Multi-class Imbalanced Learning

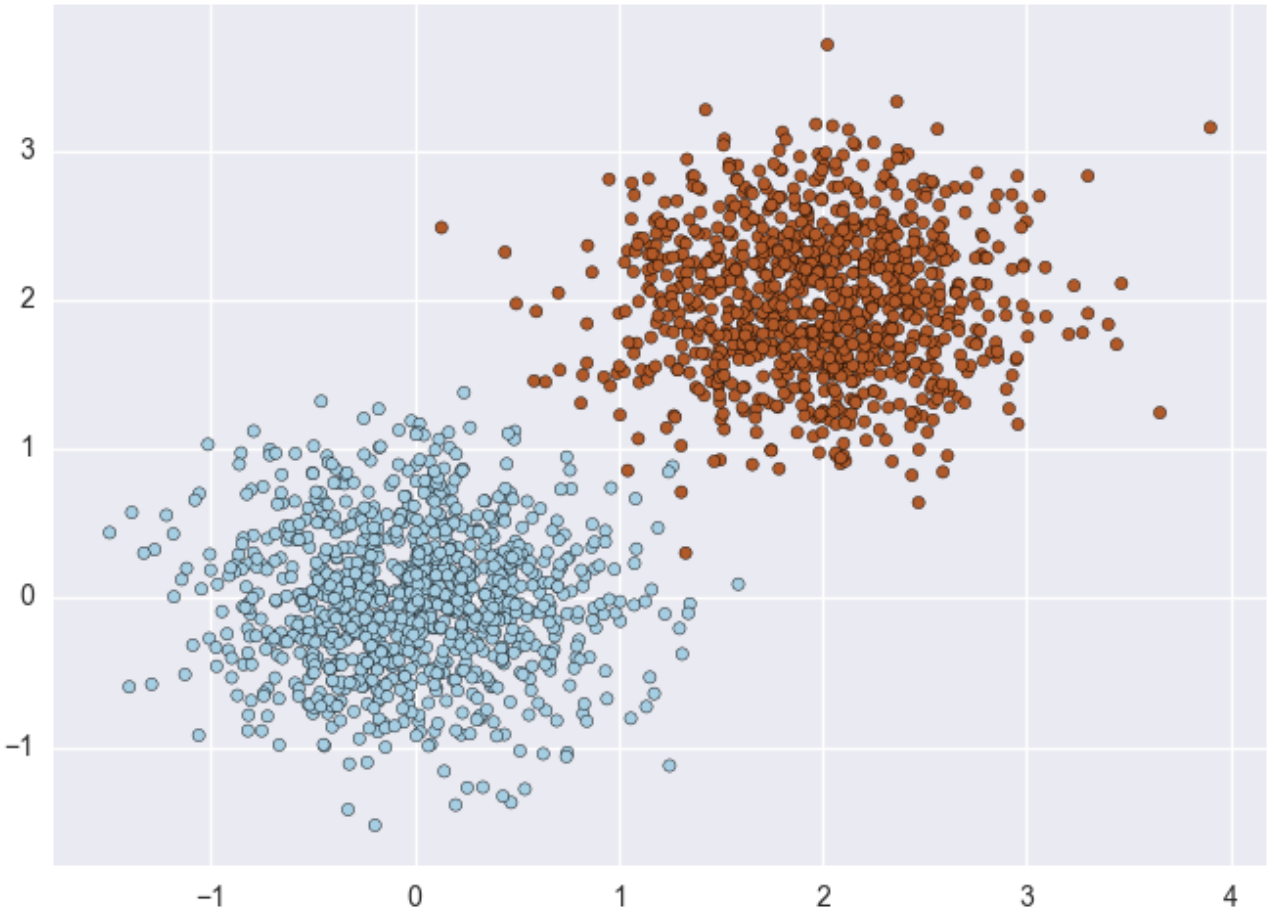
DingHao

December 5, 2016

Contents

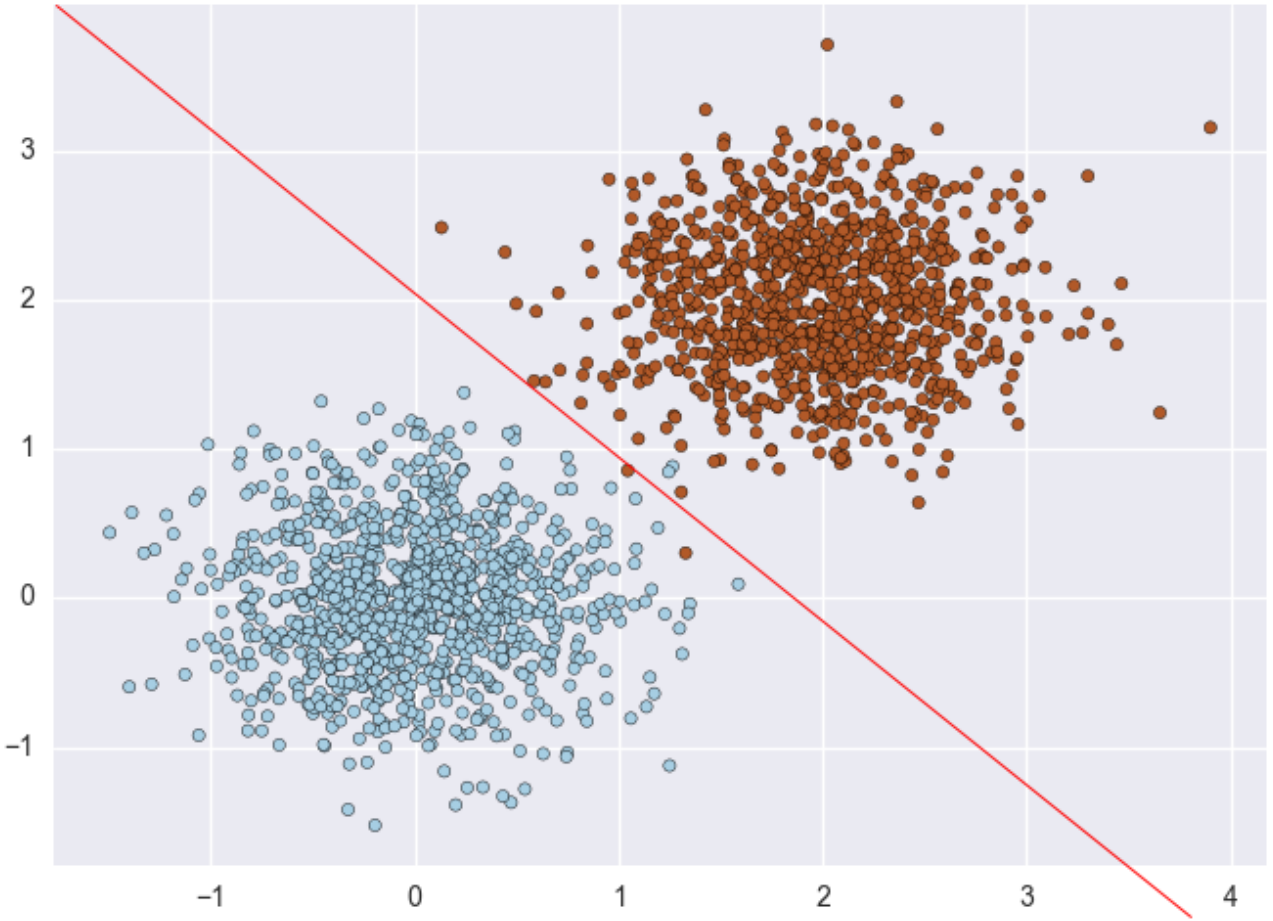
Introduction

Classification Problem



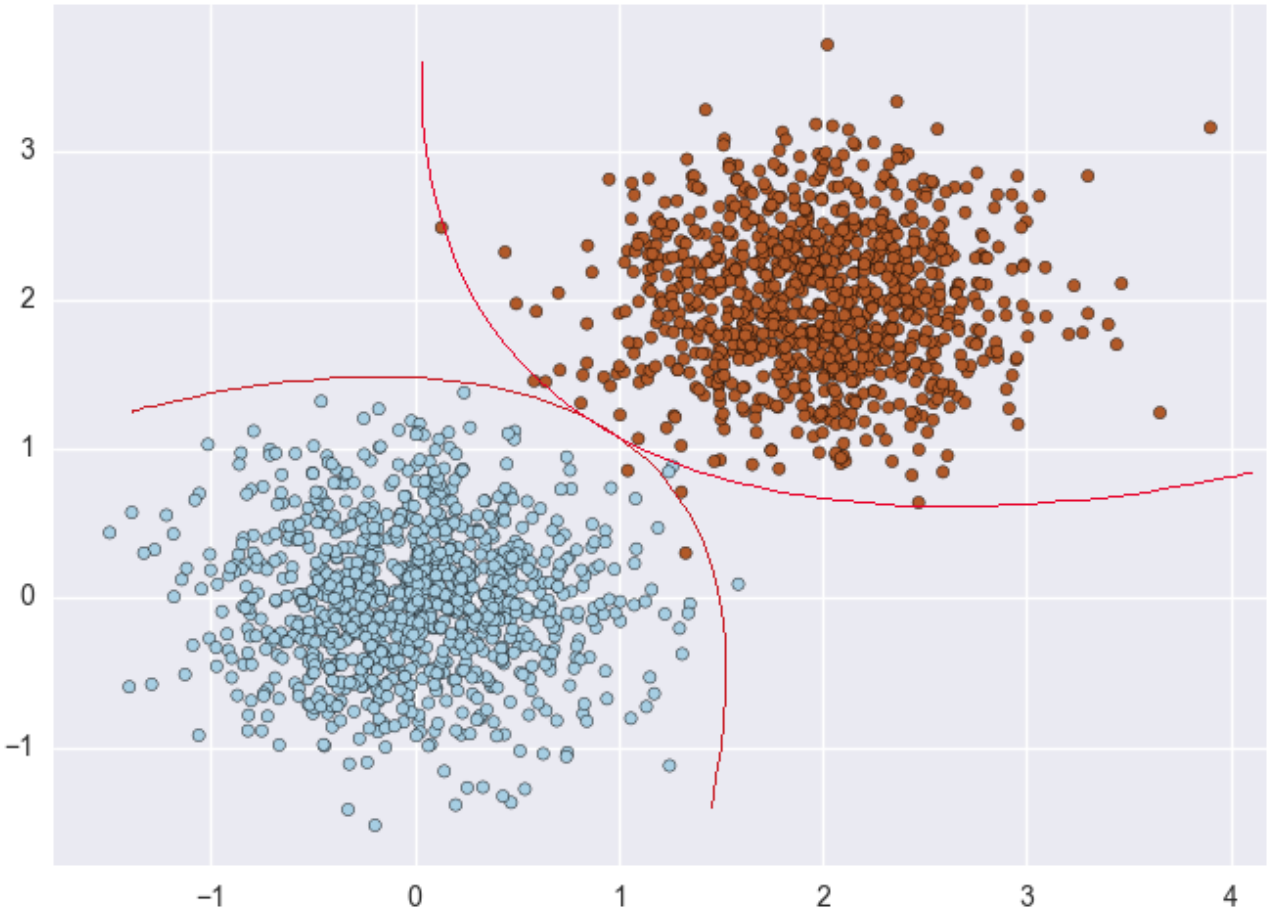
Introduction

classifiers



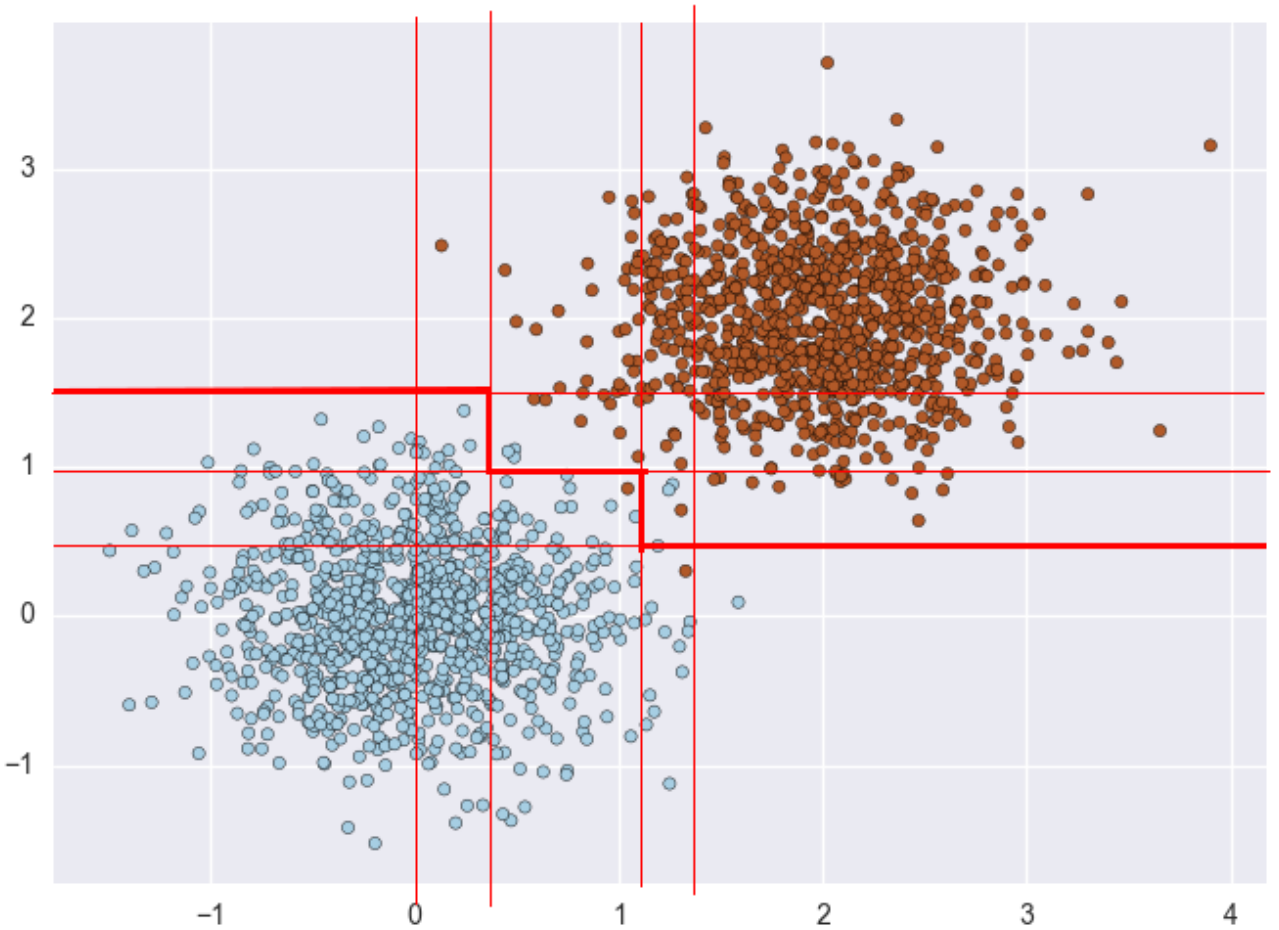
Introduction

classifiers



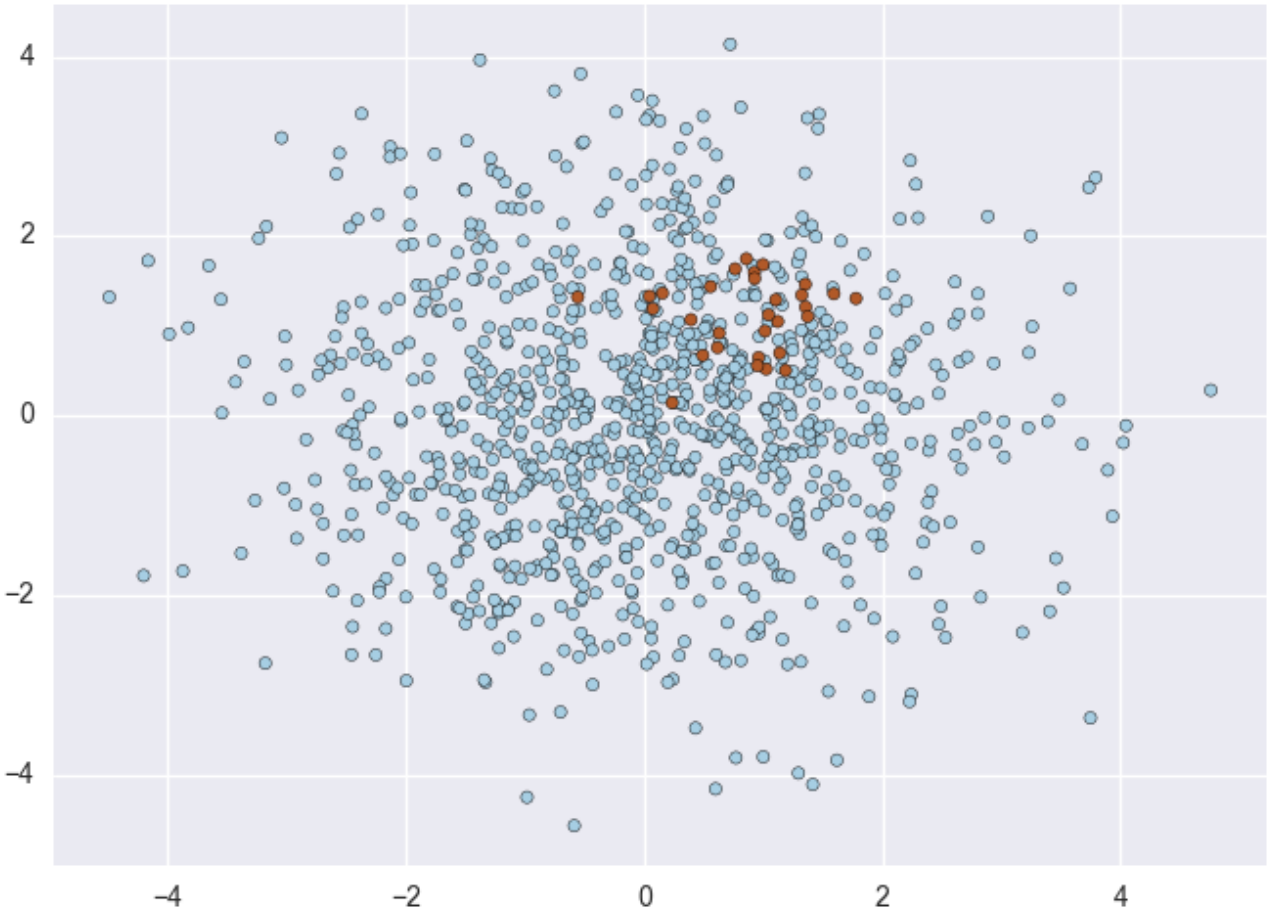
Introduction

classifiers



Introduction

Real World

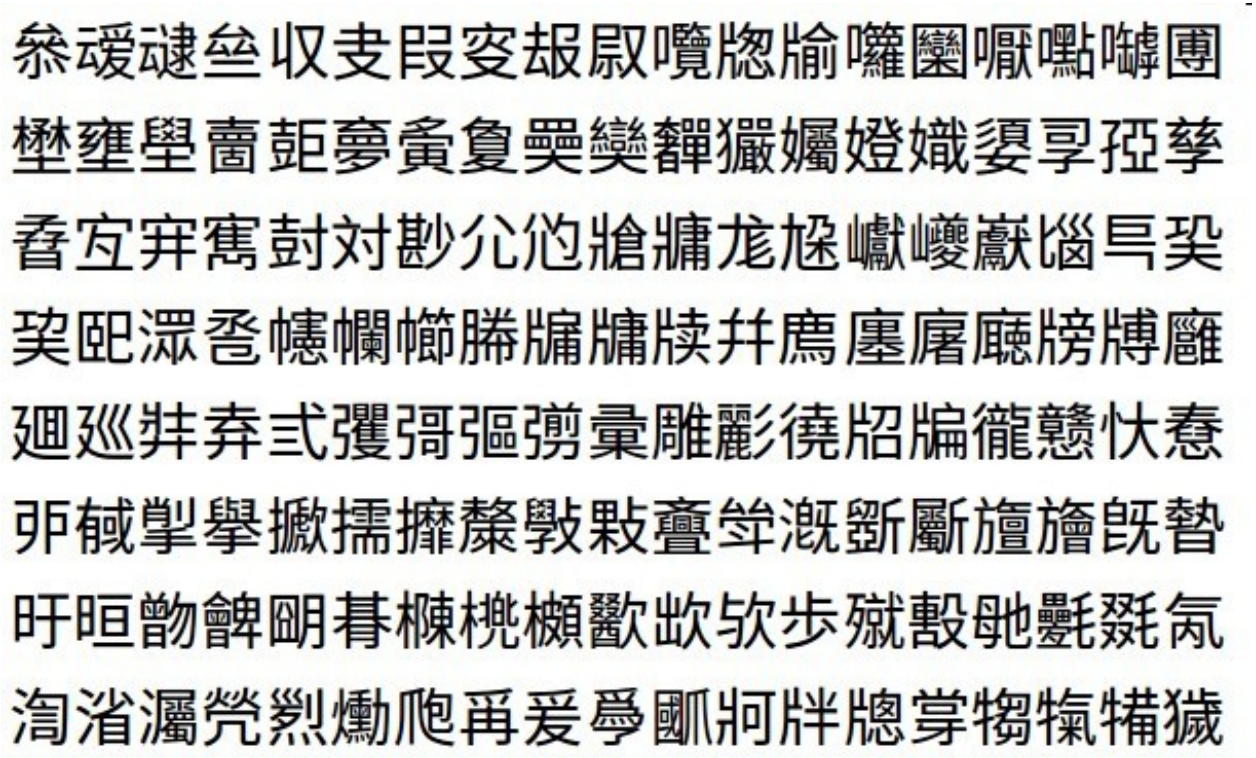


Introduction

Doggy

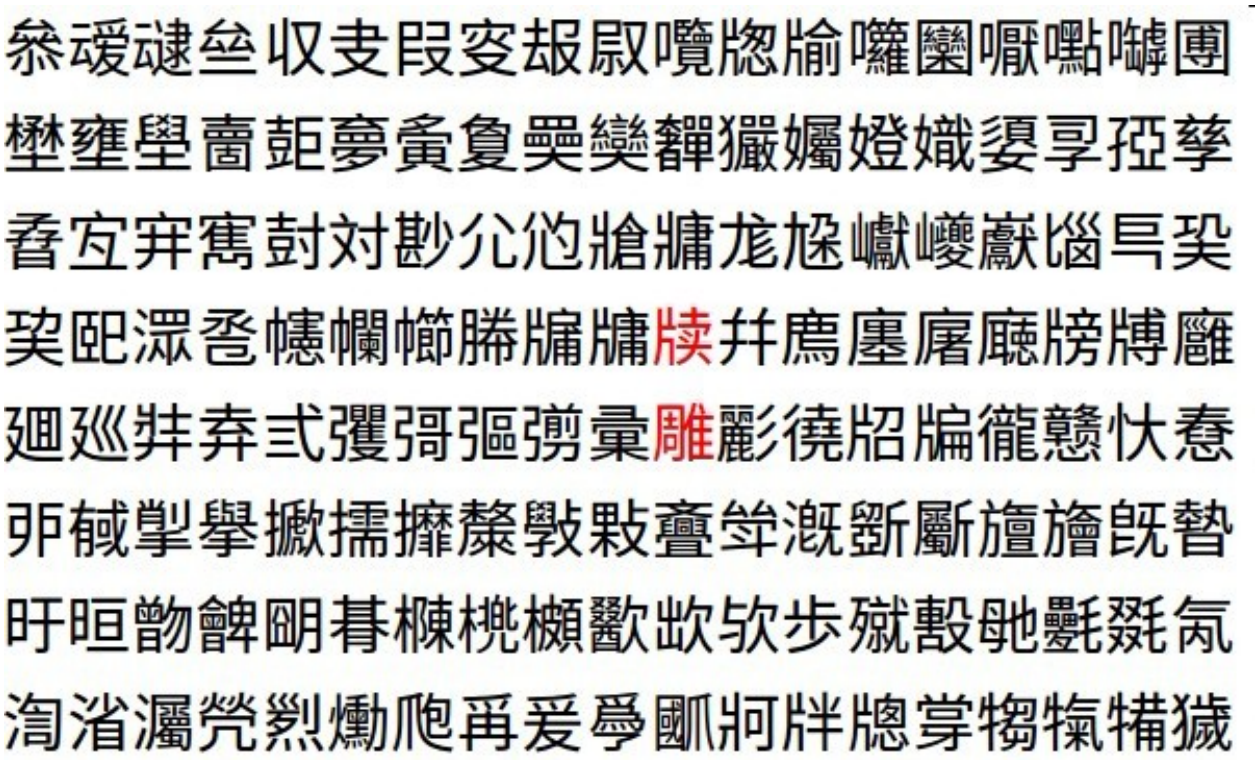
Introduction

Rare Word



Introduction

Rare Word



Introduction

Imbalanced Ratio

imbalanced ratio = majority class / minority class

ZooScan

427 / 28 = 15.25

Kaggle

1979 / 9 = 219.89

WHOI

2606720 / 4 = 651680

EVEN MORE THAN 10^8 !

WHY?

Evaluation Criteria

Name	Formula	Explanation
True Positive Rate (TP rate)	TP / (TP + FP)	The closer to 1, the better. TP rate = 1 when FP = 0. (No false positives)
True Negative Rate (TN rate)	TN / (TN + FN)	The closer to 1, the better. TN rate = 1 when FN = 0. (No false negatives)
False Positive Rate (FP rate)	FP / (FP + TN)	The closer to 0, the better. FP rate = 0 when FP = 0. (No false positives)
False Negative Rate (FN rate)	FN / (FN + TP)	The closer to 0, the better. FN rate = 0 when FN = 0. (No false negatives)

$$G - mean = \sqrt{TPr * TNr}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN} = TPr$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall}$$

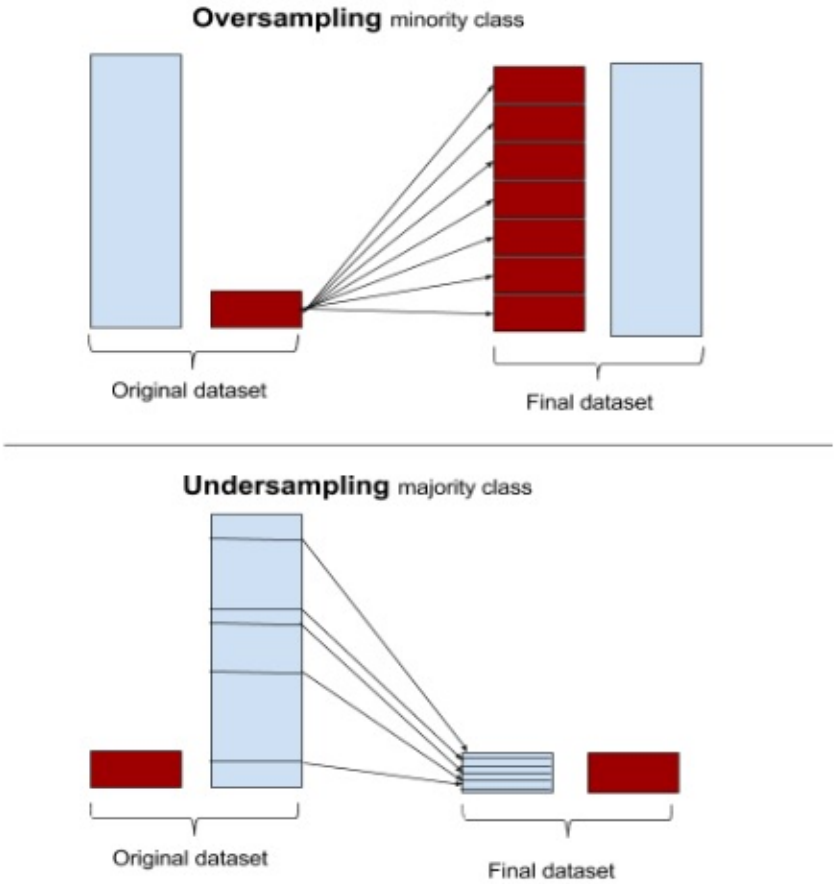
Approaches

Overview

- Sampling
  Under-sampling
  Over-samping
- Cost-sensitive learning
- Ensembled classifier
  EasyEnsemble
  BalanceCascade

Approaches

Sampling



Best approache: SMOTE

Approaches

Cost-sensitive

$$L(x, i) = \sum_j P(j|x)c(i, j)$$

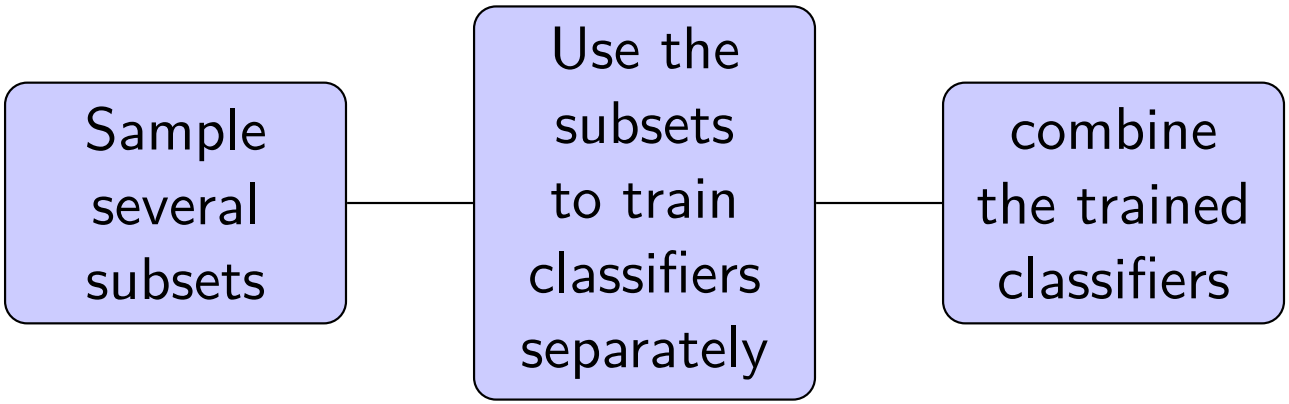
Minimize the overall cost.

- x : an example
- i : a class
- j : the jth class
- P : Probability
- c : cost matrix

Best approaches: AdaCost, AsymBoost

Approaches

Ensembled Classifier



Best approaches: EasyEnsemble, BalanceCascade, SMOTEBoost

Approaches

EasyEnsemble.M $\Rightarrow$ EasyEnsemble.D

1: Input:A set of minority class examples $\mathcal{P}$, k-1 sets of majority class examples $\mathcal{N}$, $|\mathcal{P}| < |\mathcal{N}_k|$, the number of subsets T to sample from $\mathcal{N}_k$, and s_i,the number of iterations to train an AdaBoost ensemble H_i

2: for i$\Leftarrow$ 1:T

3: $D_i = \mathcal{N}_1$

4: for t$\Leftarrow$ 1:k

5: Randomly sample a subset $\mathcal{N}_{it}$ from $\mathcal{N}_k$, N_{it},$|N_{it}| = |P| + \frac{C_1 * (|C_i - |P|)}{|C_k|}$ in the t^{th}

6: $D_i = D_i \bigcup \mathcal{N}_{it}$

7: $H_t(x) = sgn\left(\sum_{d=1}^{s_i} \alpha_{t,d} h_{t,d}(x) - \theta_i\right)$

8: $H(x) = sgn\left(\sum_{t=1}^{T} \sum_{d=1}^{s_i} \alpha_{t,d} h_{t,d}(x) - \sum_{t=1}^{T} \theta_i\right)$

Q.-Q.Li, and X.-Y.Liu. "EasyEnsemble.M for multiclass imbalance problem." In: Pattern Recognition and Artificial Int
27.2(2014):187-192.

Future Work

- Optimize the algorithm to cost less runtime
- Use Kaggle and WHOI datasets
- Increase the amount of time in each dataset


Q&A