TCS RESEARCH AND INNOVATION
MUMBAI, INDIA

TATA CONSULTANCY SERVICES
Experience certainty.

A NOVEL ADAPTIVE MINORITY OVERSAMPLING TECHNIQUE FOR IMPROVED CLASSIFICATION IN DATA IMBALANCED SCENARIOS

Ayush Tripathi, Rupayan Chakraborty, Sunil Kumar Kopparapu

{t.ayush, rupayan.chakraborty, sunilkumar.kopparapu}@tcs.com

INTRODUCTION

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
  - ★ network intrusion detection
  - ★ disease diagnosis
  - ★ software defect detection
  - ★ bioinformatics

PREVALENT TECHNIQUES

Resampling Techniques

- Re-balance the sample space.
  - ★ Oversampling : Generate minority class samples - SMOTE, ROS, ADASYN etc.
  - ★ Undersampling : Discard majority class samples.
  - ★ 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 (step 1 and step 2)

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)

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.

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

RESULTS

F1 Scores

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

F2 Scores

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

CONCLUSION

Summary

- 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:
  - ★ A balanced representation of the data distribution.
  - ★ 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.

ICPR 2020

25th International Conference on Pattern Recognition

MILAN, ITALY

JANUARY 10-15, 2021