Technique to Handle Imbalanced Datasets

Manohar Kuse
Kumar Harsh Srivastava
Abhishek Dubey
Aniroop Mathur

Problem Description

Class A

Class B

$$| N_A | \gg | N_B |$$

Motivation

- Rare Patterns give skewed datasets
  - Cancer Cell Classification
  - Gene Profiling
  - Credit Card Frauds
- Minority class considered as noise
- Improper training of classifier due to skew
- Required to balance the dataset by artificial samples

Idea

- Artificial instances by combining k-minority samples at a time

- F → Feature vectors of minority class
- M → Number of samples in minority class
- R = { r | 1 < r < M, r is a random number} such that, | R | = k

$$s_{new,p} = {}^{1}/_{k} \sum_{x_i \in R} F(x_i)$$

Pseudo Code

- Inputs : F, k, NE (number of artificial samples)
- Output : E (Artificial samples)

```
For i = 1 to NE
    <tmp> = < 0 >
    For  j = 1 to k
        r = random( 1, M )
        <tmp> = <tmp> +  <F(r)>
    End
    E(i) = <tmp> / k
End
```

References

- **Handling imbalanced datasets: A review.** S. Kotsiantis, D. Kanellopoulos, P. Pintelas.
- **SMOTE: Synthetic Minority Over-sampling Technique.** N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer.
- **Off-line, Handwritten Numeral Recognition by Perturbation Method.** T. Ha, H. Bunke