УДК 001.817

ВЕБ-ПРИЛОЖЕНИЕ ДЛЯ РЕКОМЕНДАЦИИ МЕТОДОВ БАЛАНСИРОВКИ ДАННЫХ

М. М. Лукашевич[1)], Е. Клицунова[2)]

[1)] *Белорусский государственный университет, Минск, Беларусь, LukashevichMM@bsu.by*
[2)] *Белорусский государственный университет, Минск, Беларусь, kateryna.klitsunova@gmail.com*

Статья посвящена разработке программного средства для решения проблемы выбора методов обработки несбалансированных данных. Представлено веб-приложение, интегрирующее модель машинного обучения, которая рекомендует алгоритмы балансировки на основе характеристик входного набора данных. Описаны процесс формирования обучающих данных, проектирование модели и архитектура веб-приложения, реализованного на фреймворке Streamlit.

Ключевые слова: несбалансированные данные; машинное обучение; балансировка данных; рекомендательная система; веб-приложение.

WEB APPLICATION FOR RECOMMENDING DATA BALANCING METHODS

M. M. Lukashevich[a)], K. Klitsunova[b)]

[a)] *Belarusian State University, Minsk, Belarus, LukashevichMM@bsu.by*
[b)] *The United Institute of Informatics Problems, Minsk, Belarus, kateryna.klitsunova@gmail.com*

The article is devoted to the development of a software tool for solving the problem of choosing methods for processing imbalanced data. A web application is presented that integrates a machine learning model that recommends balancing algorithms based on the characteristics of the input dataset. The process of generating training data, designing the model, and the architecture of the web application implemented on the Streamlit framework are described.

Keywords: imbalanced data; machine learning; data balancing; recommender system; web application.

1. Введение

Широкое распространение задач бинарной и мультиклассовой классификации в таких критически важных областях, как медицина, финансовая безопасность и инженерия, часто сопряжено с проблемой

несбалансированности данных – существенного преобладания объектов одного или нескольких классов.

Для решения данной проблемы предложен широкий спектр методов [1–4], которые условно можно разделить на следующие категории: методы на уровне данных (балансировка путем синтетического увеличения меньшего класса или уменьшения большего), методы на уровне алгоритмов (например, обучение с учетом издержек классификации) и ансамблевые методы. Однако отсутствие универсального подхода, эффективного для всех типов задач, и многообразие доступных техник балансировки создают сложность выбора подходящей техники на практике.

В связи с этим актуальной задачей является разработка инструмента, способного автоматизировать и оптимизировать процесс выбора метода предобработки данных на основе их объективных характеристик. Целью данной работы является описание разработки программного средства – интерактивного веб-приложения, которое интегрирует в себя модель машинного обучения для рекомендации наиболее эффективных методов и их комбинаций для балансировки конкретного набора данных, основанную на результатах предварительного комплексного экспериментального исследования.

2. Формирование обучающих данных для модели

Основой для построения модели рекомендаций послужили результаты экспериментального исследования, подробно описанного в предыдущей работе [5]. Целью эксперимента была всесторонняя оценка влияния различных методов балансировки, а также их комбинаций, на качество моделей классификации.

Было отобрано пять наборов данных с платформы Kaggle объемом от ~2 тыс. до ~254 тыс. записей и с разной степенью дисбаланса (от 2:100 до 29:100), их описание приводится в табл. 1.

Для каждого набора данных был выполнена стандартная предобработка: обработка пропущенных значений и выбросов, кодирование категориальных признаков, нормализация числовых признаков и стратифицированное разделение на обучающую и тестовую выборки в соотношении 70 к 30.

Таблица 1

Описание наборов данных

Название	Размер набора данных	Тип классификации	Соотношение классов
Классификация здоровья плода	~2 тысячи записей	мультиклассовая	8:100 и 13:100
Прогнозирование церебрального инсульта	~43 тысячи записей	бинарная	2:100

Окончание табл. 1

Название	Размер набора данных	Тип классификации	Соотношение классов
Кредитный риск	~32 тысячи записей	бинарная	22:100
Классификация кредитного рейтинга	100 тысяч записей	мультиклассовая	17:100 и 29:100
Показатели здоровья при диабете	~254 тысячи записей	мультиклассовая	2:100 и 15:100

На каждом из подготовленных наборов данных проводилось обучение и оценка качества множества моделей. Тестировались как классические алгоритмы (решающее дерево, k-ближайших соседей, метод опорных векторов, наивный байесовский классификатор), так и ансамблевые (случайный лес, градиентный бустинг, XGBoost, LightGBM, AdaBoost). Для каждой модели фиксировалось значение сбалансированной точности (Balanced Accuracy). Балансировка данных проводилась изолированно следующими методами.

Увеличение меньшего класса: Random Oversampling, SMOTE, Borderline-SMOTE, Borderline-SMOTE SVM, ADASYN.

Уменьшение большего класса: Random Undersampling, NearMiss, Tomek Links, Condensed Nearest Neighbors, Edited Nearest Neighbors, One-Sided Selection, Neighborhood Cleaning Rule.

Далее исследовалось комбинированное применение наиболее эффективных методов из обеих групп. Для каждого набора данных, каждой модели и каждого метода балансировки (как изолированного, так и комбинированного) сохранялось итоговое значение сбалансированной точности.

Для создания набора данных для обучения модели рекомендаций из результатов эксперимента были извлечены следующие признаки для каждого наблюдения:

1) Size: объем набора данных в тысячах записей;
2) Imbalance Ratio: степень дисбаланса, выраженная как отношение количества объектов в наименьшем классе к количеству объектов в наибольшем классе;
3) Type: тип классификации, бинарная или мультиклассовая;
4) Oversampling: метод увеличения меньшего класса (если применялся);
5) Undersampling: метод уменьшения большего класса (если применялся);
6) Balanced Accuracy: значение сбалансированной точности, достигнутое после применения данной комбинации методов.

Поскольку абсолютные значения метрики сильно варьировались от набора к набору, для нивелирования этого эффекта была проведена

минимаксная нормализация значения сбалансированной точности в пределах каждого исходного набора данных. Это позволило перейти от абсолютных значений точности к относительным рангам эффективности методов внутри каждого конкретного кейса.

Категориальные признаки (Type, Oversampling, Undersampling) были преобразованы в числовой формат с использованием кодирования меток (Label Encoding). Итоговый обучающий набор представлял собой таблицу, где каждой строке соответствовала уникальная комбинация характеристик набора данных и примененных методов балансировки, а целевой переменной выступало нормализованное значение эффективности этой комбинации.

3. Проектирование и обучение модели ранжирования

Задачей модели было ранжирование методов балансировки по предполагаемой эффективности для новых данных.

В качестве базовых алгоритмов были выбраны регрессоры, показавшие высокую эффективность в задачах ранжирования [6, 7]: Gradient Boosting Regressor, Extra Trees Regressor, Random Forest Regressor, XGBoost Regressor и LightGBM Regressor. Для настройки их гиперпараметров и объективного сравнения был применен фреймворк Optuna, который проводил поиск по заданному пространству гиперпараметров для каждого алгоритма.

Оценка качества моделей в процессе кросс-валидации проводилась с использованием метрики R^2 (коэффициент детерминации), так как она хорошо интерпретирует долю дисперсии, объясненную моделью. Для обеспечения репрезентативности оценки использовалась 5-кратная кросс-валидация с перемешиванием (Stratified K-Fold).

По итогам оптимизации были отобраны три алгоритма, показавшие наилучшие и наиболее стабильные результаты: XGBoost Regressor, Gradient Boosting Regressor и Extra Trees Regressor.

Для повышения обобщающей способности и устойчивости предсказаний была протестирована идея объединения лучших моделей в ансамбль. Были опробованы следующие подходы: бэггинг, стекинг и голосующий ансамбль.

Наилучшие результаты показал ансамбль на основе голосования (Voting Regressor), объединивший три лучших алгоритма, который и был выбран в качестве финальной модели.

Для количественной оценки были рассчитаны стандартные метрики регрессии: средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE) и коэффициент детерминации (R^2) для каждого набора кросс-валидации (табл. 2).

Таблица 2

Обобщение результатов экспериментальных исследований

Номер блока	MSE	MAE	Коэффициент детерминации
1	0,01135477	0,07120671	0,69812307
2	0,02629933	0,10227519	0,60890527
3	0,00695144	0,04180257	0,65440132
4	0,01836801	0,05555495	0,59469433
5	0,01021363	0,0535413	0,79880918

Полученные значения подтвердили адекватную предсказательную способность модели и ее применимость для задачи ранжирования.

4. Разработка веб-приложения

Для практического применения модели было разработано веб-приложение. Для его реализации был выбран язык программирования Python, что обеспечило бесшовную интеграцию с уже разработанной моделью машинного обучения и библиотеками для работы с данными (Pandas, NumPy, Scikit-learn, Imbalanced-learn).

Для реализации веб-интерфейса был выбран фреймворк Streamlit. Его основное преимущество – возможность создания интерактивных веб-приложений с богатыми визуальными возможностями практически целиком на языке Python, без необходимости написания клиент-серверного кода. Это значительно ускорило процесс разработки и позволило сосредоточиться на логике приложения.

Приложение развернуто в облачной среде Streamlit Community Cloud, не требует установки и доступно по публичной ссылке https://resample.streamlit.app/. Интерфейс выполнен на английском языке для универсальности и ориентирован на специалистов в области машинного обучения (рис. 1).

Работа пользователя с приложением следует линейной последовательности.

1. Загрузка данных. Пользователь выбирает файл или демонстрационный набор данных.

2. Предобработка. Приложение определяет наличие пропусков и предлагает варианты их обработки.

3. Рекомендация. Пользователь получает список лучших комбинаций методов наилучшей предполагаемой эффективностью (рис. 2).

4. Балансировка. Пользователь выбирает желаемый метод и параметры балансировки из выпадающих списков.

5. Визуализация и экспорт. Приложение показывает диаграммы соотношения классов до и после балансировки и предоставляет кнопку для скачивания сбалансированного набора данных.

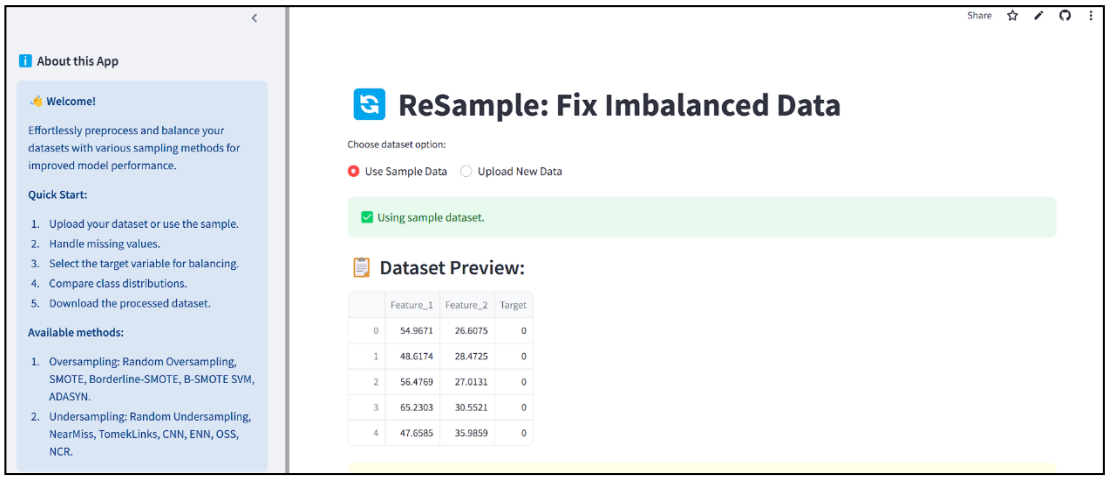

Рис. 1. Интерфейс веб-приложения

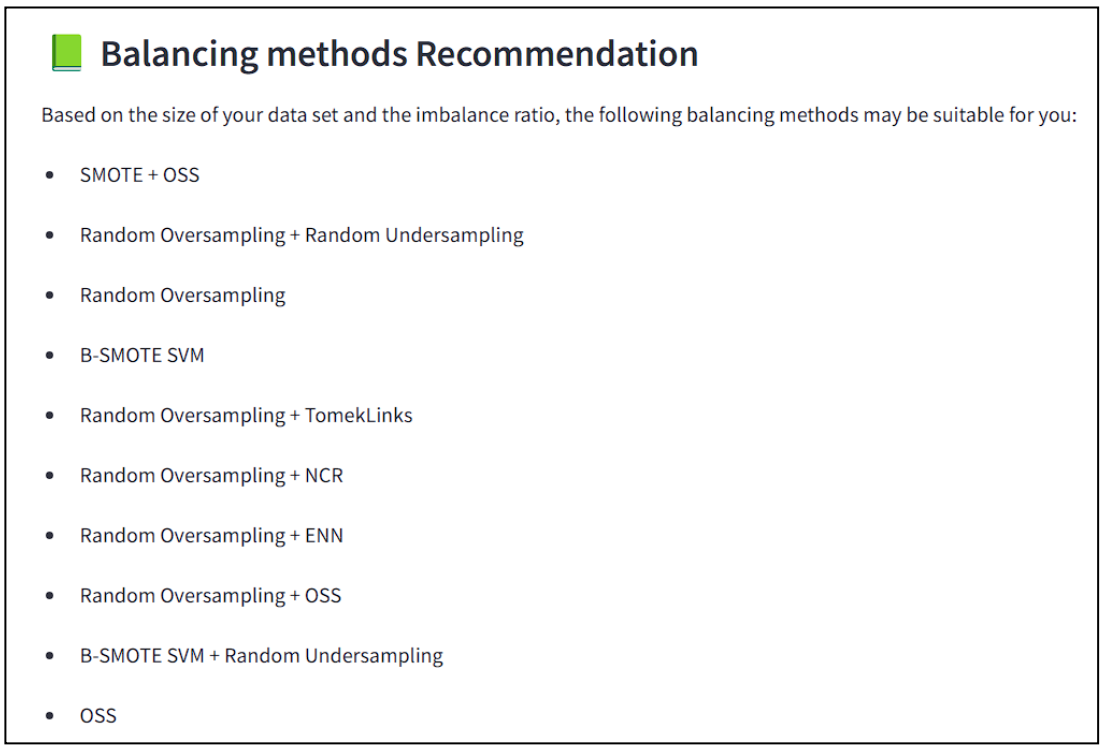

Рис. 2. Рекомендация методов балансировки в веб-приложении

Приложение развернуто в облачной среде и доступно онлайн, что обеспечивает его кроссплатформенность и удобство использования.

5. Заключение

В данной статье представлена разработка программного средства, предназначенного для решения актуальной проблемы выбора методов обработки несбалансированных данных в машинном обучении. Основное преимущество предлагаемого решения заключается в интеграции модели

машинного обучения, обученной на результатах комплексного экспериментального исследования, в удобный и доступный веб-интерфейс.

Разработанное программное средство может быть использовано специалистами по анализу данных и машинному обучению в различных предметных областях для быстрого старта работы с несбалансированными наборами данных и повышения качества построенных классификационных моделей.

Библиографические ссылки

1. Classification of imbalanced data: review of methods and applications / P. Kumar [et al.] // IOP conference series: materials science and engineering. IOP Publishing, 2021. Vol. 1099, №. 1. Article no. 012077.

2. *Krawczyk B.* Learning from imbalanced data: open challenges and future directions // Progress in artificial intelligence. 2016. Vol. 5, №. 4. P. 221–232.

3. *Branco P., Torgo L., Ribeiro R.* A survey of predictive modeling on imbalanced domains // ACM computing surveys (CSUR). 2016. Vol. 49, №. 2. P. 1–50.

4. *Sun Y., Wong A. K. C., Kamel M. S.* Classification of imbalanced data: A review // International journal of pattern recognition and artificial intelligence. 2009. Vol. 23, №. 4. P. 687–719.

5. *Клищунова Е., Лукашевич М. М*. Сравнительный анализ методов балансировки данных для задач машинного обучения // BIG DATA и анализ высокого уровня : сборник научных статей XI Международной научно-практической конференции / Белорус. гос. ун-т информатики и радиоэлектроники ; редкол.: В. А. Богуш [и др.]. Минск : БГУИР, 2025. С. 74–83.

6. *Li H.* A short introduction to learning to rank // IEICE TRANSACTIONS on Information and Systems. 2011. Vol. 94, №. 10. P. 1854–1862.

7. Learning to rank for information retrieval / T. Y. Liu [et al.] // Foundations and Trends in Information Retrieval. 2009. Vol. 3, №. 3. P. 225–331.