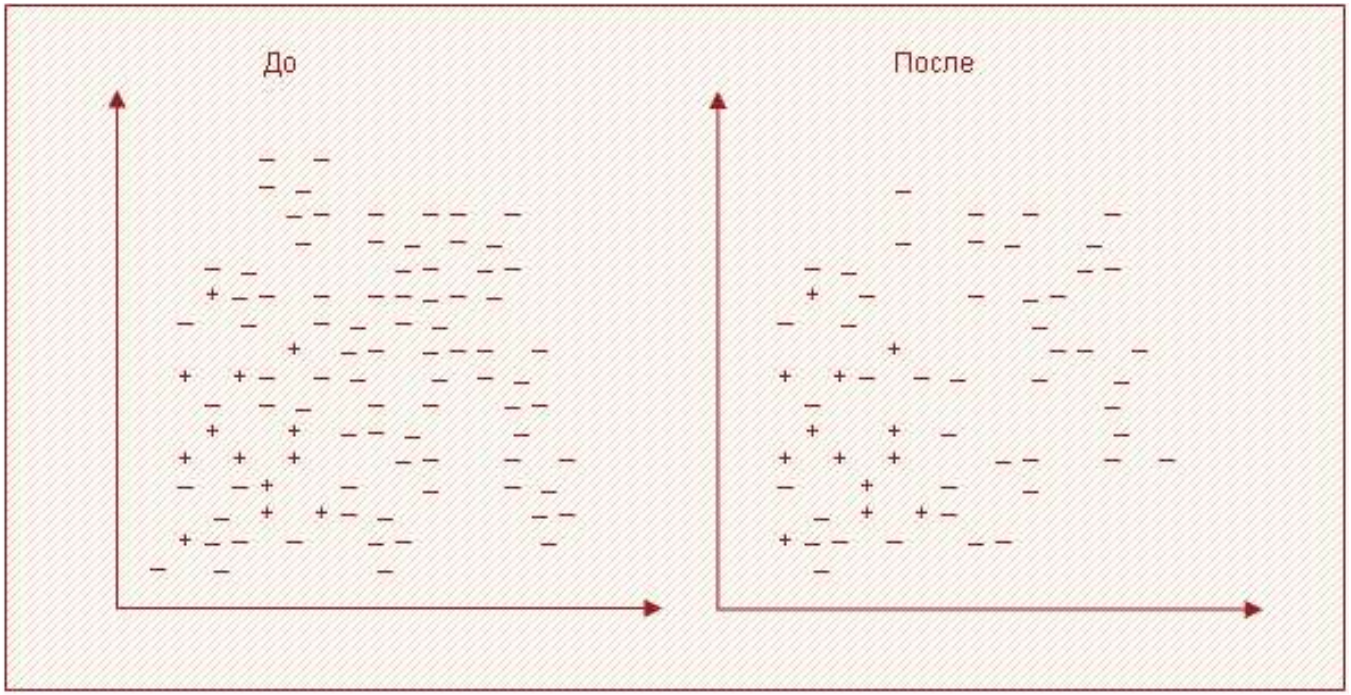
Oversampling / Undersampling

Random Undersampling



Источник: URL

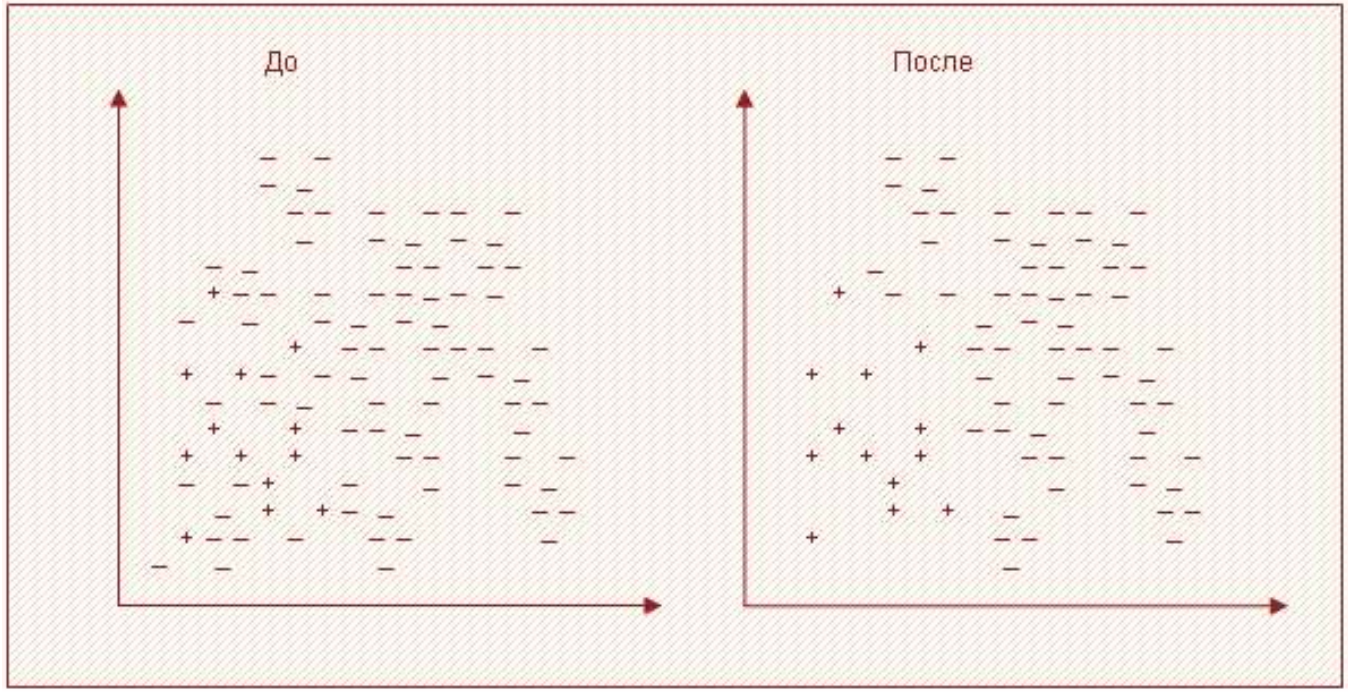
Tomek Links

Пусть примеры E_i и E_j принадлежат к различным классам, $d(E_i, E_j)$ – расстояние между указанными примерами. Пара (E_i, E_j) называется связью Томека, если не найдется ни одного примера E_l такого, что будет справедлива совокупность неравенств:

$$\begin{cases} d(E_i, E_l) < d(E_i, E_j) \\ d(E_j, E_l) < d(E_i, E_j)' \end{cases}$$

Согласно данному подходу, все мажоритарные записи, входящие в связи Томека, должны быть удалены из набора данных. Этот способ хорошо удаляет записи, которые можно рассматривать в качестве «зашумляющих»

Tomek Links

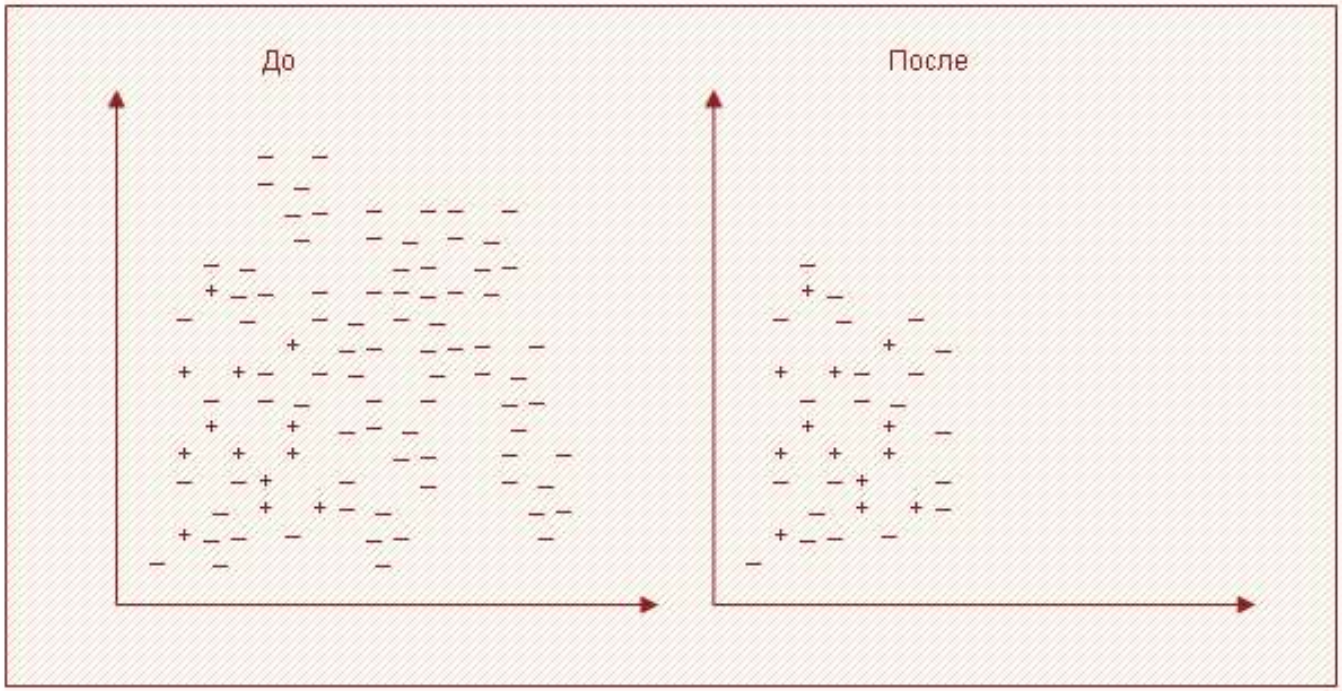


Condensed Nearest Neighbor

Пусть L – исходный набор данных. Из него выбираются все миноритарные примеры и (случайным образом) один мажоритарный. Обозначим это множество как S. Все примеры из L классифицируются по правилу одного ближайшего соседа (1-NN). Записи, получившие ошибочную метку, добавляются во множество S

Таким образом, мы будем учить классификатор находить отличие между похожими примерами, но принадлежащими к разным классам.

Condensed Nearest Neighbor



One-side sampling

Главная идея этой стратегии – это последовательное сочетание предыдущих двух, рассмотренных выше.

Для этого на первом шаге применяется правило сосредоточенного ближайшего соседа, а на втором – удаляются все мажоритарные примеры, участвующие в связях Томека.

Таким образом, удаляются большие «сгустки» мажоритарных примеров, а затем область пространства со скоплением миноритарных очищается от потенциальных шумовых эффектов.

Neighborhood cleaning rule

Эта стратегия также направлена на то, чтобы удалить те примеры, которые негативно влияют на исход классификации миноритарных наблюдений.

Для этого все примеры классифицируются по правилу трех ближайших соседей. Удаляются следующие мажоритарные примеры:

- получившие верную метку класса;
- являющиеся соседями миноритарных примеров, которые были неверно классифицированы.

Oversampling

Самый простой метод – это дублирование примеров миноритарного класса. В зависимости от того, какое соотношение классов необходимо, выбирается количество случайных записей для дублирования.

SMOTE

- Synthetic Minority Oversampling Technique

Алгоритм:

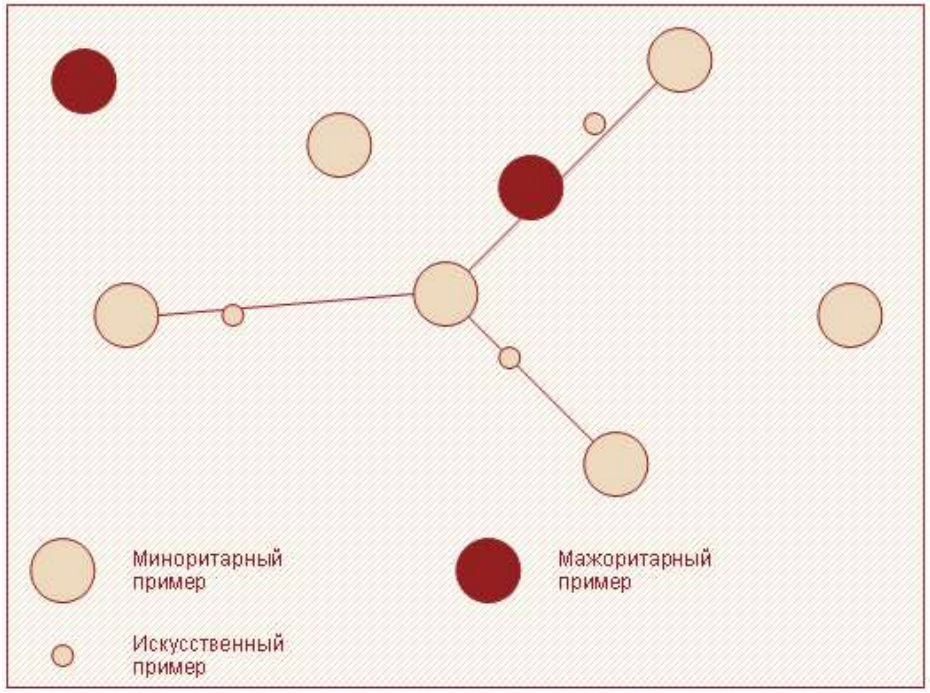
- Для создания новой записи находят разность $\mathrm{d} = X_b - X_a$, где X_a, X_b– векторы признаков «соседних» примеров a и b из миноритарного класса.
- Их находят, используя алгоритм ближайшего соседа (KNN). В данном случае необходимо и достаточно для примера b получить набор из k соседей, из которого в дальнейшем будет выбрана запись b. Остальные шаги алгоритма KNN не требуются.

SMOTE

- Далее из d путем умножения каждого его элемента на случайное число в интервале $(0, 1)$ получают $\hat{d}$. Вектор признаков нового примера вычисляется путем сложения X_a и $\hat{d}$.

Алгоритм SMOTE позволяет задавать количество записей, которое необходимо искусственно сгенерировать. Степень сходства примеров a и b можно регулировать путем изменения значения k (числа ближайших соседей).

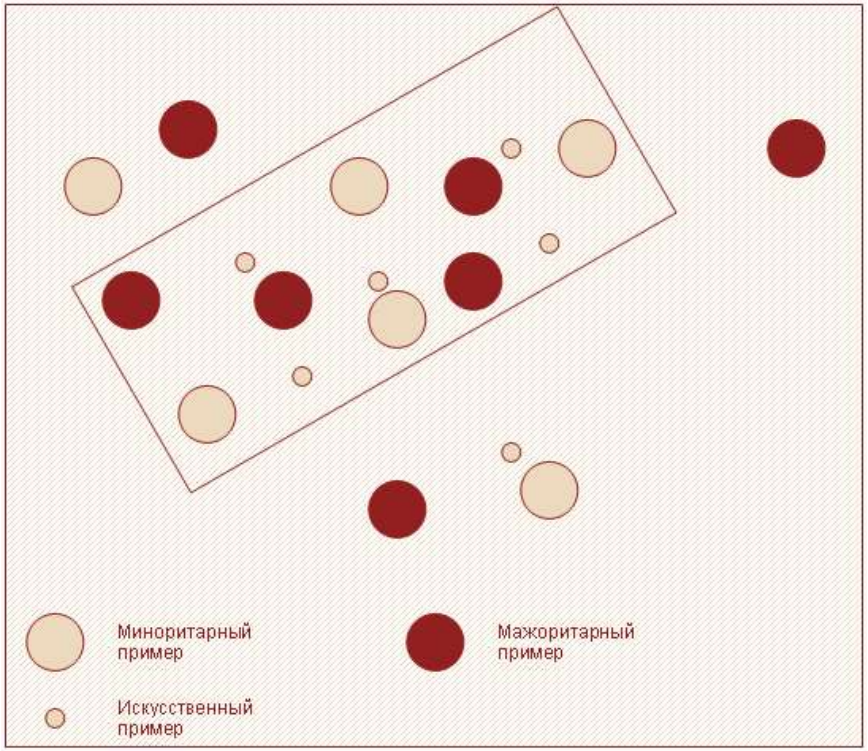
SMOTE


Миноритарный пример
Мажоритарный пример
Искусственный пример

SMOTE

- Данный подход имеет недостаток в том, что «вслепую» увеличивает плотность примерами в области слабо представленного класса
- В случае, если миноритарные примеры равномерно распределены среди мажоритарных и имеют низкую плотность, алгоритм SMOTE только сильнее перемешает классы.

SMOTE


Миноритарный пример
Мажоритарный пример
Искусственный пример

ASMO

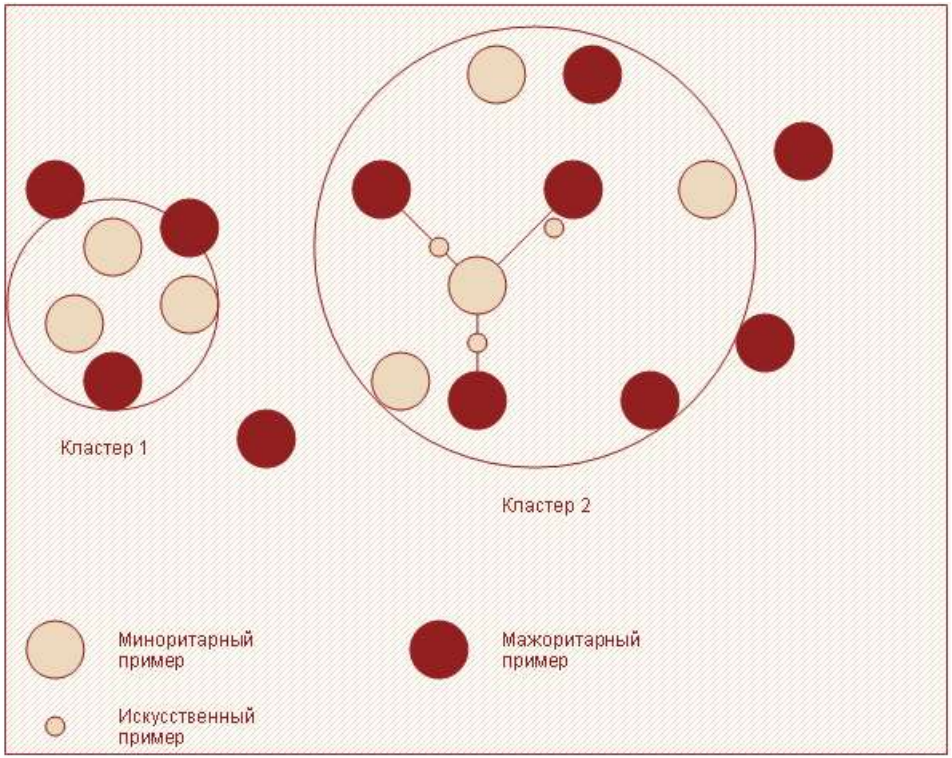
- Adaptive Synthetic Minority Oversampling
- Если для каждого i-ого примера миноритарного класса из k ближайших соседей g ($g \leq k$) принадлежит к мажоритарному, то набор данных считается «рассеянным».
- В этом случае используют алгоритм ASMO, иначе применяют SMOTE (как правило, g задают равным 20).

ASMO

Алгоритм:

1. Используя только примеры миноритарного класса, выделить несколько кластеров (например, алгоритмом k-means).
2. Сгенерировать искусственные записи в пределах отдельных кластеров на основе всех классов. Для каждого примера миноритарного класса находят m ближайших соседей, и на основе них (также как в SMOTE) создаются новые записи.

ASMO


Кластер 1
Кластер 2
Миноритарный пример
Мажоритарный пример
Искусственный пример