# Borderline-SMOTE: Новый Метод Избыточной выборки при изучении несбалансированных наборов данных

**Автор:** Aleksandr Gambarov  
**Дата:** November 2024

## Аннотация

В последние годы обучение на несбалансированных наборах данных привлекает все больше внимания как в теоретическом, так и в практическом плане. В данной статье подчеркивается важность несбалансированных наборов данных и их широкая область применения в области интеллектуального анализа данных, а также приводится обзор метрик оценки и существующих методов для решения проблемы дисбаланса. Синтетическая техника увеличения выборки меньшинства (SMOTE) является одним из методов увеличения выборки, направленных на решение этой проблемы. Основываясь на методе SMOTE, в этой статье представлены два новых метода увеличения выборки меньшинства: Borderline-SMOTE1 и Borderline-SMOTE2, в которых увеличиваются только примеры меньшинства, находящиеся близко к границе раздела классов. Для класса меньшинства эксперименты показывают, что наши подходы достигают лучшего показателя TP (true positive, истинно положительные) и значения F-меры по сравнению с методами SMOTE и случайного увеличения выборки.

## Введение

Существуют два типа дисбаланса в наборе данных. Первый — это межклассовый дисбаланс, при котором одни классы содержат значительно больше примеров, чем другие. Второй — это внутриклассовый дисбаланс, при котором некоторые подмножества одного класса имеют значительно меньше примеров, чем другие подмножества того же класса. По принятой терминологии в несбалансированных наборах данных классы с большим числом примеров называют классами большинства, а классы с меньшим числом примеров — классами меньшинства.

Проблема дисбаланса привлекает все больше внимания в последние годы. Несбалансированные наборы данных встречаются во многих областях реального мира, таких как выявление ненадежных клиентов телекоммуникационных компаний, обнаружение утечек нефти на спутниковых радиолокационных изображениях, обучение произношению слов, классификация текстов, обнаружение мошеннических телефонных звонков, задачи информационного поиска и фильтрации, и многие другие. В этих областях нас в основном интересует класс меньшинства, а не класс большинства, поэтому необходимо достигать достаточно высокой точности предсказания для класса меньшинства. Однако традиционные алгоритмы интеллектуального анализа данных показывают неудовлетворительные результаты на несбалансированных наборах данных, так как их распределение не учитывается при разработке таких алгоритмов.

Структура данной статьи организована следующим образом. В разделе 2 дается краткий обзор недавних разработок в области обучения на несбалансированных наборах данных. Раздел 3 описывает наши методы увеличения выборки для решения проблемы дисбаланса. Раздел 4 представляет эксперименты и сравнивает наши методы с другими методами увеличения выборки. Раздел 5 подводит итоги.

## Недавние разработки в области обучения на несбалансированных наборах данных

### Метрики оценки в условиях дисбаланса

Большинство исследований в условиях дисбаланса данных сосредоточены на задачах с двумя классами, так как многоклассовую задачу можно упростить до задачи с двумя классами. По принятой конвенции метка класса меньшинства обозначается как положительная, а метка класса большинства — как отрицательная. Таблица 1 иллюстрирует матрицу ошибок для задачи с двумя классами. Первая колонка таблицы содержит фактические метки классов примеров, а первая строка показывает их предсказанные метки. TP (true positive) и TN (true negative) обозначают количество положительных и отрицательных примеров, которые классифицированы правильно, в то время как FN (false negative) и FP (false positive) обозначают количество неверно классифицированных положительных и отрицательных примеров соответственно.

**Таблица 1: Матрица ошибок для задачи с двумя классами**

|  | **Предсказанный положительный** | **Предсказанный отрицательный** |
| --- | --- | --- |
| **Положительный** | TP | FN |
| **Отрицательный** | FP | TN |

Метрика точности вычисляется по формуле:

Метрика FP rate:

Метрика TP rate (Recall):

Метрика Precision:

Метрика F-меры:

При использовании для оценки производительности алгоритма в условиях дисбаланса данных метрика точности, как правило, склонна лучше предсказывать класс большинства и показывает неудовлетворительные результаты для класса меньшинства. Это следует из её определения: если набор данных сильно несбалансирован, даже если классификатор правильно классифицирует все примеры класса большинства и ошибочно классифицирует все примеры класса меньшинства, точность классификатора останется высокой из-за преобладания класса большинства. В таких случаях точность не может достоверно отражать предсказание для класса меньшинства. Поэтому требуются более разумные метрики оценки.

Кривая ROC (Receiver Operating Characteristic) — одна из популярных метрик для оценки алгоритмов на несбалансированных наборах данных. Это двумерный график, на котором показатель TP rate отображается по оси y, а FP rate — по оси x. Точка (0,1) является идеальной точкой для классификатора. ROC-кривая показывает относительные соотношения между преимуществами (TP rate) и затратами (FP rate). Площадь под кривой (AUC) также может использоваться для оценки несбалансированных данных. Кроме того, F-мера является популярной метрикой для оценки в задачах с дисбалансом. F-мера высока, когда показатели Recall и Precision высоки, и её можно настроить, изменяя значение (), которое обычно равно 1 и соответствует относительной важности Precision по сравнению с Recall.

Приведенные выше метрики оценки могут обоснованно оценивать алгоритм для несбалансированных данных, так как их формулы ориентированы на класс меньшинства.

### Методы решения проблемы дисбаланса в обучении

Решения проблемы дисбаланса данных можно разделить на методы на уровне данных и методы на уровне алгоритмов. Методы на уровне данных изменяют распределение несбалансированных наборов данных, после чего сбалансированные данные передаются алгоритму для улучшения обнаружения класса меньшинства. Методы на уровне алгоритмов модифицируют существующие алгоритмы интеллектуального анализа данных или предлагают новые алгоритмы для решения проблемы дисбаланса.

#### Методы на уровне данных

На уровне данных были предложены различные формы методов повторной выборки. Самые простые методы повторной выборки — это случайное увеличение выборки и случайное уменьшение выборки. Первый метод увеличивает количество примеров класса меньшинства путем точного дублирования примеров, тогда как второй метод случайным образом удаляет некоторые примеры класса большинства. Однако случайное увеличение выборки может сделать области принятия решений алгоритма более узкими и специфичными, что может привести к переобучению. Случайное уменьшение выборки может удалить полезную информацию из набора данных. Поэтому были разработаны улучшенные методы повторной выборки, такие как эвристические методы, комбинации методов увеличения и уменьшения выборки, а также методы, которые интегрируют повторную выборку непосредственно в алгоритмы интеллектуального анализа данных.

Некоторые из улучшенных методов повторной выборки включают:

Кубат и др. представили эвристический метод уменьшения выборки, который балансирует набор данных за счет удаления шумов и избыточных примеров класса большинства. Нитеш и соавторы предложили метод SMOTE (Synthetic Minority Over-sampling Technique), который генерирует новые синтетические примеры вдоль линии между примерами класса меньшинства и их выбранными ближайшими соседями. Преимущество SMOTE заключается в том, что он делает области принятия решений более крупными и менее специфичными. Нитеш и соавторы также интегрировали SMOTE в стандартную процедуру бустинга, что улучшило предсказание для класса меньшинства, не ухудшая точность на всем наборе данных. Густаво и соавторы комбинировали методы увеличения и уменьшения выборки для решения проблемы дисбаланса. Эстабрукс и др. предложили метод множественной повторной выборки, который адаптивно выбирает наиболее подходящий уровень повторной выборки. Тэхо Джо и Натали Джапкович предложили метод увеличения выборки на основе кластеров, который решает как межклассовый, так и внутриклассовый дисбаланс. Хонгью Гуо и Херна Виктор нашли сложные примеры классов большинства и меньшинства в процессе бустинга, а затем сгенерировали новые синтетические примеры из сложных примеров и добавили их в набор данных.

#### Методы на уровне алгоритмов

Методы на уровне алгоритмов работают с самими алгоритмами, а не с набором данных. Стандартный алгоритм бустинга, например Adaboost, увеличивает веса ошибочно классифицированных примеров и уменьшает веса правильно классифицированных с использованием той же пропорции, не учитывая дисбаланс набора данных. Таким образом, традиционные алгоритмы бустинга не показывают хороших результатов на классе меньшинства. Чтобы решить этот недостаток, Джоши и соавторы предложили улучшенный алгоритм бустинга, который обновляет веса положительных и отрицательных предсказаний по-разному. Новый алгоритм может достичь лучшего предсказания для класса меньшинства. При обработке несбалансированных наборов данных граница классов, определенная методом опорных векторов (SVM), склонна смещаться в сторону класса меньшинства, что увеличивает количество ошибок для этого класса. Ганг Ву и Эдвард Чанг предложили алгоритм выравнивания границы классов, который модифицирует границу классов, изменяя ядро функции SVM. Кайчжу Хуан и соавторы представили смещенную минимаксную вероятностную машину (Biased Minimax Probability Machine, BMPM) для решения проблемы дисбаланса. Опираясь на надежные средние и ковариационные матрицы классов большинства и меньшинства, BMPM может вычислить гиперплоскость принятия решений путем корректировки нижней границы точности для тестового набора. Кроме того, существуют и другие эффективные методы, такие как обучение с учетом затрат, корректировка вероятности предсказаний, обучение с одним классом и так далее.

## Новый метод увеличения выборки: Borderline-SMOTE

Большинство алгоритмов классификации стремятся максимально точно изучить границу каждого класса в процессе обучения, чтобы улучшить предсказание. Примеры, расположенные на границе классов и рядом с ней (в данной статье они называются пограничными примерами), имеют большую вероятность быть классифицированными неверно по сравнению с примерами, находящимися далеко от границы. Поэтому они более важны для задачи классификации.

На основе этого анализа можно предположить, что примеры, находящиеся далеко от границы, вносят незначительный вклад в процесс классификации. В связи с этим в данной работе предложены два новых метода увеличения выборки для класса меньшинства: Borderline-SMOTE1 и Borderline-SMOTE2, в которых увеличиваются только пограничные примеры класса меньшинства. Наши методы отличаются от существующих методов увеличения выборки, при которых увеличиваются либо все примеры класса меньшинства, либо случайное подмножество примеров этого класса.

Наши методы основаны на алгоритме SMOTE (Synthetic Minority Over-sampling Technique). SMOTE генерирует синтетические примеры для увеличения выборки класса меньшинства. Для каждого примера из класса меньшинства вычисляются его (k) ближайших соседей того же класса (в SMOTE значение (k) обычно равно 5). Из них случайным образом выбираются несколько примеров в соответствии с коэффициентом увеличения выборки. Затем новые синтетические примеры создаются вдоль линии между оригинальным примером и его выбранными ближайшими соседями. В отличие от существующих методов увеличения выборки, наши методы увеличивают или усиливают только пограничные примеры класса меньшинства. Сначала определяются пограничные примеры, после чего на их основе создаются синтетические примеры, которые добавляются к исходному обучающему набору.

Пусть ( T ) — обучающий набор данных, ( P ) — класс меньшинства, а ( N ) — класс большинства, и:

где () и () обозначают количество примеров классов меньшинства и большинства соответственно.

### Подробная процедура Borderline-SMOTE1:

1. **Поиск ближайших соседей**: Для каждого примера () из класса меньшинства вычисляем его ближайших соседей из всего обучающего набора . Обозначим количество примеров класса большинства среди этих ближайших соседей как где .
2. **Определение типов примеров**: Если (то есть все ближайших соседей относятся к классу большинства), то считается шумовым и не участвует в дальнейших шагах. Если , то есть количество ближайших соседей из класса большинства больше, чем из класса меньшинства, считается легко ошибочно классифицируемым и помещается в множество DANGER. Если , считается безопасным и не участвует в следующих шагах.
3. **Выбор пограничных примеров**: Примеры в множестве DANGER являются пограничными для класса меньшинства , поэтому . Обозначим:

где .

1. **Генерация синтетических примеров**: Для каждого примера в DANGER вычисляем его ближайших соседей из класса меньшинства . Затем создаем синтетических положительных примеров, где — целое число от 1 до . Для каждого примера из DANGER случайным образом выбираем ближайших соседей из его ближайших соседей в . Затем вычисляем разности () между и каждым из его ближайших соседей. Умножаем каждую из этих разностей на случайное число (), чтобы создать новые синтетические примеры между и его ближайшими соседями:

Мы повторяем эту процедуру для каждого в DANGER, получая синтетических примеров. Этот шаг аналогичен методу SMOTE. Для получения подробностей см. оригинал SMOTE.

### Метод Borderline-SMOTE2

Метод Borderline-SMOTE2 похож на Borderline-SMOTE1, но с добавлением новой схемы для создания синтетических примеров. В Borderline-SMOTE2 синтетические примеры создаются не только на основе каждого пограничного примера из класса меньшинства и его ближайших соседей из этого же класса, но и на основе его ближайшего соседа из класса большинства.

**Процедура Borderline-SMOTE2:**

1. **Определение пограничных примеров**: Процесс выбора пограничных примеров такой же, как в Borderline-SMOTE1. Мы выделяем примеры из класса меньшинства, находящиеся в зоне DANGER, то есть такие, у которых больше половины ближайших соседей относятся к классу большинства. Эти примеры считаются пограничными для класса меньшинства.
2. **Генерация синтетических примеров**: Для каждого пограничного примера из зоны DANGER генерируются синтетические примеры двумя способами:
   * **С использованием ближайших соседей из класса меньшинства**: Как и в Borderline-SMOTE1, случайным образом выбираются ближайших соседей из ближайших соседей класса меньшинства для каждого примера в DANGER. На основе этих соседей создаются новые синтетические примеры по той же формуле:
   * $ \_j = p\_i’ + r\_j \_j, j = 1, 2, , s $
   * где — случайное число между 0 и 1.
   * **С использованием ближайшего соседа из класса большинства**: Для каждого примера в DANGER определяется его ближайший сосед из класса большинства. Разность между пограничным примером из класса меньшинства и его ближайшим соседом из класса большинства умножается на случайное число в диапазоне от 0 до 0.5, чтобы новые синтетические примеры находились ближе к классу меньшинства, но в направлении границы:
   * $ *{} = p\_i’ + 0.5 r* {} $
   * где — разность между пограничным примером меньшинства и его ближайшим соседом из класса большинства, а — случайное число в диапазоне [0, 0.5].

Таким образом, Borderline-SMOTE2 создает синтетические примеры на основе как ближайших соседей из класса меньшинства, так и ближайшего соседа из класса большинства. Это позволяет еще сильнее усилить пограничные примеры, но может также привести к некоторому перекрытию между классами, что влияет на точность.

## Эксперименты

В наших экспериментах для оценки результатов использовались показатели TP (доля истинно положительных классификаций) и F-мера для класса меньшинства. Показатель TP отражает точность для класса меньшинства, а значение в формуле F-меры установлено равным 1.

Использовались четыре набора данных, представленные в таблице 2. Среди них набор данных Circle — это смоделированный нами набор, изображенный на рисунке 1, а остальные взяты из репозитория UCI. Все атрибуты в наборах данных являются количественными. В наборе Satimage класс с меткой “4” рассматривается как класс меньшинства, а остальные метки объединены в класс большинства, так как в данной статье рассматривается только задача с двумя классами.

**Таблица 2: Описание наборов данных**

| Название набора данных | Число примеров | Число атрибутов | Метки классов (меньшинство : большинство) | Доля класса меньшинства |
| --- | --- | --- | --- | --- |
| Circle (симуляция) | 1600 | 2 | 1 : 0 | 6,25% |
| Pima (UCI) | 768 | 8 | 1 : 0 | 34,77% |
| Satimage (UCI) | 6435 | 36 | 4 : остальные | 9,73% |
| Haberman (UCI) | 306 | 3 | Умер : Выжил | 26,47% |

В наших экспериментах использовались четыре метода увеличения выборки: SMOTE, случайное увеличение выборки и два предложенных нами метода — Borderline-SMOTE1 и Borderline-SMOTE2. Случайное увеличение выборки увеличивает класс меньшинства путем полного или частичного дублирования положительных примеров. За счет увеличения числа примеров класса меньшинства методы увеличения выборки позволяют сбалансировать распределение в наборе данных и улучшить обнаружение класса меньшинства.

Чтобы удобно сравнить результаты, значение в наших методах установлено таким образом, чтобы количество примеров класса меньшинства в зоне DANGER составляло примерно половину всех примеров класса меньшинства. Значение установлено равным 5, как в SMOTE. Для каждого метода показатели TP и F-меры получены с помощью 10-кратной перекрестной проверки. Чтобы снизить случайные колебания, для методов SMOTE и наших методов TP и F-меры усредняются по трем независимым 10-кратным перекрестным проверкам. После увеличения исходных обучающих наборов с помощью указанных методов валидация проводилась с использованием классификатора C4.5.

Поскольку задача дисбаланса заключается в улучшении точности предсказаний для класса меньшинства, мы приводим результаты только для этого класса. Сравнение результатов проводилось по показателям TP и F-меры для класса меньшинства. TP отражает производительность классификатора на классе меньшинства тестового набора, тогда как F-мера демонстрирует производительность классификатора на всем тестовом наборе.

На рисунке 2 показаны результаты экспериментов. Графики (a), (b), (c) и (d) отображают F-меру и TP для класса меньшинства при применении четырех методов увеличения выборки на наборах данных Circle, Pima, Satimage и Haberman соответственно. По оси x в каждом графике отложено количество новых синтетических примеров. На графиках также показаны значения F-меры и TP для исходных наборов данных с использованием классификатора C4.5.

Результаты, представленные на рисунке 2, показывают следующее:

1. Все четыре метода увеличения выборки улучшают показатель TP для класса меньшинства. Для наборов Circle, Pima и Haberman показатели TP у наших методов лучше, чем у SMOTE и случайного увеличения выборки. В сравнении с исходными наборами данных, максимальное улучшение TP для Borderline-SMOTE1 и Borderline-SMOTE2 составило 20% и 22% на наборе Circle, 21,3% и 20,5% на Pima, 10,1% и 10,0% на Satimage, и по 45,2% на Haberman. Для Satimage показатели TP у наших методов выше, чем у случайного увеличения выборки, и сопоставимы с SMOTE.
2. F-мера для Borderline-SMOTE1 обычно выше, чем для SMOTE и случайного увеличения выборки, а F-мера для Borderline-SMOTE2 также сопоставима с другими методами. В сравнении с исходными наборами данных, наибольшее улучшение F-меры для Borderline-SMOTE1 и Borderline-SMOTE2 составило 12,1% и 10,3% на Circle, 2,3% и 1,3% на Pima, 2,3% и 1,4% на Satimage, и 24,7% и 23,0% на Haberman.

В целом Borderline-SMOTE1 показывает отличные результаты как по TP, так и по F-мере, а Borderline-SMOTE2 достигает высоких значений TP, так как генерирует синтетические примеры не только из пограничных примеров класса меньшинства, но и с использованием их ближайших соседей из класса большинства. Тем не менее, это может привести к перекрытию между классами и, соответственно, к небольшому снижению F-меры.

## Заключение

В последние годы обучение на несбалансированных наборах данных привлекает всё больше внимания как в теоретической, так и в практической областях. Однако традиционные методы интеллектуального анализа данных показывают неудовлетворительные результаты в условиях дисбаланса данных. В данной статье предложены два новых метода синтетического увеличения выборки для класса меньшинства — Borderline-SMOTE1 и Borderline-SMOTE2. Мы сравнили показатели TP и F-меры наших методов с методами SMOTE, случайного увеличения выборки и оригинальной версии C4.5 на четырёх наборах данных.

Пограничные примеры класса меньшинства чаще ошибочно классифицируются, чем те, что находятся дальше от границы. Наши методы увеличивают выборку только пограничных примеров класса меньшинства, в отличие от SMOTE и случайного увеличения выборки, которые увеличивают либо все примеры из класса меньшинства, либо случайное подмножество. Эксперименты показывают, что наши методы обеспечивают лучшие результаты, подтверждая их эффективность.

Существует несколько направлений для дальнейших исследований в этой области. Разработка различных стратегий для определения пограничных примеров и автоматическое адаптивное определение количества таких примеров будут полезны. Стоит также рассмотреть комбинацию наших методов с методами уменьшения выборки и их интеграцию с некоторыми алгоритмами интеллектуального анализа данных.

## Сведения об авторе



Фотография автора

**Александр Гамбаров** — магистрант 2-го года обучения кафедры Корпоративных Информационных Систем МИРЭА – Российского Технологического Университета.