увеличению точности, и поэтому её следует рассмотреть более подробно.

4.2. Аугментация данных

Алгоритм SMOTE включает в себя следующие шаги [7]:

1. **Определение наименее представленного класса.**
2. **Определение количества синтетических образцов.**
3. **Определение соседей:** для каждого образца в миноритарном классе выбираются ближайшие соседи – схожие образцы из того же класса, используя метод k-ближайших соседей.
4. **Генерация синтетических образцов:** выбирается один из ближайших соседей, и на основе разности между ним и текущим образцом генерируется новый синтетический образец. Эта разность умножается на случайное число от 0 до 1, и результат добавляется к исходному образцу (рис. 3).

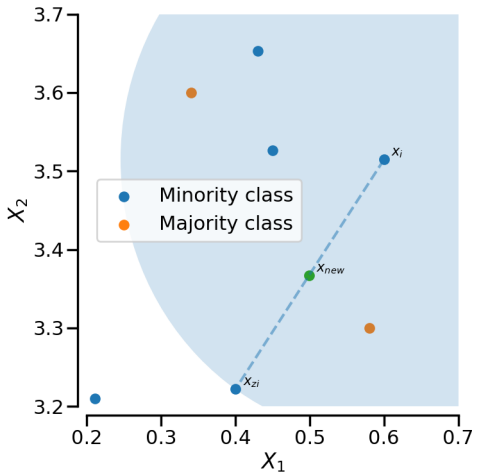

Рис. 3. Визуализация алгоритма SMOTE

Шаги 3 и 4 повторяются до тех пор, пока количество образцов в классах не станет равным.

Существует несколько модификаций данного алгоритма, среди которых SMOTE, ADASYN, Border-SMOTE.

- **ADASYN** работает аналогично обычному SMOTE, однако количество сгенерированных образцов для каждого наблюдения пропорционально количеству образцов, которые не принадлежат тому же классу, что и данное наблюдение, в заданном окружении. Это приводит к большему количеству сгенерированных образцов в областях, где не соблюдается правило ближайшего соседа [6].

- **Border-SMOTE** также генерирует синтетические образцы между ближайшими соседями, но сосредоточен на наблюдениях, находящихся на границе классов. Генерация синтетических образцов выполняется только для таких граничных наблюдений, что помогает минимизировать ошибки классификации.

В результате применения алгоритма SMOTE на новой выборке с синтезированными образцами была достигнута лучшая точность классификации для всех моделей (табл. 4).

Таблица 4. Результаты на выборке с синтезированными данными

Модель	**Train, %**	**Test, %**	**Описание**
Логистическая регрессия	86,88	79,63	SMOTE после удаления выбросов
Решающее дерево	92,91	79,63	Border-SMOTE после удаления выбросов
MLPClassifier	100,00	90,74	SMOTE после удаления выбросов
Случайный лес	88,33	83,33	SMOTE после удаления выбросов
XGBoost	100,00	87,27	SMOTE без нормирования и удаления выбросов
Нейросеть Keras	97,50	87,03	SMOTE после удаления выбросов

В ходе этого этапа исследования было установлено, что метод ADASYN не подходит для данного набора данных, поскольку в нем существуют области, где ни один из ближайших соседей не принадлежит к классу большинства. Это приводит к возникновению ситуации, когда происходит деление на ноль и возникают значения NaN.

В результате простой реализации SMOTE продемонстрировала лучшие результаты. Однако, учитывая, что часть данных является искусственно сгенерированной, был проведён дополнительный эксперимент. В этом эксперименте выборка была разделена на тренировочную и тестовую так, чтобы тестовая выборка вовсе не содержала синтезированных образцов. Таким образом, модели обучались на данных с искусственными образцами, а затем тестировались на части исходных данных, которые не участвовали в аугментации. Результаты этого эксперимента представлены в табл. 5.

Как видно из результатов, точность на тестовой выборке значительно ниже, чем на тренировочной, причём разница почти вдвое. Это свидетельствует о том, что метод SMOTE не применим для нашей задачи, так как сгенерированные данные не обеспечивают адекватной обобщающей способности моделей.

Заключение

В данной работе была рассмотрена задача повышения качества классификации фенотипов заболеваний ЖКТ с использованием различных методов обработки дан-

Таблица 5. Результаты на выборке без искусственных данных

Модель	**Точность на тренировочной выборке, %**	**Точность на тестовой выборке, %**
Логистическая регрессия	83,94	55,17
Решающее дерево	94,51	44,82
MLPClassifier	100,00	58,62
Случайный лес	88, 61	68,96
XGBoost	100, 00	55,17

ных. В результате тщательного поиска были выявлены ключевые проблемы в датасете, такие как наличие выбросов и дисбаланс классов, которые отрицательно сказались на точности классификации. Использование методов, таких как логистическая регрессия и градиентный бустинг CatBoost, позволило достичь значительных результатов, особенно после очистки данных, что подтверждает важность обработки данных для повышения качества предсказаний.

Одним из основных достижений работы стало успешное применение подхода к удалению аномальных данных на основе вероятностей логистической регрессии, что позволило улучшить точность классификации до 82,60 % для модели MLPClassifier. Несмотря на то, что метод аугментации данных с использованием SMOTE не дал ожидаемых результатов, он всё же привлёк внимание к возможности улучшения классификации с помощью синтетических образцов. Этот опыт подчёркивает, что не все подходы могут быть универсальными, и важно находить методы, подходящие конкретному набору данных.

Полученные результаты подтверждают, что более глубокий анализ данных и правильный выбор методов их обработки могут существенно влиять на качество предсказаний. Для дальнейшего исследования в данной области необходимо собрать дополнительные данные и привлечь медицинских специалистов для более точного определения классов и выбора признаков, имеющих наибольшее значение для диагностики. Это позволит значительно повысить эффективность моделей и улучшить качество диагностики заболеваний ЖКТ.

Литература

1. Shevlyakov A.N., Berezin A.A. Recognition of the MNIST dataset with defective rows // Journal of Physics: Conference Series: 15, Virtual, Online, 09–11 November 2021. Virtual, Online, 2022. P. 012031.
2. Агалаков С.А. Статистические методы анализа данных : учеб. пособие. Омск: Изд-во Ом. гос. ун-та, 2017. 92 с.
3. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер, 2018. 480 с.
4. Официальная документация CatBoost. URL: `https://catboost.ai/en/docs/` (дата обращения: 25.02.2024).

5. Anomaly Detection. URL: https://vs.inf.ethz.ch/edu/HS2011/CPS/papers/chandola09_anomaly-detection-survey.pdf (дата обращения: 10.11.2024).
6. Eight tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. URL: https://machinelearningdev.blogspot.com/2016/08/8-tactics-to-combat-imbalanced-classes.html (дата обращения: 10.11.2024).
7. Официальная документация imbalanced-learn (SMOTE). URL: https://imbalanced\-learn.org/stable/over_sampling.html\#smote-variants (дата обращения: 10.11.2024).

IMPROVING THE EFFICIENCY OF CLASSIFICATION OF GASTROINTESTINAL DISEASE PHENOTYPES USING DATA PROCESSING METHODS

S.A. Agalakov[1]
Ph.D. (Phys.-Math.), Associate Professor, e-mail: agalakovsa@gsuite.omsu.ru
A.A. Berezin[2]
Ph.D. Student, e-mail: andreyberezin55@gmail.com

[1]Dostoevsky Omsk State University, Omsk, Russia
[2]Omsk State Technical University, Omsk, Russia

Abstract. The article is devoted to the problem of improving the accuracy of classification of phenotypes of diseases of the gastrointestinal tract. In the course of preliminary research, the CatBoost model was identified, which demonstrates the best results in this task, achieving an accuracy of 92.85 % in the training sample and 79.31 % in the test sample. The main goal of this work is to improve the accuracy of predictions of the model solving the problem of classification of diseases by studying the characteristics of the original data set. Data processing methods aimed at increasing the quality of model predictions, including the search and removal of abnormal data in the sample, are proposed. The issue of choosing the criterion of "abnormality" for specific data is also considered. In addition, the article discusses methods of dealing with the problem of class imbalance, which is an important aspect for improving the overall efficiency of classification.

Keywords: machine learning, classification problem, logistic regression, neural networks, search for abnormal data, class imbalance, data augmentation.

Дата поступления в редакцию: 11.11.2024