Imblearn

A look Beyond Over Sampling

rocketmat

Leandro Ferreira

/ferreiramleandro

@leozimmelo

/leandromferreira

rocketmat


we are
reshaping
the way HR selects candidates
using artificial intelligence
rocketmat

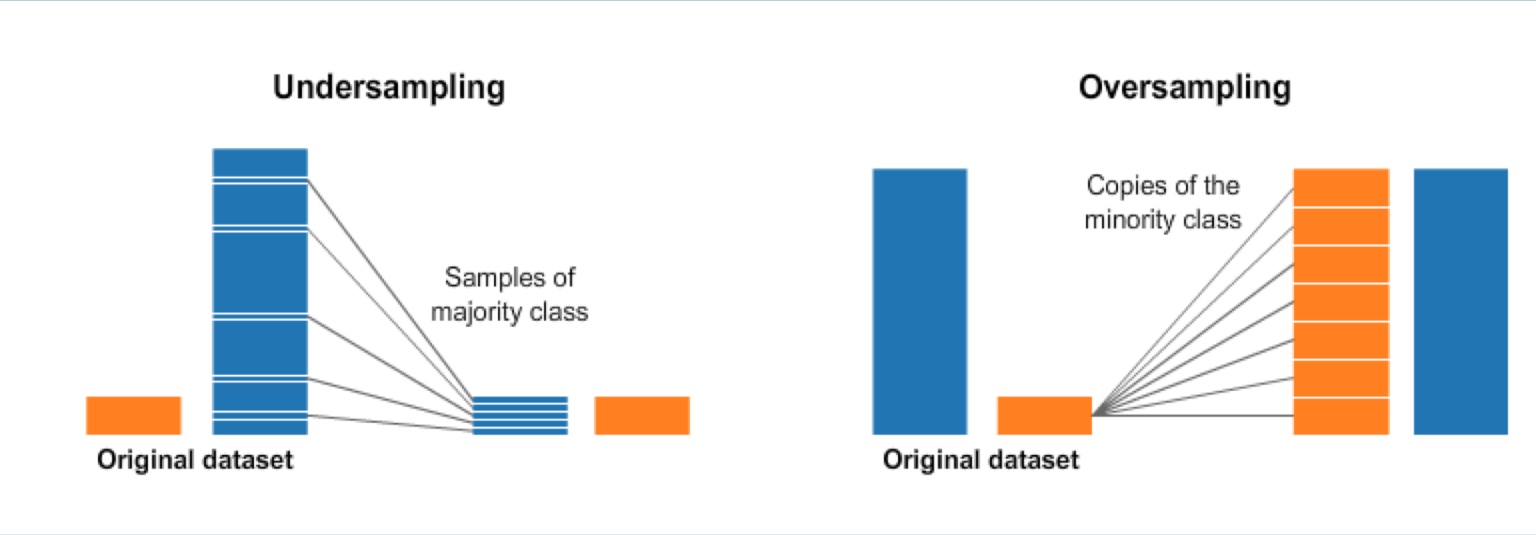
imblearn

- ❏ Version: 0.4.3
- ❏ Started at 2014
- ❏ Compatible with scikit-learn

Docs: https://imbalanced-learn.readthedocs.io/

rocketmat [🚀]

Resampling



Source: Resampling Strategies for Imbalanced Datasets

rocketmat

Over VS Under Sampling

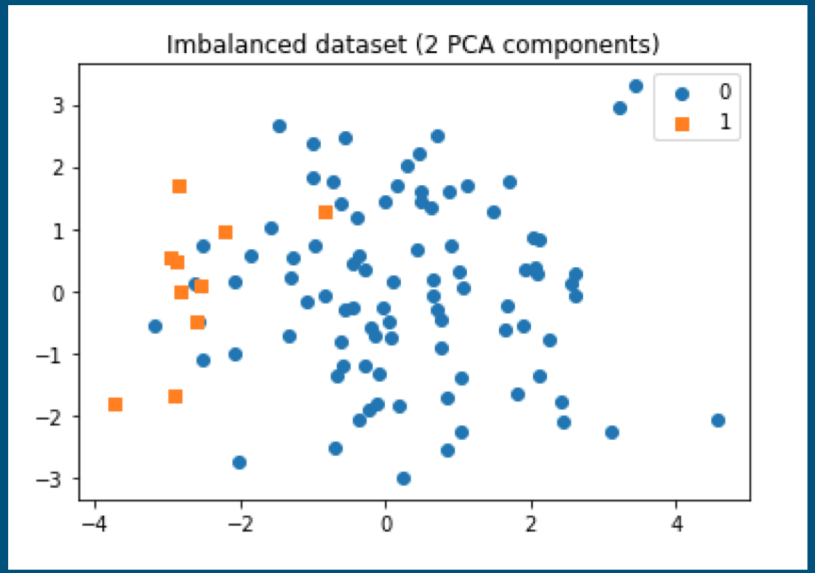
Can cause Overfitting

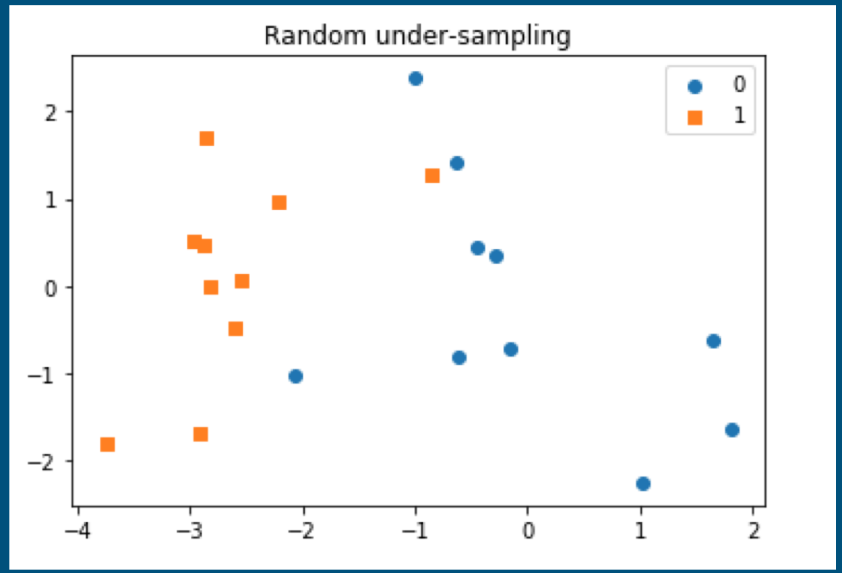
Discart useful information

Over Sampling Techniques

- ❏ Random Over Sampling
- ❏ SMOTE (Synthetic Minority Over Sampling technique)
- ❏ ADASYN (Adaptive Synthetic Sampling Approach)

Random Under Sampling





Near Miss

1: Retain points for the majority class whose mean distance to the k nearest point in minority class is lowest

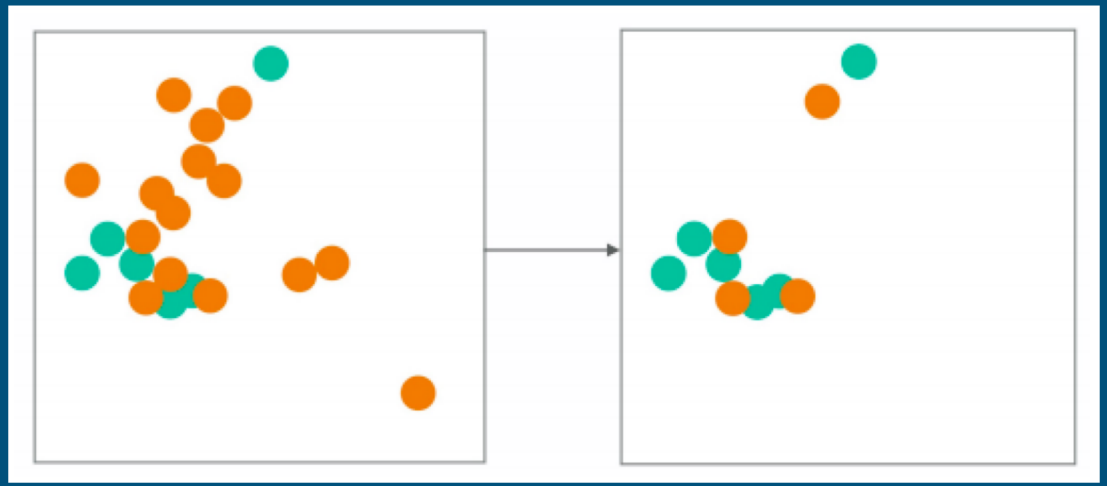
rocketmat

Near Miss

2: Keep points of the majority class whose mean distance to the k farthest points in the minority class

Near Miss

3: Select K nearest neighbors in the majority class for every point in the minority class



rocketmat

Condensed NN

rocketmat [🚀]

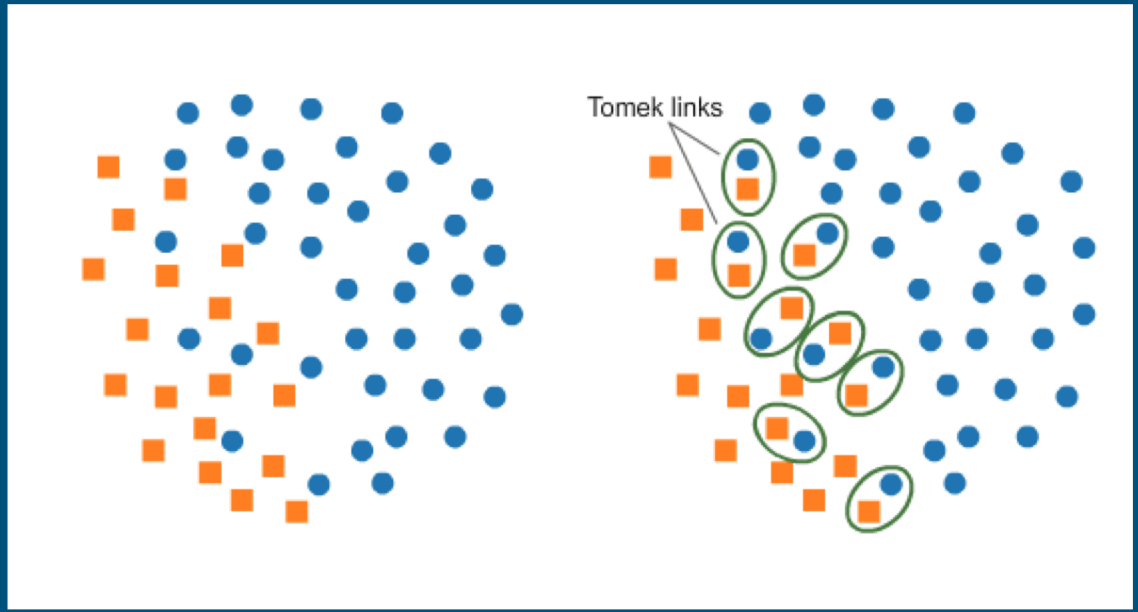
Edited NN

rocketmat [🚀]

Repeated ENN

rocketmat [🚀]

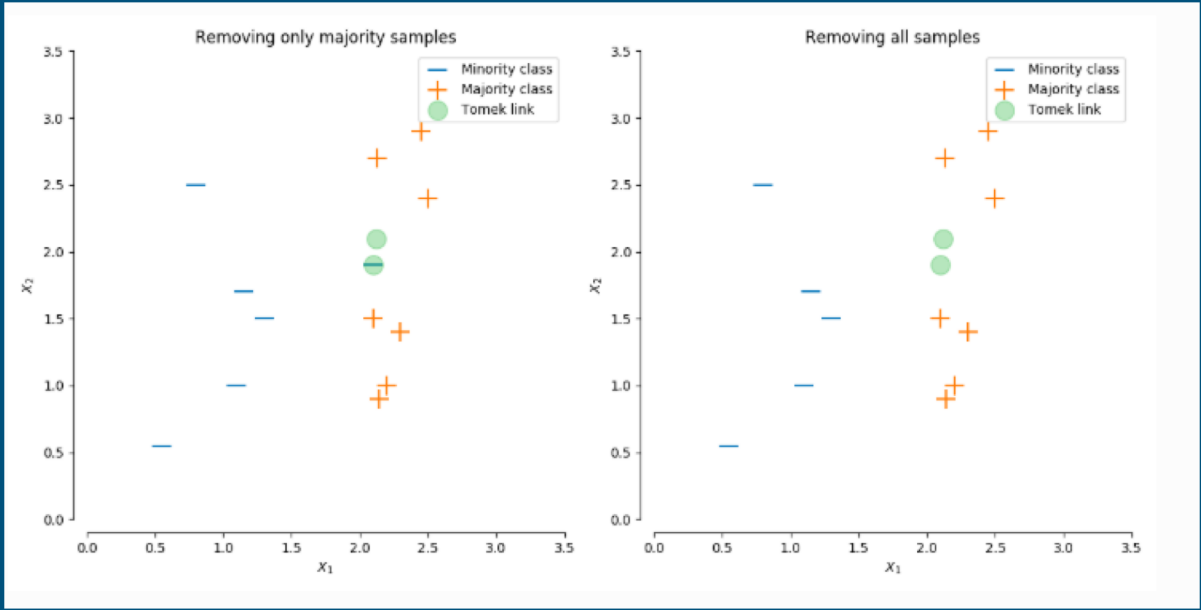
Tomek Links



Source: Resampling Strategies for Imbalanced Datasets

rocketmat

Tomek Links



rocketmat

Easy Ensemble

1. For $i = 1, ..., N$:
   (a) Randomly sample a subset L_i of L such that $|L_i| = |S|$.
   (b) Learn an AdaBoost ensemble using L_i and S

$$F_i(x) = sgn(\Sigma_{j=1}^{n_i} w_{ij} f_{ij}(x) - b_i)$$

2. Combine the above classifiers into a meta-ensemble

$$F(x) = sgn(\Sigma_{i=1}^{N}(\Sigma_{j=1}^{n_i} w_{ij} f_{ij}(x) - b_i))$$

Balance Cascade

1. Set $t = r^{\frac{1}{N-1}}$
2. For $i = 1, ..., N$:
   (a) Randomly sample a subset L_i of L such that $|L_i| = |S|$.
   (b) Learn an AdaBoost ensemble using L_i and S

$$F_i(x) = sgn(\Sigma_{j=1}^{n_i} w_{ij} f_{ij}(x) - b_i)$$

   (c) Tune b_i such that the false positive rate for F_i is t.
3. Undersample L to remove points correctly classified by F_i.
4. Combine the above classifiers into a meta-ensemble

$$F(x) = sgn(\Sigma_{i=1}^{N}(\Sigma_{j=1}^{n_i} w_{ij} f_{ij}(x) - b_i))$$

Ensemble Methods

- ❑ AllKNN
- ❑ Instance Hardness Threshold
- ❑ Neighbourhood Cleaning Rule
- ❑ One Sided Selection

Ensemble Methods Classifier

- ❑ Balanced Random Forest Classifier
- ❑ Balanced Bagging Classifier
- ❑ Easy Ensemble Classifier
- ❑ RUS Boost Classifier

Combine

- SMOTE+ENN
- SMOTE+TOMEK

References and helpful links

[Article] Survey of resampling techniques for improving classification performance in unbalanced datasets

[PyData Talk] Ajinkya More | Resampling techniques and other strategies

[Article] Exploratory Undersampling for Class-Imbalance Learning

[PyData Talk] Imbalanced Data, Mehrdad Yazdan

[Article] Evaluation Measures for Models Assessment over Imbalanced Data Sets