DOI 10.37882/2223-2966.2026.01.38

ИССЛЕДОВАНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ ОБРАБОТКИ ДИСБАЛАНСА ДАННЫХ НА СИНТЕТИЧЕСКИХ НАБОРАХ

INVESTIGATION OF THE EFFECTIVENESS OF DATA IMBALANCE PROCESSING METHODS ON SYNTHETIC DATASETS

K. Shakirov

Summary. The article discusses the problem of class imbalance in machine learning. Various resampling methods for solving this problem are compared. A study using synthetically generated data with varying degrees of imbalance from 10 % to 90 % of the minority class is presented. The data was trained on a random forest model. Various methods of resampling to the training sample were analyzed: without processing, random oversampling (Random Over), SMOTE, random reduction of the sample (Random Under) and SMOTETomek. The effectiveness of the methods was evaluated using the following metrics: Accuracy, area under the ROC curve (ROC-AUC), Precision, Recall, and F1-measure. The results showed that the SMOTETomek method demonstrates the best performance among the considered approaches.

Keywords: data imbalance, imbalance processing methods, synthetic data, Random Over, SMOTE, Random Under, SMOTETomek, quality metrics, machine learning.

Шакиров Кирилл Фаридович
старший преподаватель, Федеральное государственное бюджетное образовательное учреждение высшего образования Российский экономический университет имени Г.В. Плеханова
SHakirov.KF@rea.ru

Аннотация. В статье рассматривается проблема дисбаланса классов в машинном обучении. Приводится сравнение различных методов ресемплинга для решения данной проблемы. Представлено исследование с использованием синтетически сгенерированных данных с варьированием степени дисбаланса от 10 % до 90 % миноритарного класса. Данные обучались на модели случайного леса. Были проанализированы различные методы ресемплинга применительно к обучающей выборке: без обработки, случайное передискретизирование (Random Over), SMOTE, случайное уменьшение выборки (Random Under) и SMOTETomek. Оценка эффективности методов проводилась по метрикам: точность (Accuracy), площадь под ROC-кривой (ROC-AUC), прецизионность (Precision), полнота (Recall) и F1-мера. Результаты показали, что метод SMOTETomek демонстрирует наилучшие показатели среди рассмотренных подходов.

Ключевые слова: дисбаланс данных, методы обработки дисбаланса, синтетические данные, Random Over, SMOTE, Random Under, SMOTETomek, метрики качества, машинное обучение.

Введение

При решении задач одной из распространенных проблем является дисбаланс классов данных. Она характеризуется превышением количества объектов одного класса количества объектов другого класса. Подобный дисбаланс не может не отразиться на эффективности работы моделей машинного обучения. Модели машинного обучения игнорируют минорный класс, а предпочтение отдают мажоритарному классу [8, с. 1].

Данная работа посвящена эксперименту, направленному на исследование эффективности различных подходов к обработке дисбаланса классов на синтетических данных. Использование синтетических данных позволяет контролировать параметры исследования и детально анализировать результаты работы различных методов обработки дисбаланса классов.

Цель эксперимента — сравнение популярных методов обработки дисбаланса данных и определение их эффективности при различных уровнях дисбаланса.

Методология

Основные этапы эксперимента и используемые методы были следующими:

Первоначально была выполнена генерация синтетических данных с использованием функции make_classification из библиотеки scikit-learn [10].

Затем данные были разделены на обучающую и тестовую выборки с сохранением стратификации [2, с. 112].

К данным применялись методы обработки дисбаланса классов [6, с. 112]: Random Over, SMOTE, Random Under, SMOTETomek.

После этого осуществлялось обучение модели Random Forest Classifier на обработанных и необработанных данных.

Оценка качества моделей проводилась с использованием следующих метрик [9, с. 4]: Accuracy, ROC-AUC, Precision, Recall, F1-мера.

Синтезированные данные имели следующие параметры: количество генерируемых объектов — 10 000; количество признаков — 20 (из них 15 информативных и 5 избыточных); уровни дисбаланса изменялись от 10 % до 90 % миноритарного класса с шагом 10 %; количество повторений для каждого уровня дисбаланса — 10.

Литературный обзор

Несмотря на актуальность темы дисбаланса классов, в российском научном сегменте за последние пять лет опубликовано сравнительно немного работ, посвящённых проблемам его обработки. Исследования затрагивают различные сферы — от экономики [3, с. 26; 5, с. 145] и географии [4, с. 89] до задач распознавания эмоций по изображениям [7, с. 685] и медицинских исследований [1, с. 82]. Однако большинство работ сосредоточено на решении частных задач в конкретных отраслях и не охватывает более общие аспекты борьбы с дисбалансом. В публикациях, как правило, анализируются существующие проблемы дисбаланса классов на определённых датасетах и выявляются наиболее эффективные методы его обработки только для данных наборов данных.

Следует отметить, что практически отсутствуют исследования, в которых рассматривались бы методы обработки дисбаланса данных в зависимости от степени его выраженности. Подобные работы могли бы существенно облегчить решение практических задач, связанных с дисбалансом классов данных.

Данная работа призвана восполнить существующий пробел в отечественных исследованиях.

Результаты исследования

Полученные по завершении исследования результаты эффективности методов обработки классов данных были сведены в единую таблицу. В ней представлены средние значения метрик для каждого из подходов к обработке классов в зависимости от уровня дисбаланса.

Анализ результатов позволяет сделать следующие выводы:

- При низком уровне дисбаланса (10–30 % миноритарного класса) методы SMOTE и SMOTETomek показывают более высокие значения Recall и Precision по сравнению с другими методами. При этом Random Under в некоторых случаях демонстрирует хорошие результаты по Precision, но уступает по Recall.
- При среднем уровне дисбаланса (30–70 % миноритарного класса) методы SMOTE и SMOTETomek продолжают показывать устойчиво высокие результаты, сохраняя баланс между Precision и Recall. Метод Random Over также демонстрирует удовлетворительные показатели, особенно по Precision.
- При высоком уровне дисбаланса (70–90 % миноритарного класса) методы SMOTE и SMOTETomek обладают значительным преимуществом по сравнению с другими методами, особенно по метрике Recall. Метод Random Over показывает хорошие результаты, но всё же уступает SMOTE и SMOTETomek.

Проведённый эксперимент показал, что наиболее эффективными методами при различных уровнях дисбаланса классов являются SMOTE и SMOTETomek. При этом следует отметить, что для повышения точности предсказаний положительного класса (Precision) целесообразно использовать метод Random Under, тогда как по метрике Recall более стабильные результаты демонстрирует метод Random Over, хотя он и уступает SMOTE и SMOTETomek.

Таблица 1.

Значения метрик и уровня дисбаланса

№	Disbalance	Method	Repeat	Accuracy	ROC_AUC	Precision	Recall
0	0.1	Original	4.5	0.95335	0.962173	0.983979	0.560451
1	0.1	Random Over	4.5	0.95800	0.968548	0.959036	0.623424
2	0.1	Random Under	4.5	0.91890	0.956993	0.573902	0.893263
3	0.1	SMOTE	4.5	0.96630	0.967833	0.874317	0.789367
4	0.1	SMOTETomek	4.5	0.96595	0.968231	0.876255	0.783145
5	0.2	Original	4.5	0.94600	0.977171	0.968890	0.757998
6	0.2	Random Over	4.5	0.95400	0.980067	0.947500	0.818615
7	0.2	Random Under	4.5	0.93350	0.975318	0.789078	0.921392
8	0.2	SMOTE	4.5	0.95625	0.980230	0.906893	0.874568
9	0.2	SMOTETomek	4.5	0.95745	0.980378	0.907617	0.880245

№	Disbalance	Method	Repeat	Accuracy	ROC_AUC	Precision	Recall
10	0.3	Original	4.5	0.94755	0.983117	0.965607	0.856765
11	0.3	Random Over	4.5	0.95250	0.984516	0.950342	0.889047
12	0.3	Random Under	4.5	0.94350	0.981609	0.884716	0.935587
13	0.3	SMOTE	4.5	0.95290	0.984409	0.931682	0.910921
14	0.3	SMOTETomek	4.5	0.95275	0.984158	0.933875	0.907936
15	0.4	Original	4.5	0.94970	0.985392	0.962504	0.910078
16	0.4	Random Over	4.5	0.95115	0.985690	0.954836	0.921810
17	0.4	Random Under	4.5	0.94835	0.984761	0.931003	0.941140
18	0.4	SMOTE	4.5	0.95135	0.985735	0.947936	0.929791
19	0.4	SMOTETomek	4.5	0.95120	0.985803	0.949409	0.927792
20	0.5	Original	4.5	0.95340	0.986362	0.956066	0.950606
21	0.5	Random Over	4.5	0.95250	0.986199	0.955148	0.949706
22	0.5	Random Under	4.5	0.95195	0.986230	0.955584	0.948105
23	0.5	SMOTE	4.5	0.95250	0.986351	0.955679	0.949107
24	0.5	SMOTETomek	4.5	0.95210	0.985939	0.955921	0.948009
25	0.6	Original	4.5	0.94795	0.983995	0.945097	0.969444
26	0.6	Random Over	4.5	0.95060	0.984462	0.954809	0.963099
27	0.6	Random Under	4.5	0.94615	0.983120	0.964113	0.945232
28	0.6	SMOTE	4.5	0.95255	0.984791	0.959191	0.961679
29	0.6	SMOTETomek	4.5	0.94905	0.984780	0.957037	0.957921
30	0.7	Original	4.5	0.94940	0.983265	0.946970	0.982589
31	0.7	Random Over	4.5	0.95225	0.984199	0.959107	0.973131
32	0.7	Random Under	4.5	0.93470	0.980826	0.974732	0.930493
33	0.7	SMOTE	4.5	0.95390	0.984605	0.967028	0.966895
34	0.7	SMOTETomek	4.5	0.95370	0.984300	0.966647	0.967039
35	0.8	Original	4.5	0.94885	0.976215	0.945261	0.993538
36	0.8	Random Over	4.5	0.95560	0.978427	0.959057	0.986510
37	0.8	Random Under	4.5	0.93035	0.973026	0.982271	0.929347
38	0.8	SMOTE	4.5	0.95850	0.977385	0.971722	0.976346
39	0.8	SMOTETomek	4.5	0.95875	0.977592	0.972215	0.976158
40	0.9	Original	4.5	0.95045	0.961470	0.949724	0.997601
41	0.9	Random Over	4.5	0.95650	0.965975	0.957550	0.995648
42	0.9	Random Under	4.5	0.90580	0.954997	0.987531	0.906306
43	0.9	SMOTE	4.5	0.96345	0.966552	0.976542	0.982813
44	0.9	SMOTETomek	4.5	0.96280	0.966210	0.975843	0.982813

Обсуждение результатов

Результаты исследования показывают, что эффективность методов SMOTE и SMOTETomek для обработки дисбаланса классов остаётся достаточно высокой даже при изменении уровня дисбаланса. Применение данных методов способствует повышению эффективности работы модели. Однако выбор конкретного подхода во многом зависит от условий задачи и метрик качества, которым отдается предпочтение. Так, если требуется высокий уровень метрики Recall, предпочтительным будет использование SMOTE или SMOTETomek. Метод Random Under целесообразно применять, если приоритетом является высокая Precision.

Заключение

В ходе эксперимента были проанализированы различные подходы к обработке дисбаланса классов в зависимости от уровня дисбаланса. Результаты показали, что методы SMOTE и SMOTETomek являются наиболее эффективными для решения проблемы дисбаланса в широком диапазоне его значений. Тем не менее выбор метода обработки дисбаланса должен учитывать специфику задачи и приоритеты в метриках качества модели.

ЛИТЕРАТУРА

1. Агалаков С.А., Березин А.А. Повышение эффективности классификации фенотипов заболевания желудочно-кишечного тракта с помощью методов обработки данных //Математические структуры и моделирование. — 2025. — №. 1 (73). — С. 81–91.
2. Бобоназаров Р.Ч. Проблема дисбаланса классов в задаче противодействия мошенничеству: метрики, семплирование и свёрточные нейронные сети // Безопасность информационных технологий. — 2025. — Т. 32. — №. 2. — С. 102–121.
3. Демидова Л.А., Шаршатов М.А., Шыхыев А.А. Методы решения проблемы дисбаланса классов в задаче бинарной классификации // Электронный научный журнал «ИТ-Стандарт». — 2023. — № 1. — С. 22–33.
4. Иглин С.М., Морейдо В.М., Головнин К.И. Прогнозирование редких гидрологических явлений методами машинного обучения на примере ледовых заторов на реке Печоре //Вестник Московского университета. Серия 5. География. — 2025. — №. 1. — С. 87–97.
5. Константинов А.Ф., Дьяконова Л.П. Сравнительный анализ методов снижения дисбаланса классов при построении моделей машинного обучения в финансовом секторе //Известия Кабардино-Балкарского научного центра РАН. — 2025. — Т. 27. — №. 1. — С. 143–151.
6. Проневич О.Б., Зайцев М.В. Интеллектуальные методы повышения точности прогнозирования редких опасных событий на железнодорожном транспорте //Надежность. — 2021. — Т. 21. — №. 3. — С. 54–64.
7. Рюмина Е.В., Карпов А.А. Сравнительный анализ методов устранения дисбаланса классов эмоций в видеоданных выражений лиц //Научно-технический вестник информационных технологий, механики и оптики. — 2020. — Т. 20. — №. 5. — С. 683–691.
8. Fatih Sağlam, Mehmet Ali Cengiz, MCSMOTE: A transition matrix-driven oversampling technique for class imbalance,Applied Soft Computing, Volume 185, Part B, 2025
9. Rezvani S., Wang X. A broad review on class imbalance learning techniques //Applied Soft Computing. — 2023. — Т. 143. — С. 110415.
10. Scikit-learn: machine learning in Python [Электронный ресурс] / F. Pedregosa, G. Varoquaux, A. Gramfort [и др.] // URL: (дата обращения: 19.10.2025).

Журнал «Современная наука: актуальные проблемы теории и практики»