Recap

29 April 2024 13:55

Imbalan~d Sess→2

[Imbalanced data] → classification → binary (1000)

1000 pts
3 classes
A, B, C

0 | 1
95% | 5%
[950] | [50] → imbalanced
0 | 1

[A = 800
B = 100
C = 100]

model

What is the problem

↳ ml algorithms
↳ data → imbalanced
↳ 0 (instances) 950 } bias
↳ 1 (instance) 50

↳ metrics
↳ [accuracy]

misleading accuracy – 99%

1000
990 No | 10 Yes

ml model

No ← function [990 correct, 10 wrong]

Precision / Recall } confusion matrix } FP TP FN TN

What is solution

- Undersampling
  - Random undersampling
- Oversampling
- SMOTE
- Cost sensitive learning
  - loss function
- Class weighting
- Threshold balance

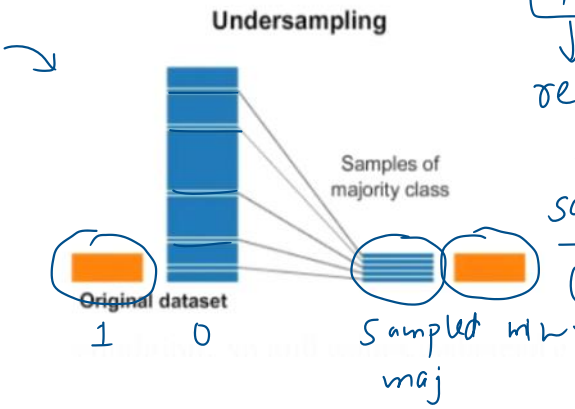
- Random undersampling

learning

loss function.

Random Undersampling

29 April 2024 13:56



imbalanced data

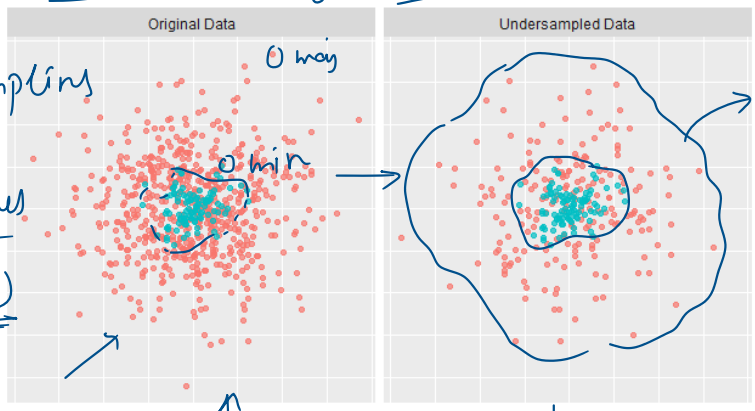
maj(0) min(1) → maj = min

900 100

800 data points

resampling

sample (100)



code → imblearn

balanced

computation

Disadv

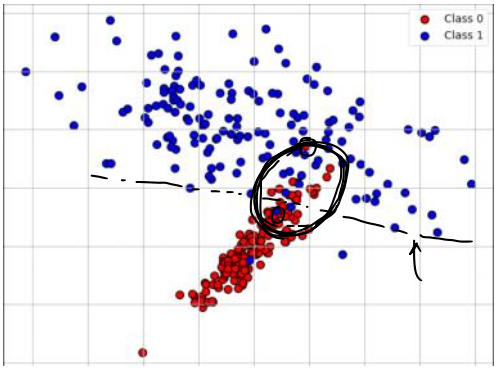
info loss

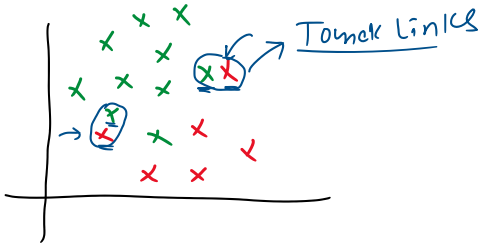
Tomek Links

29 April 2024 13:56

The Problem

In many real-world classification tasks, the boundary between different classes may not be clear-cut. Instances of different classes can be very close to each other in the feature space, creating overlap. This overlap can confuse learning algorithms, leading to poorer generalization and increased misclassification, especially near the boundaries.





What are Tomek Links

- Nearest Neighbors: A pair of instances forms a Tomek Link if each instance in the pair is the closest data point to the other in the dataset.
- Opposing Classes: These nearest neighbors must belong to different classes.

How does it work?

Step 1: Identify Nearest Neighbors

- What to Do: For each data point in the dataset, find its nearest neighbor. This means finding the closest other data point in terms of distance, typically using a standard metric like Euclidean distance.

Step 2: Determine if They Are Tomek Links

- What to Do: Check each pair of nearest neighbors to see if they are Tomek Links. A pair is considered a Tomek Link if:
- The two points are mutual nearest neighbors (each is the closest to the other).
- The two points belong to different classes.

Step 3: Remove Relevant Data Points

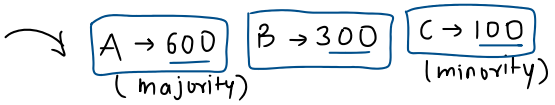
- What to Do: Once you identify Tomek Links, decide on the removal strategy. Commonly, you would remove the data point from the majority class involved in the Tomek Link. This helps to:
- Reduce the overlap between classes by removing points that are likely contributing to boundary confusion.
- Clean up the dataset by potentially removing noisy or mislabeled instances.

Step 4: Update the Dataset

- What to Do: After removing the selected data points, update the dataset. This new dataset should have clearer class boundaries, with less overlap and noise.

Step 5: Proceed with Further Data Processing or Modeling

- What to Do: Use the cleaned and updated dataset for further data processing or directly for training your machine learning models. The improved dataset should help in achieving better classification performance, especially in tasks where class imbalance is an issue.

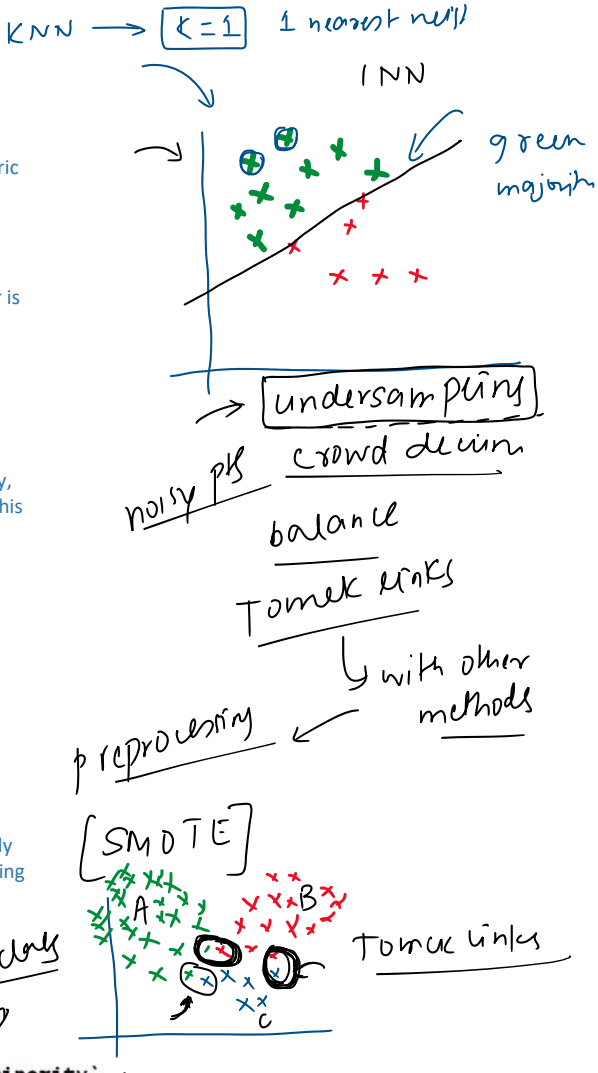


1. `majority`
   - **Effect on AB Links**: Removes instances from Class A when linked with Class B.
   - **Effect on BC Links**: No effect, as neither B nor C is the majority class.

2. `not minority`

In this example, let's assume Class C is the smallest minority class, so it's protected.

- **Effect on AB Links**: Removes instances from Class A and possibly Class B, depending on



- **Effect on AB Links**: Removes instances from Class A when linked with Class B.
- **Effect on BC Links**: No effect, as neither B nor C is the majority class.
- **Effect on CA Links**: Removes instances from Class A when linked with Class C.

3. `not majority`

- **Effect on AB Links**: Removes instances from Class B.
- **Effect on BC Links**: Removes instances from Classes B and C.
- **Effect on CA Links**: Removes instances from Class C.

In this example, let's assume Class C is the smallest minority class, so it's protected.

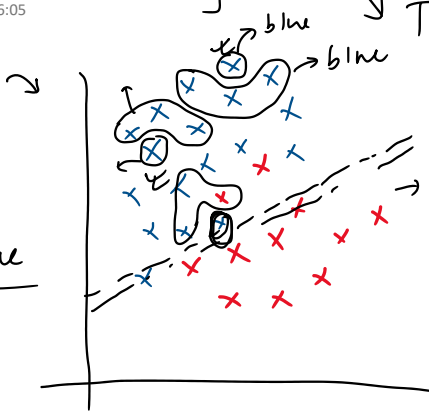
- **Effect on AB Links**: Removes instances from Class A and possibly Class B, depending on whether B is considered non-minority based on the context (if B is significantly larger than C).
- **Effect on BC Links**: Removes instances from Class B but protects Class C.
- **Effect on CA Links**: Removes instances from Class A, protecting Class C.

4. `all`

- **Effect on AB Links**: Removes instances from both Class A and Class B.
- **Effect on BC Links**: Removes instances from both Class B and Class C.
- **Effect on CA Links**: Removes instances from both Class C and Class A.

Edited Nearest Neighbors
29 April 2024 16:05

Tomek Links → ENN → decision boundary → clear
→ select points → remove



ENN

↳ clear
noisy.
overlapping

step 1 → KNN → K=3 default

{ 2 blue
1 red → blue

step 2 → move to every data point

2 red → red
1 blue

{ all → mode }

↳ data point

Tomek Links → aggressive
↳ undersampling

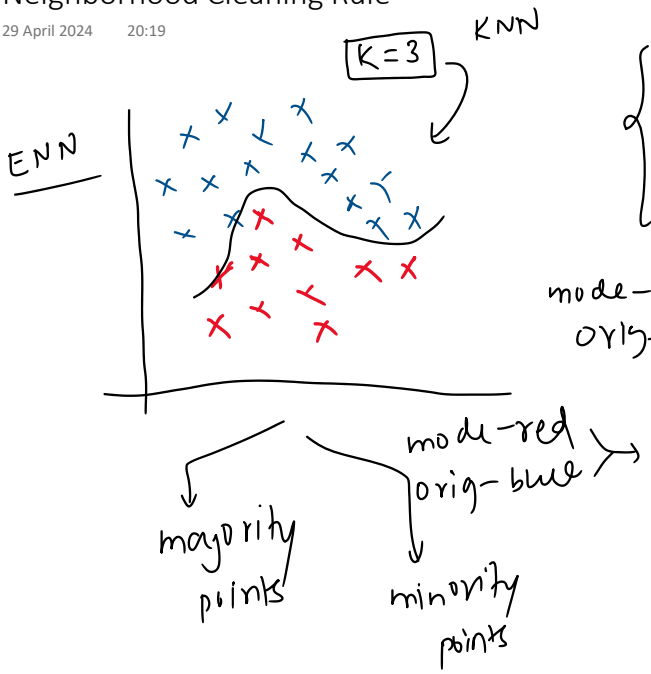
+100

decision boundary → declutter → undersampling

↓
SMOTE

Neighborhood Cleaning Rule

29 April 2024 20:19



mode-blue
orig-red > mismatch

mode-red
orig-blue >

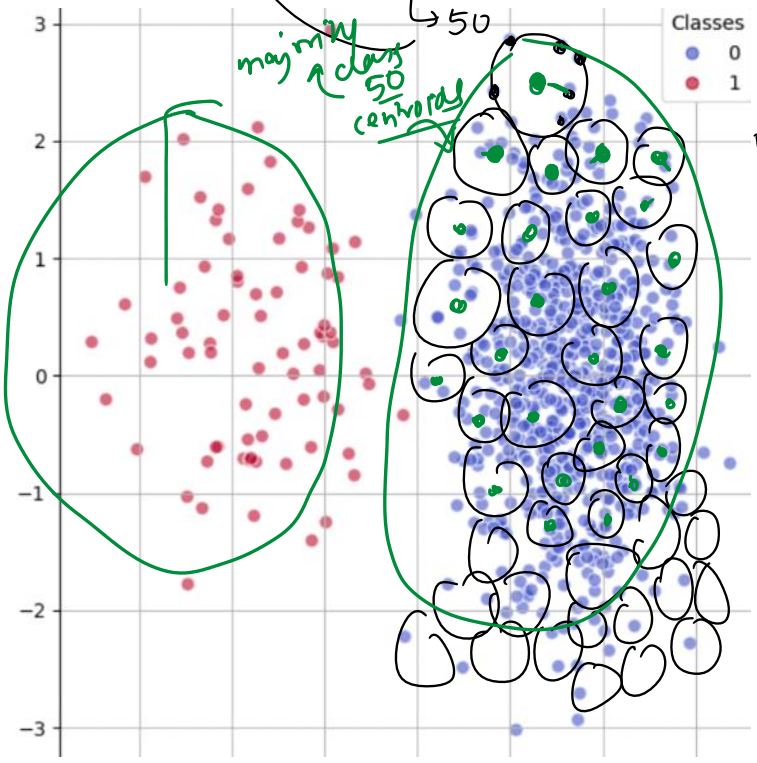
majority points

minority points

The Neighborhood Cleaning Rule (NCL) is another instance reduction technique that refines datasets for classification tasks, particularly focusing on balancing class distributions in imbalanced datasets. It extends the concept of the Edited Nearest Neighbor (ENN) rule by incorporating aspects of both undersampling the majority class and cleaning the data around the minority class to enhance classification accuracy and reduce bias.

Cluster Centroids

30 April 2024 01:06



0 1 Undersampling

950 50

↳ 50

majority class 50 centroids

Prototype generation

↓

undersampling

↳ new data points

K-means also

↓

clustering

K = 50

majority dp → new dp

↓

represent

undersampling

Instance Hardness Threshold

30 April 2024 16:19

https://towardsdatascience.com/instance-hardness-threshold-an-undersampling-method-to-tackle-imbalanced-classification-problems-6d80f91f0581

file:///C:/Users/Nitish/Downloads/s10994-013-5422-z.pdf