УДК 004.06:004.08

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ БАЛАНСИРОВКИ ДАННЫХ ДЛЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

Е. Клицунова
Студентка кафедры информационных систем управления БГУ
kateryna.klitsunova@gmail.com

М.М. Лукашевич
Доцент кафедры информационных систем управления БГУ, ст. науч. сотр. лаборатории №222 ОИПИ, кандидат технических наук, доцент
LukashevichMM@bsu.by

Е. Клицунова

Является студенткой четвертого курса Белорусского государственного университета. Область научных интересов связана с решением практических задач в области машинного обучения.

М.М. Лукашевич

Является доцентом кафедры информационных систем управления БГУ. ст. науч. сотр. лаборатории №222 ОИПИ. Область научных интересов связана с разработкой методов и алгоритмов машинного обучения для решения задач анализа данных и компьютерного зрения.

Аннотация. В работе представлены результаты сравнительного анализа методов и алгоритмов работы с несбалансированными данными при построении моделей машинного обучения. Представлены результаты экспериментальных исследований алгоритмов балансировки данных, основанных на увеличении меньшего класса и на уменьшении большего класса с использованием библиотеки imbalanced-learn. По результатам экспериментов оценено влияние изученных методов и алгоритмов на качество моделей, полученных в результате обучения с классическими классификаторами и ансамблевыми алгоритмами.

Ключевые слова: машинное обучение, несбалансированные данные, балансировка данных, oversampling, undersampling.

Введение. В актуальных задачах машинного обучения как во время исследований, так и на реальных практических примерах приходится сталкиваться с несбалансированным набором данных (Imbalanced Dataset). Набор данных называется несбалансированным, когда в нем наблюдается значительная диспропорция между количеством объектов различных классов. При этом соотношение классов может варьироваться от сравнительно небольшого смещения до значительных диспропорций, когда на один объект меньшего класса приходится сотня, тысяча или даже миллион объектов большего класса. Определение наличия дисбаланса классов в наборе данных можно осуществить визуально с помощью гистограммы, которая отражает количество объектов в каждом классе, или путем визуализации на графике при небольшой размерности пространства признаков [1]. Визуализация искусственно сгенерированного набора данных с соотношением объектов разных классов в пропорции 1:100 представлена на рисунке 1.

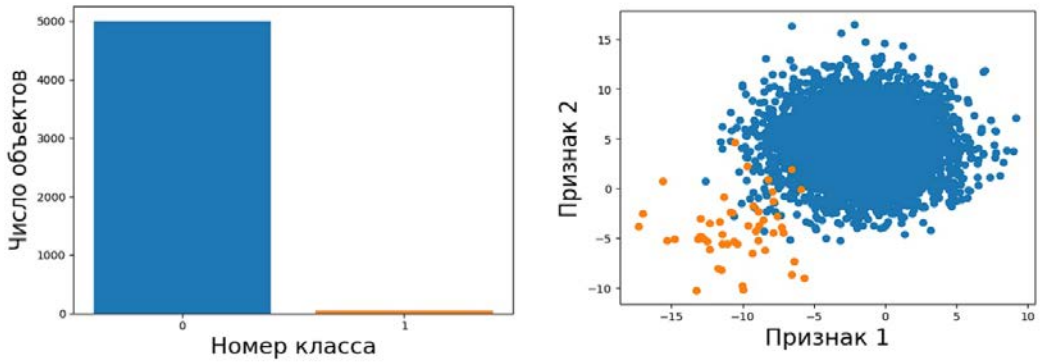

Рисунок 1. Визуализация набора данных с балансом классов 1:100 на гистограмме (слева) и графике (справа)

Наиболее характерна проблема дисбаланса классов для задач классификации. Она может привести к значительным трудностям при обучении моделей машинного обучения, поскольку классические алгоритмы классификации рассчитаны на равное соотношение между классами, и при работе с несбалансированными данными их эффективность значительно снижается [2]. Особенно важным это становится в задачах, где крайне важно получить несмещенные в сторону большего класса результаты — например, не пропустить пациента с серьезным заболеванием или мошеннические действия.

Для сравнительного анализа качества моделей в задачах классификации, как бинарной, так и мультиклассовой, используются стандартные метрики [3 - 5]. Однако некоторые из них применимы только для задачи со сбалансированным набором данных, но при использовании в задачах с несбалансированными данными могут привести к чрезмерно оптимистичным результатам и дать ложное представление об эффективности модели. К примеру, использование доли верно классифицированных объектов (Accuracy) для данных с выраженным дисбалансом может дать высокий уровень точности, приближающийся к 100%, даже если все объекты классифицируются в более крупный класс, а все объекты меньшего класса классифицированы неверно.

Согласно [6, 7] более подходящим является использование для задач классификации с несбалансированным набором данных сбалансированную точность (Balanced Accuracy), определяемую по формулам (1)-(3).

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specifity}}{2} \tag{1}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Specifity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

где TP (True Positive) – количество корректных идентификаций положительного класса;

TN (True Negative) – количество корректных идентификаций отрицательного класса;

FN (False Negative) – количество случаев, когда объектам отрицательного класса были присвоены метки положительного класса;

FP (False Positive) – количество случаев, когда объектам отрицательного класса были присвоены метки положительного класса;

Целью исследования являлось экспериментальное сравнение качества некоторых моделей классификации до и после применения методов балансировки данных в зависимости от размера набора данных и величины дисбаланса классов. Необходимость исследования обусловлена тем, что каждый из методов имеет свои преимущества и недостатки, и их комбинации могут приводить к улучшению качества моделей.

Методы балансировки данных. Согласно [1, 8], методы и алгоритмы работы с несбалансированными данными можно разделить на следующие категории:

1 *Подходы на уровне алгоритма.* Данный тип методов заключается в модификации существующих классических алгоритмов таким образом, чтобы адаптировать их к несбалансированному набору данных и уменьшить смещение в сторону большего класса при обучении. Также сюда можно отнести обучение с учетом издержек классификации, что представляет собой модификацию алгоритмов путем добавления затрат на неправильную классификацию.

2 *Подходы на уровне данных.* Этот подход направлен на уменьшение дисбаланса непосредственно в наборе данных до применения к ним уже существующих алгоритмов классификации.

3 *Использование для обучения ансамблевых алгоритмов.* Такие алгоритмы лучше адаптируются к дисбалансу данных благодаря вычислению среднего результата нескольких классификаторов одного или разных типов.

4 *Комбинированные алгоритмы.* Приведенные выше методы можно комбинировать в разных соотношениях, чтобы получить модель с наилучшим качеством классификации. Также можно комбинировать различные инструменты из одной категории методов, к примеру, использовать разные варианты балансировки набора данных.

Рассмотрим основные методы подхода на уровне данных, а именно – балансировки данных.

Увеличение меньшего класса. Подход с увеличением количества объектов меньшего класса особенно актуален для работы с небольшими наборами данных, где удаление объектов большего класса может существенно сказаться на точности классификации. Выделяют следующие основные алгоритмы:

1 Алгоритм случайного увеличения (Random Oversampling) заключается в дублировании случайно выбранных объектов меньшего класса.

2 Алгоритм SMOTE создает синтетические примеры для меньшего класса, основываясь на существующими экземплярами, что увеличивает количество объектов редкого класса без создания копий существующего.

3 Алгоритм Borderline-SMOTE генерирует синтетические примеры только на границе классов, что улучшает способность модели различать классы, близкие к границе решения.

4 Алгоритм Borderline-SMOTE SVM использует векторные машины (SVM) для определения границ между классами и генерации синтетических данных на этих границах.

5 Алгоритм ADASYN создает синтетические примеры для меньшего класса с учетом их распределения и весов классов.

Уменьшение большего класса. Алгоритмы в рамках данного подхода можно разделить на три группы: ориентированные на выбор объектов, которые необходимо сохранить в выборке, ориентированные на выбор объектов для удаления и комбинированные методы.

1 Алгоритм случайного удаления (Random Undersampling) заключается в дублировании случайно выбранных объектов меньшего класса.

2 Алгоритм CNN направлен на поиск подмножества в выборке, который не приведет к потере качества модели.

3 Алгоритм Near Miss имеет три варианта (NearMiss-1, NearMiss-2 и NearMiss-3), в каждом из которых выбираются объекты большего класса, которые нужно сохранить в выборке, на основании метрики в пространстве признаков. В первом варианте остаются

объекты с минимальным средним расстоянием к трем ближайшим объектам меньшего класса, во втором варианте — к трем наиболее дальним объектам меньшего класса, в третьем варианте — к каждому объекту меньшего класса.

4 Алгоритм Tomek Links находит пары ближайших соседей из разных классов, после чего удаляется объект из большего класса.

5 Алгоритм ENN для каждого объекта вычисляет класс по трем ближайшим соседям, и если объект принадлежит большему классу и идентифицирован неправильно, то он удаляется.

6 Алгоритм OSS комбинирует алгоритмы Tomek Links для удаления шума и объектов большего класса на границе и CNN для удаления объектов, далеких от границы решения.

7 Алгоритм NCR комбинирует алгоритмы CNN для удаления избыточных объектов большего класса и ENN для удаления шума.

Методы увеличения меньшего класса следует применять с осторожностью во избежание переобучения модели. Также не обязательно по результатам увеличения меньшего класса получать именно столько объектов, сколько присутствует в преобладающем классе: для некоторых задач достаточно убрать дисбаланс до умеренного или небольшого. Уменьшение большего класса не во всех случаях приводит к значительному увеличению качества моделей (а иногда может привести и к ухудшению), однако его важно рассмотреть, поскольку алгоритмы из этой категории могут улучшить эффективность в случае комбинации их с методами увеличения меньшего класса.

Постановка эксперимента. Для проведения экспериментов были выбраны публичные наборы данных, представленные на платформе Kaggle. Подбор данных выполнялся по следующим критериям: различная величина наборов данных; различная величина дисбаланса; различная величина признакового пространства; различная тематика наборов данных. По данным критериям были отобраны пять различных наборов данных для следующих задач (таблица 1).

Таблица 1. Наборы данных для проведения экспериментов

Название	Размер набора данных	Тип классификации	Дисбаланс
Классификация здоровья плода [9]	~2 тыс. записей	мультиклассовая	умеренный
Прогнозирование церебрального инсульта [10]	~43 тыс. записей	бинарная	чрезвычайный
Кредитный риск [11]	~32 тыс. записей	бинарная	умеренный
Классификация кредитного рейтинга [12]	100 тыс. записей	мультиклассовая	умеренный
Показатели здоровья при диабете [13]	~254 000 записей	мультиклассовая	чрезвычайный

Соотношение классов в выбранных наборах данных наглядно отображено на графиках, которые приводятся на рисунке 2.

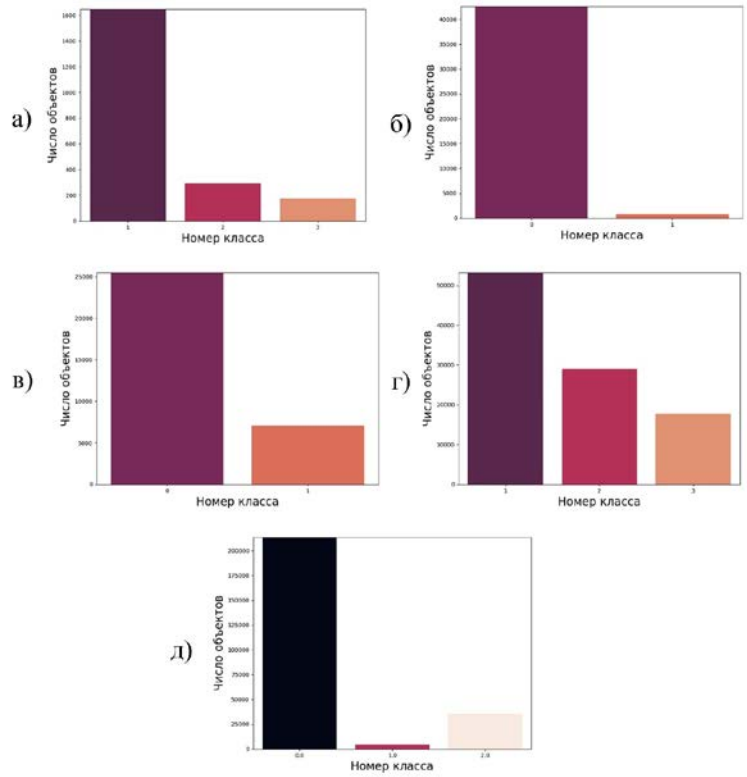

Рисунок 2. Соотношение классов в наборе данных: а) «Классификация здоровья плода» б) «Прогнозирование церебрального инсульта» в) «Кредитный риск» г) «Классификация кредитного рейтинга» д) «Показатели здоровья при диабете»

Для каждого из выбранных наборов данных выполнен разведочный анализ данных, замена пропусков средними значениями или их удаление, преобразование категориальных признаков в числовые и масштабирование данных. Выполнен отбор признаков с удалением на основе корреляционной матрицы и процента влияния на целевую переменную. Каждый набор был разделен на обучающую (70%) и тестовую (30%) выборки.

После выполнения предобработки на каждом выбранном наборе данных проводилось обучение модели с использованием следующих методов классификации:

– решающие деревья (Decision Trees, DT);
– метод k-ближайших соседей (k-Nearest Neighbors, kNN);
– Гауссовский наивный байесовский классификатор (Gaussian Naive Bayes, GaussianNB);
– многослойный перцептрон (Multilayer Perceptron, MLP).

Также было обучение проводилось на ансамблях с гиперпараметрами по умолчанию, а именно:

– случайный лес (Random Forest, RF);
– градиентный бустинг (Gradient Boosting, GB);
– AdaBoost;
– XGBoost;
– LightGBM.

В качестве языка программирования выбран Python благодаря наличию специализированной библиотеки imbalanced-learn [14], которая предоставляет алгоритмы обработки несбалансированных данных. Также использовалась библиотека алгоритмов машинного обучения scikit-learn, а также библиотеки XGBoost и LightGBM для реализации ансамблевых методов машинного обучения. Для всех методов классификации использовались значения параметров, установленные по умолчанию в указанных библиотеках. Для сравнения моделей в качестве метрики использовалась сбалансированная

точность (Balanced Accuracy). Каждая модель была обучена на обучающем наборе данных и выполнена оценка точности на тестовом наборе данных с применением указанной метрики (исходный набор). Далее для обучающего набора данных независимо друг от друга применялись указанные выше техники балансировки данных (увеличение меньшего класса и уменьшение большего класса). В рамках использования подхода по уменьшению большего класса эксперименты были проведены на значительном числе алгоритмов, но далее приведены результаты только некоторых более эффективных (для каждого набора данных перечень алгоритмов разный).

В таблицах 2-11 приводятся результаты экспериментов. «Полужирным» выделены максимальные значения точности при использовании исходного набора данных и различных техник балансировки данных. Голубой заливкой обозначены максимальные значения точности для набора в целом. Золотая заливка использовалась для акцента в тех случаях, когда наблюдалось снижение точности классификации по сравнению с точностью для исходного набора данных.

Таблица 2. Результаты экспериментов для набора данных «Классификация здоровья плода» (увеличение меньшего класса)

Модель	**Исходный набор**	**Random Oversampling**	**SMOTE**	**B-SMOTE**	**B-SMOTE SVM**	**ADASYN**
kNN	0,79	0,81	0,85	0,85	0,85	0,83
DT	0,87	0,88	0,86	0,86	0,86	0,87
MLP	0,67	0,69	0,85	0,80	0,85	0,82
GaussianNB	0,78	0,79	0,79	0,83	0,77	0,84
RF	0,89	**0,93**	**0,93**	**0,93**	0,91	**0,92**
GB	0,91	0,86	0,90	0,89	0,88	0,86
AdaBoost	0,82	0,91	0,92	0,90	0,91	**0,92**
XGBoost	0,89	0,92	0,92	0,92	**0,92**	**0,92**
LightGBM	**0,92**	0,92	0,92	0,92	**0,92**	**0,92**

Таблица 3. Результаты экспериментов для набора данных «Классификация здоровья плода» (уменьшение большего класса)

Модель	**Исходный набор**	**Random Undersampling**	**NM-1**	**NM-3**	**CNN**	**Tomek Links**	**OSS**
kNN	0,79	0,85	0,75	0,71	0,75	0,79	0,83
DT	0,87	0,87	0,85	0,83	0,80	0,84	0,87
MLP	0,67	0,82	0,77	0,76	0,77	0,67	0,84
GaussianNB	0,78	0,79	0,69	0,75	0,77	0,78	0,77
RF	0,89	0,91	0,90	0,88	0,87	0,88	0,91
GB	0,91	**0,93**	**0,92**	**0,90**	**0,90**	0,90	**0,93**
AdaBoost	0,82	0,78	0,70	0,84	0,79	0,81	0,88
XGBoost	0,89	**0,93**	0,91	**0,90**	0,87	0,89	0,94
LightGBM	**0,92**	0,92	0,88	0,87	0,86	**0,93**	**0,93**

Таблица 4. Результаты экспериментов для набора данных «Прогнозирование церебрального инсульта» (увеличение меньшего класса)

Модель	**Исходный набор**	**Random Oversampling**	**SMOTE**	**B-SMOTE**	**B-SMOTE SVM**	**ADASYN**
kNN	0,50	0,54	0,61	0,57	0,56	0,60
DT	0,51	0,52	0,55	0,54	0,54	0,54
MLP	0,50	0,73	0,65	0,66	0,56	0,64
GaussianNB	**0,65**	0,74	**0,72**	**0,72**	**0,71**	**0,72**
RF	0,50	0,51	0,53	0,52	0,53	0,53
GB	0,50	0,77	0,65	0,63	0,60	0,67
AdaBoost	0,50	**0,78**	0,67	0,67	0,62	0,67
XGBoost	0,50	0,60	0,56	0,54	0,53	0,57
LightGBM	0,50	0,66	0,57	0,55	0,53	0,56

Таблица 5. Результаты экспериментов для набора данных «Прогнозирование церебрального инсульта» (уменьшение большего класса)

Модель	**Исходный набор**	**Random Undersampling**	**NM-2**	**NM-3**	**Tomek Links**	**ENN**	**OSS**	**NCR**
kNN	0,50	0,74	0,51	0,65	0,50	0,50	0,50	0,50
DT	0,51	0,69	0,50	0,54	0,52	0,52	0,53	0,53
MLP	0,50	0,75	**0,67**	0,66	0,50	0,50	0,50	0,50
GaussianNB	**0,65**	0,74	0,52	**0,71**	**0,65**	**0,65**	**0,59**	**0,66**
RF	0,50	0,76	0,51	0,66	0,50	0,50	0,50	0,50
GB	0,50	**0,77**	0,50	0,70	0,50	0,50	0,50	0,50
AdaBoost	0,50	**0,77**	0,51	0,63	0,50	0,50	0,50	0,50
XGBoost	0,50	0,74	0,50	0,66	0,50	0,50	0,50	0,51
LightGBM	0,50	0,74	0,50	0,64	0,50	0,50	0,50	0,50

Таблица 6. Результаты экспериментов для набора данных «Кредитный риск» (увеличение меньшего класса)

Модель	**Исходный набор**	**Random Oversampling**	**SMOTE**	**B-SMOTE**	**B-SMOTE SVM**	**ADASYN**
kNN	0,72	0,74	0,73	0,74	0,74	0,73
DT	0,84	0,84	0,81	0,81	0,82	0,80
MLP	0,50	0,72	0,73	0,73	0,72	0,75
GaussianNB	0,67	0,72	0,69	0,68	0,71	0,68
RF	0,85	0,86	0,85	0,84	0,85	0,84
GB	0,84	0,84	0,83	0,84	0,84	0,84
AdaBoost	0,81	0,83	0,79	0,78	0,80	0,78
XGBoost	**0,86**	**0,88**	**0,86**	**0,86**	**0,86**	**0,87**
LightGBM	**0,86**	0,87	**0,86**	**0,86**	**0,86**	0,86

Таблица 7. Результаты экспериментов для набора данных «Кредитный риск» (уменьшение большего класса)

Модель	**Исходный набор**	**Random Undersampling**	**NM-3**	**CNN**	**Tomek Links**	**ENN**	**OSS**	**NCR**
kNN	0,72	0,74	0,70	0,73	0,73	0,75	0,72	0,74
DT	0,84	0,82	0,79	0,78	0,84	0,83	0,83	0,83
MLP	0,50	0,52	0,50	0,50	0,69	0,69	0,50	0,50
GaussianNB	0,67	0,71	0,71	0,71	0,68	0,70	0,69	0,70
RF	0,85	0,84	0,85	0,84	0,85	0,84	0,85	0,84
GB	0,84	0,84	0,84	0,84	0,84	0,84	0,84	0,84
AdaBoost	0,81	0,82	0,82	0,83	0,82	0,82	0,82	0,82
XGBoost	**0,86**	**0,87**	**0,86**	**0,86**	**0,87**	**0,87**	**0,87**	**0,87**
LightGBM	**0,86**	0,86	**0,86**	**0,86**	0,86	0,86	0,86	0,86

Таблица 8. Результаты экспериментов для набора данных «Классификация кредитного рейтинга» (увеличение меньшего класса)

Модель	**Исходный набор**	**Random Oversampling**	**SMOTE**	**B-SMOTE**	**B-SMOTE SVM**	**ADASYN**
kNN	0,71	0,75	0,76	0,76	0,75	0,77
DT	0,73	0,73	0,75	0,74	0,75	0,74
MLP	0,59	0,50	0,53	0,59	0,45	0,67
GaussianNB	0,64	0,63	0,63	0,63	0,63	0,63
RF	**0,80**	**0,82**	**0,82**	**0,82**	**0,82**	**0,82**
GB	0,71	0,72	0,73	0,73	0,73	0,73
AdaBoost	0,63	0,68	0,70	0,69	0,70	0,69
XGBoost	0,76	0,77	0,78	0,78	0,77	0,78
LightGBM	0,74	0,75	0,76	0,76	0,76	0,75

Таблица 9. Результаты экспериментов для набора данных «Классификация кредитного рейтинга» (уменьшение большего класса)

Модель	**Исходный набор**	**Random Undersampling**	**NM-3**	**CNN**	**Tomek Links**	**ENN**	**OSS**	**NCR**
kNN	0,71	0,68	0,58	0,56	0,71	0,67	0,64	0,71
DT	0,73	0,75	0,66	0,64	0,76	0,76	0,72	0,78
MLP	0,59	0,58	0,39	0,50	0,59	0,57	0,65	0,45
GaussianNB	0,64	0,63	0,67	0,65	0,64	0,64	0,64	0,64
RF	**0,80**	**0,83**	**0,78**	**0,76**	**0,82**	**0,78**	**0,78**	**0,81**
GB	0,71	0,73	0,71	0,70	0,72	0,72	0,71	0,73
AdaBoost	0,63	0,72	0,67	0,64	0,68	0,69	0,66	0,70
XGBoost	0,76	0,78	0,74	0,72	0,78	0,77	0,75	0,79
LightGBM	0,74	0,76	0,73	0,72	0,75	0,75	0,74	0,76

Таблица 10. Результаты экспериментов для набора данных «Показатели здоровья при диабете» (увеличение меньшего класса)

Модель	Исходный набор	Random Oversampling	SMOTE	B-SMOTE	B-SMOTE SVM	ADASYN
kNN	0,38	0,38	0,44	0,45	0,44	0,44
DT	0,40	0,39	0,40	0,40	0,39	0,40
MLP	0,39	0,43	0,40	0,44	0,47	0,42
GaussianNB	**0,45**	**0,47**	**0,49**	**0,48**	**0,49**	**0,49**
RF	0,39	0,38	0,39	0,39	0,39	0,39
GB	0,39	0,45	0,45	0,45	0,45	0,44
AdaBoost	0,39	0,45	0,46	0,47	0,46	0,46
XGBoost	0,39	0,44	0,40	0,40	0,40	0,40
LightGBM	0,38	0,45	0,41	0,40	0,41	0,40

Таблица 11. Результаты экспериментов для набора данных «Показатели здоровья при диабете» (уменьшение большего класса)

Модель	Исходный набор	Random Undersampling	CNN	Tomek Links	ENN	OSS	NCR
kNN	0,38	0,43	0,34	0,38	0,35	0,36	0,37
DT	0,40	0,40	0,36	0,40	0,39	0,39	0,39
MLP	0,39	0,43	0,39	0,38	0,37	0,38	0,37
GaussianNB	**0,45**	**0,47**	**0,45**	**0,45**	**0,44**	**0,45**	**0,44**
RF	0,39	0,41	0,38	0,38	0,36	0,38	0,37
GB	0,39	0,42	0,39	0,38	0,37	0,38	0,37
AdaBoost	0,39	0,41	0,39	0,38	0,38	0,38	0,37
XGBoost	0,39	0,42	0,38	0,38	0,38	0,38	0,37
LightGBM	0,38	0,42	0,37	0,38	0,38	0,38	0,37

Обсуждение полученных результатов. Сравнение полученных результатов позволяет сделать следующие выводы. В большинстве случаев классификаторы, обученные на сбалансированном наборе данных, показывали результаты не хуже, чем на исходном наборе. Во всех случаях удаётся достичь повышения точности классификации путём комбинации определенной модели классификатора и техники балансировки данных. Чем меньше степень дисбаланса, тем больше можно подобрать успешных комбинаций модели классификации и техники балансировки данных (таблица 12).

Таблица 12. Обобщение результатов экспериментальных исследований

Набор данных	Исходный набор		Сбалансированный набор	
	Balanced Accuracy	Лучшая модель	Balanced Accuracy	Лучшая модель и техника балансировки
«Классификация здоровья плода»	0,92	XGBoost LightGBM	0,93	RF – Random Oversampling RF – SMOTE RF – B-SMOTE GB – Random Undersampling

				GB – OSS XGBoost – Random Oversampling LightGBM – TomekLinks LightGBM – OSS
«Прогнозирование церебрального инсульта»	0,65	GaussianNB	0,78	AdaBoost – Random Oversampling
«Кредитный риск»	0,86	XGBoost LightGBM	0,88	XGBoost – Random Oversampling
«Классификация кредитного рейтинга»	0,80	RF	0,83	RF – Random Oversampling RF – SMOTE RF – B-SMOTE RF – B-SMOTE SVM RF – ADASYN RF – Random Undersampling
«Показатели здоровья при диабете»	0,45	GaussianNB	0,49	GaussianNB – SMOTE GaussianNB – B-SMOTE SVM GaussianNB – ADASYN

Наиболее противоречивым можно назвать алгоритм уменьшения большей выборки Near Miss. Наиболее чувствительным к нему оказался средний набор данных с крайне выраженным дисбалансом, причем для него вариант NearMiss-1 показал ухудшение сбалансированной точности вплоть до 0,24, а вот вариант NearMiss-3 улучшил сбалансированную точность в среднем на 0,13 — с 0,52 до 0,65. Для остальных же классификаторов и наборов данных применение этого метода дало результаты не лучше или даже существенно хуже, чем при обучении на исходном наборе.

Особенно хорошо показали себя методы балансировки данных на небольшом наборе данных и на наборе данных среднего размера, но с ярко выраженных дисбалансом классов. Для них удалось добиться значительного повышения качества моделей. А вот для больших наборов данных положительный эффект от балансирования класса оказался менее выраженным. При этом использование алгоритмов увеличения меньшего класса дало больший прирост сбалансированной точности, особенно для не-ансамблевых алгоритмов, чем использование алгоритмов уменьшения большего класса. Ансамблевые алгоритмы, как и ожидалось, оказались менее подверженными влиянию балансировки данных, однако и для них удалось добиться небольшого повышения точности. Тем не менее, для больших наборов данных со значительным дисбалансом результат классификации все равно является крайне низким — значение сбалансированной точности близко к 0,5, что не позволяет применять полученную модель в реальной практике.

Заключение. Сравнение результатов работы моделей на классических классификаторах показало улучшение качества моделей после балансировки данных. При этом методы увеличения меньшего класса дали более заметные результаты, а улучшение на наборах данных небольшого (до 10 тыс. записей) и среднего (до 50 тыс. записей) размера с большим дисбалансом в целом оказалось более выраженным. Также отмечено небольшое улучшение качества моделей при балансировке данных перед обучением с ансамблями такими как случайный лес, градиентный бустинг, XGBoost, AdaBoost и LightGBM. Наилучшие показатели сбалансированной точности были получены в результате комбинации балансирования данных и ансамблевых методов для всех наборов данных за исключением большого набора (250 тыс. записей) с чрезвычайным дисбалансом.

Список литературы

[1] Brownlee J. Imbalanced classification with Python: better metrics, balance skewed classes, cost-sensitive learning. Machine Learning Mastery, 2020. — 446 p.

[2] Kumar P., Bhatnagar R., Gaur K. Classification of imbalanced data: review of methods and applications. IOP conference series: materials science and engineering. 2021; 1099(1). DOI: 10.1088/1757-899X/1099/1/012077

[3] Михайличенко А.А. Аналитический обзор методов оценки качества алгоритмов классификации в задачах машинного обучения. Вестник Адыгейского государственного университета. 2022; 311(4). DOI: 10.53598/2410-3225-2022-4-311-52-59

[4] Ferri C., Hernendez-Orallo J., Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognition Letters. 2009; 30(1). DOI: 10.1016/j.patrec.2008.08.010

[5] Sokolova M., Lapalme G. A systematic analysis of performance measures for classification tasks. Information processing & management. 2009; 45(4). DOI: 10.1016/j.ipm.2009.03.002

[6] Старовойтов В.В., Голуб Ю.И. Об оценке результатов классификации несбалансированных данных по матрице ошибок. Информатика. 2021; 18(1). DOI: 10.37661/1816-0301-2021-18-1-61-71

[7] Старовойтов В.В., Голуб Ю.И Как оценивать результаты классификации несбалансированных больших данных? BIG DATA and Advanced Analytics = BIG DATA и анализ высокого уровня. Минск: Бестпринт. 2021 — с. 272–283.

[8] Krawczyk, B. Learning from imbalanced data: open challenges and future directions. Progress in artificial intelligence. 2016; 5(4). DOI: 10.1007/s13748-016-0094-0

[9] Ayres-de-Campos D., Bernardes, J., Garrido A., Marques-de-Sá J., Pereira-Leite L. A program for automated analysis of cardiotocograms. J. Matern. Fetal Med. 2000; 9. DOI: 10.1002/1520-6661(200009/10)9:5<311::AID-MFM12>3.0.CO;2-9

[10] Liu T., Fan W., Wu Ch. Data for A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets. Mendeley Data, 2019. 1. DOI: 10.17632/x8ygrw87jw.1

[11] Credit Risk Dataset [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

[12] Dataset for Credit Score Classification [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/datasets/ayushsharma0812/dataset-for-credit-score-classification

[13] Diabetes Health Indicators Dataset [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

[14] Lemaître G., Nogueira F., Aridas C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning; Journal of machine learning research. 2017; 18(17). DOI: 10.48550/arXiv.1609.06570

COMPARATIVE ANALYSIS OF DATA BALANCING METHODS

K. Klitsunova
Student of the Department of Information Management Systems of BSU

M.M. Lukashevich
Associate Professor of the Department of Information Management Systems of BSU, Senior Researcher of the Laboratory No. 222 of the UIIP PhD of Technical sciences, Associate Professor

Abstract. The paper presents the results of a comparative analysis of methods and algorithms for working with imbalanced data in machine learning models. Experimental results of data balancing algorithms based on increasing a smaller class and decreasing a larger class using the imbalanced-learn library are presented. Based on the results of the experiments, the influence of the studied methods and algorithms on the quality of models obtained as a result of training with classical classifiers and ensemble algorithms was assessed.

Keywords: machine learning, imbalanced data, data balancing, oversampling, undersampling.