



# A Novel Adaptive Minority Oversampling Technique for Improved Classification in Data Imbalanced Scenarios

Ayush Tripathi, Rupayan Chakraborty, Sunil Kumar Kopparapu

ICPR - 2020, Milan, Italy 10-15 January, 2021

# Introduction



©2020

### Class Imbalance

- ▶ Number of samples corresponding to each class is not proportionate.
- Minority class is underrepresented i.e. it has fewer samples compared to the majority class.
- Performance of conventional classifiers tends to get biased towards the majority class.
- ▶ Real-world Applications :
  - ✤ fraud detection
  - $\star$  network intrusion detection
  - disease diagnosis
  - $\star$  software defect detection
  - $\leftarrow$  bioinformatics

## Prevalent Techniques



©2020

#### Resampling Techniques

- ▶ Re-balance the sample space.
  - ★ Oversampling : Generate minority class samples SMOTE, ROS, ADASYN etc.
  - $\star$  Undersampling : Discard majority class samples.
  - $\star$  Hybrid Sampling : Use a combination of oversampling and undersampling.

### Cost Sensitive Learning

• Assign higher misclassification cost to minority class samples compared to the majority class samples.

#### Ensemble based classifiers

• Use a combination of multiple base classifiers.

## Motivation



#### Requirements

- Mitigate the woes of class imbalance.
- ▶ Ensure that the minority samples are diverse enough to maintain the original data distribution.

### Proposed Algorithm

- ▶ Local distribution characteristics of the data are often ignored during the oversampling process. So, propose an algorithm that makes use of the local distribution of the data.
- ▶ 2 step approach is used i.e. oversampling followed by an undersampling.
- ▶ Conventional SMOTE is used for oversampling.
- ▶ Probability density estimation based undersampling.



# Proposed Algorithm



 $\odot 2020$ 

### Step 1 : Sample Generation

- ▶ Preliminary step : Identical to SMOTE oversampling technique.
- ▶ Given N minority and M majority class instances, obtain an intermediate oversampled distribution consisting of  $\hat{N}_1$  additional minority samples.

### Step 2 : GMM Clustering

▶ The minority distribution consisting of  $(N + \hat{N}_1)$  samples are clustered into C clusters using a Gaussian-Mixture Model (**GMM**) based clustering algorithm.

# Proposed Algorithm



### Step 3 : Adaptive Sample Selection

- ▶ The majority sample distribution is employed to adaptively select M N samples from the generated  $\hat{N}_1$  minority samples.
- ▶ Compute the number of majority class samples  $(q_i)$  for which  $p(m_j \in C_i | M)$  exceeds the probability threshold  $(p_t)$ .
- ▶ Cluster weight is defined as

$$w_i = \begin{cases} (1 - \frac{q_i}{M}), & \text{if } w_i > w_t \\ 0, & \text{otherwise} \end{cases}$$

Select  $N_i$  points belonging to the synthetically generated samples  $\hat{N}_1$  lying in the  $i^{th}$  cluster.

$$N_i = \left( (M - N) \times \frac{w_i}{\sum_{i=1}^C w_i} \right)$$

©2020

## Results: $F_1$ Scores obtained using the proposed approach



Classification Task	$F_1$ Score
Pima	$0.6727 \pm 0.0023$
Glass0	$0.7530 \pm 0.0025$
Vehicle0	$0.9498 \pm 0.0004$
Ecoli1	$0.8261 \pm 0.0046$
Yeast3	$0.7591 \pm 0.0020$
Pageblock	$0.9600 \pm 0.0063$
Glass5	$0.6666 \pm 0.1333$
Yeast5	$0.6853 \pm 0.0040$
Yeast6	$0.5219 \pm 0.0107$
Abalone	$0.0688 \pm 0.0002$
Anger-Anxiety	$0.9501 \pm 0.0002$
Anger-Disgust	$0.8952 \pm 0.0021$
Anger-Happy	$0.7338 \pm 0.0027$
Anger-Sad	$0.9913 \pm 0.0003$
Anxiety-Disgust	$0.8570 \pm 0.0064$
Boredom-Disgust	$0.8978 \pm 0.0006$
Happy-Disgust	$0.9210 \pm 0.0020$
Neutral-Anger	$0.9353 \pm 0.0021$
Neutral-Disgust	$0.8646 \pm 0.0003$
Sad-Disgust	$0.9632 \pm 0.0024$
AD-MCI	$0.5773 \pm 0.0081$
HC-MCI	$0.6220 \pm 0.0033$
04clover5z-600-5-70-BI	$0.5268 \pm 0.0041$
04clover5z-600-5-60-BI	$0.5414 \pm 0.0017$
04clover5z-600-5-50-BI	$0.5493 \pm 0.0024$
04clover5z-600-5-30-BI	$0.5419 \pm 0.0003$
04clover5z-600-5-0-BI	$0.5674 \pm 0.0004$



### Results: $F_2$ Scores obtained using the proposed approach



Classification Task	$F_2$ Score
Pima	$0.6892 \pm 0.0032$
Glass0	$0.8267 \pm 0.0024$
Vehicle0	$0.9734 \pm 0.0002$
Ecoli1	$0.8588 \pm 0.0061$
Yeast3	$0.7986 \pm 0.0003$
Pageblock	$0.9428 \pm 0.0130$
Glass5	$0.6222 \pm 0.1362$
Yeast5	$0.8004 \pm 0.0027$
Yeast6	$0.6350 \pm 0.0049$
Abalone	$0.1465 \pm 0.0011$
Anger-Anxiety	$0.9532 \pm 0.0004$
Anger-Disgust	$0.8669 \pm 0.0064$
Anger-Happy	$0.7071 \pm 0.0073$
Anger-Sad	$0.9864 \pm 0.0007$
Anxiety-Disgust	$0.8124 \pm 0.0091$
Boredom-Disgust	$0.8774 \pm 0.0032$
Happy-Disgust	$0.9027 \pm 0.0050$
Neutral-Anger	$0.9287 \pm 0.0037$
Neutral-Disgust	$0.8405 \pm 0.0010$
Sad-Disgust	$0.9446 \pm 0.0055$
AD-MCI	$0.5563 \pm 0.0087$
HC-MCI	$0.6003 \pm 0.0053$
04clover5z-600-5-70-BI	$0.6577 \pm 0.0080$
04clover5z-600-5-60-BI	$0.7141 \pm 0.0023$
04clover5z-600-5-50-BI	$0.7285 \pm 0.0033$
04clover5z-600-5-30-BI	$0.7153 \pm 0.0005$
04clover5z-600-5-0-BI	$0.7478 \pm 0.0007$



### Summary



 $\odot 2020$ 

- ▶ A three step hybrid sampling technique that adaptively selects specific data points during the process of oversampling is proposed.
- ▶ The local distribution characteristics of the data has been given due importance in the process of oversampling.
- ▶ By the proposed technique, we obtain:
  - $\bigstar$  A balanced representation of the data distribution.
  - $\star$  The synthetically generated samples are adequately diverse and representative of the original distribution.
- ▶ The efficacy of proposed algorithm has been validated on 27 binary classification tasks with varying imbalance ratios.

Thank You



# A Novel Adaptive Minority Oversampling Technique for Improved Classification in Data Imbalanced Scenarios

Ayush Tripathi, Rupayan Chakraborty, Sunil Kumar Kopparapu

ICPR - 2020, Milan, Italy 10-15 January, 2021

