# Tabular Data Pre-Processing

Ninghao Liu

University of Georgia

*ninghao.liu@uga.edu*

February 1, 2024

# Overview

Data Imbalance

Missing Values

# Data Imbalance

- A dataset is imbalanced if the classes are not approximately equally represented.
- Imbalance on the order of 100 to 1 is prevalent in applications such as fraud detection.
- The performance of machine learning algorithms is typically evaluated using predictive accuracy.
  - Achieving a low loss value does not mean you have trained a good model.
- Costs more time training on useless samples.

We limit our discussion to binary classification scenarios.

# Data Augmentation - Effect on Models

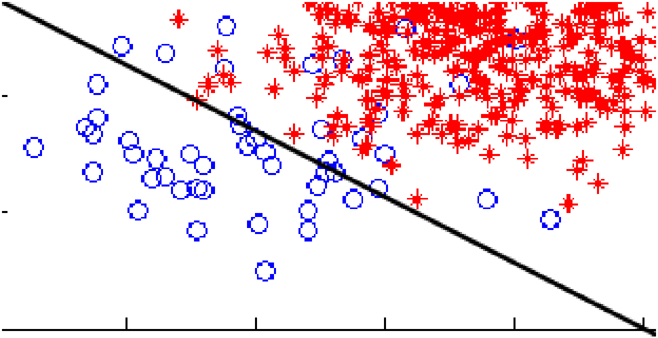


Figure: Linear classification of imbalanced data showing bias towards the majority class.

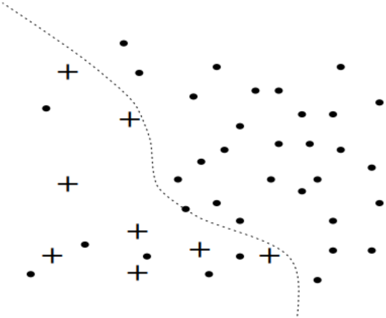
# One-Sided Selection

**Possible Solution 1**

- One-Sided Selection: Under-sampling the majority class [1].

[1]"Addressing the Curse of Imbalanced Training Sets: One Sided Selection". ICML 1997.

# One-Sided Selection



**Three-Types** of negative samples (dot points)

- Noisy samples: whose labels are problematic.
- Redundant samples: can be taken over by other samples.
- Safe samples: worth being kept.

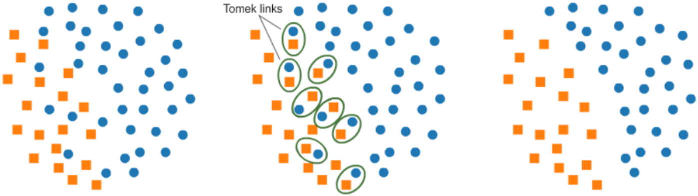
# One-Sided Selection

We try to eliminate examples suffering from the noise. These can easily be detected using the concept of *Tomek links* (Tomek, 1976).

**Tomek links**

- Take two examples, $\mathbf{x}$ and $\mathbf{y}$, so that each has a different label.
- Let $d(\cdot, \cdot)$ denote a distance metric.
- $(\mathbf{x}, \mathbf{y})$ is called a Tomek link if no example $\mathbf{z}$ exists such that $d(\mathbf{x}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$ or $d(\mathbf{y}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$.
- Examples participating in Tomek links are noisy.

# One-Sided Selection




**Tomek links**

- Take two examples, $\mathbf{x}$ and $\mathbf{y}$, so that each has a different label.
- Let $d(\cdot, \cdot)$ denote a distance metric.
- $(\mathbf{x}, \mathbf{y})$ is called a Tomek link if no example $\mathbf{z}$ exists such that $d(\mathbf{x}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$ or $d(\mathbf{y}, \mathbf{z}) < d(\mathbf{x}, \mathbf{y})$.
- Examples participating in Tomek links are noisy.

# One-Sided Selection

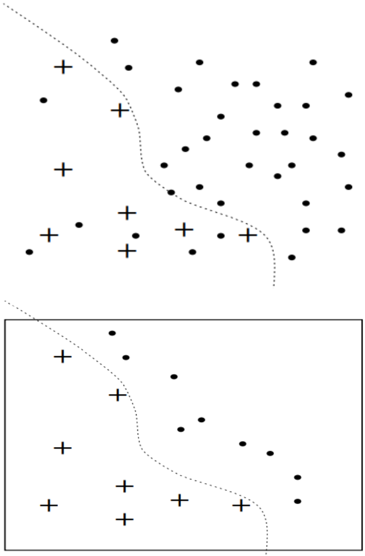
We also try to reduce the number of redundant data points.

- Let $\mathcal{D}$ denote the original dataset. We aim to find a **consistent subset** $\mathcal{C}$ from $\mathcal{D}$.
- An set $\mathcal{C} \subseteq \mathcal{D}$ is a consistent subset of $\mathcal{D}$ if, when applying the 1-NN rule, it correctly classifies samples in $\mathcal{D}$.
- Any training set is a consistent subset of itself.

# One-Sided Selection

1. Let $D$ be the original training set.
2. Initially, $C$ contains all positive examples from $D$ and one randomly selected negative example.
3. Classify $D$ with the 1-NN rule using the examples in $C$, and compare the assigned concept labels with the original ones. Move all misclassified examples into $C$ that is now consistent with $D$ while being smaller.
4. Remove from $C$ all negative examples participating in Tomek links. This removes those negative examples that are believed borderline and/or noisy. All positive examples are retained. The resulting set is referred to as $T$.

# One-Sided Selection



# Re-Sampling

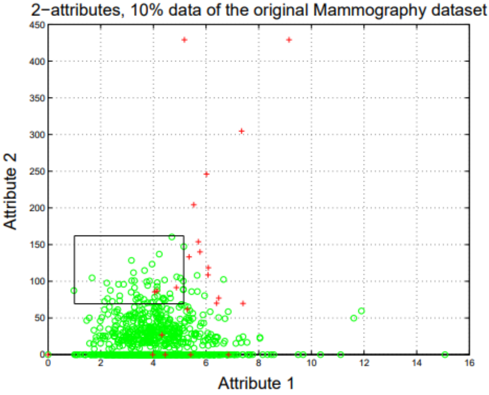
**Possible Solution 2**[2]

- Re-sampling: Random re-sampling consisted of re-sampling the smaller class at random until it consisted of as many samples as the majority class.
- Focused Re-sampling: Re-samples only those minority examples that occurred on the boundary between the minority and majority classes.

Limitation: Does not significantly improve minority class recognition.

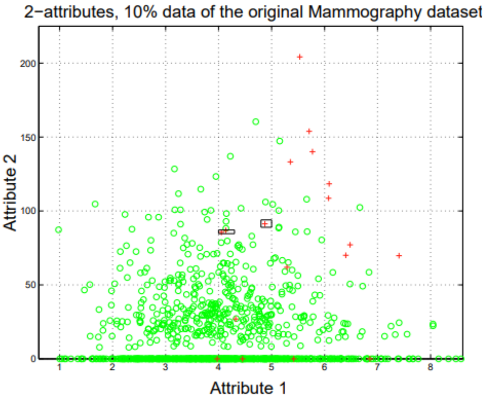
[2] "Learning from Imbalanced Data Sets: A Comparison of Various Strategies"

# Re-Sampling




(a)

Decision region in which the three minority class samples (shown by '+') reside after building a decision tree. The '+' samples are mis-classified.




(b)

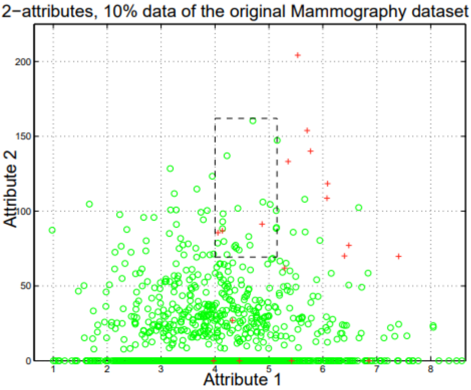
New decision regions after re-sampling. The regions are very specific. Replication of the minority class does not cause its decision boundary to spread into the majority class region.

**Possible Solution 3**

- Synthetic Sampling [3]
- To identify similar but more specific regions in the feature space as the decision region for the minority class.
- The minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with replacement.

[3] SMOTE: synthetic minority over-sampling technique

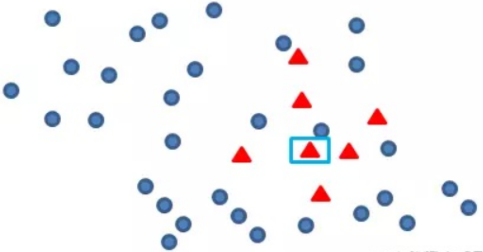
# SMOTE




(c)

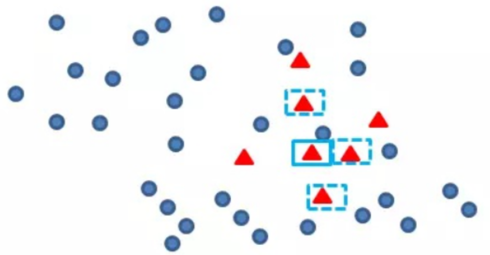
Minority-class samples are correctly classified. And we expand the decision region for the minority class.

# SMOTE



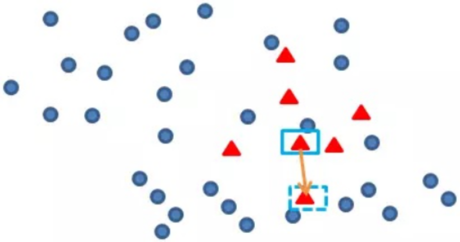
Step 1: Randomly choose minority-class samples.

# SMOTE



Step 2: Given a minority-sample $\mathbf{x}_i$, find its neighbors $\{\mathbf{x}_i^1, ..., \mathbf{x}_i^k, ..., \mathbf{x}_i^K\}$.

# SMOTE



Step 3: Creat a sample along the path between $\mathbf{x}_i$ and one of its neighbors, say $\mathbf{x}_i^k$.

# Missing Values in Data[4]

| A | B | C |
|---|---|---|
| 1 | a | 4 |
| 2 | — | 7 |
| — | — | 5 |

- Feature values are missing or meaningless for some data samples.
- Some models are not compatible with missing values, e.g., linear models.

[4] Jiawei Han et al., "Data Mining", Section 3.2

# Missing Values in Data

**Possible Solution 1**

- Just delete the sample with missing values!
- Pros: Easy!
- Cons: When the portion of missing-value samples is large, we cannot afford abandoning those data.

**Possible Solution 2**

- Filling: Complete the missing values.
- Mean Completer: Fill the missing values with the mean value of the feature on other samples. (You can also use the median value.)
- Conditional Mean Completer: Fill the missing values with the mean value of the feature on other samples having the same label.
- Mode Completer: Similar to the mean completer, but used to handle non-numeric features, e.g., categorical features.

**Possible Solution 2**

- Hot deck imputation: Given a missing-value sample, find its neighbors and use their feature (mean) value to predict the missing value.
- Regression: Build a regression model based on observed data to predict the missing values.