Dr.Subhankar Mishra

CS460 : Machine Learning

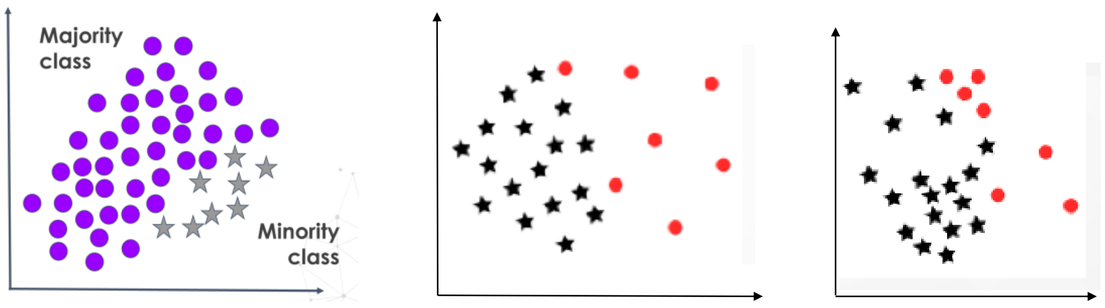
Pradeep Kumar Baisakh
Pradeepkumar.Baisakh@niser.ac.in

DIVING DEEPER INTO SMOTE:

Understanding and Applying Advanced Techniques for Class Imbalance Handling

What is Class Imbalance?

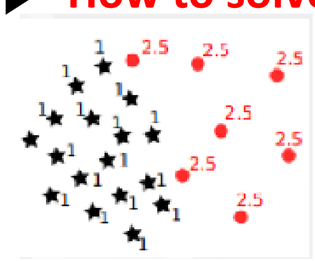
Class imbalance occurs in a classification problem when the distribution of instances among different classes is highly skewed, with one or more classes having significantly fewer instances than others.



Issues due to class imbalance:

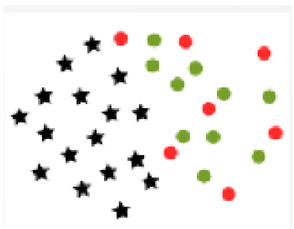
- **Biased models:** algorithms prioritize the majority class
- **Poor Generalization:** resulting models generalize poorly to unseen data
- **Misleading evaluation:** A model may achieve high accuracy by simply predicting the majority class, even if it performs poorly on the minority class

How to solve:



Reweighing

Undersampling



Oversampling

Oversampling:

Artificially adding instances to minority class by creating synthetic data points by using the existing minority data points.

SMOTE stands for **S**ynthetic **M**inority **O**versampling **Te**chnique, which creates new data points in the feature space based on existing minority points.

1. SMOTE(2004):

It generates new synthetic points based on the feature space similarity between existing instances of minority class points.

It uses the KNN algorithm to generate new data points.

- Identify the feature and its K nearest neighbor
- Find distance between both points
- Multiply distance with r, where r ∈ [0,1]
- Identify a new point between the points at computed distance
- Repeat for other feature vectors

$$n = xi + r \cdot (xj - xi)\,, xi, xj \in \mathbb{R}^d, r \in [0.1]$$

Advantages:

- Data diversity
- Lower scope of overfitting
- Versatility

Disadvantages:

- Doesn.t help with improving intra class balance
- Non safe space oversampling
- Impacts the decision boundary

2. ProWSyn(2013):

Proximity Weighted Synthetic Minority Oversampling, generates synthetic samples for the minority class by incorporating proximity weights, where samples closer to the decision boundary are given higher weights

- Calculate the number of synthetic samples that need to be generated for the minority class

$$G = (m_l - m_s) \times \beta \qquad \beta \in [0, 1]$$

- For each majority sample find its nearest K minority samples based on proximity
- Generate synthetic samples for each minority class instance based on its proximity weight. Instances with higher weights are given more emphasis during sampling, leading to a more focused oversampling approach.

3. G-SMOTE(2014):

Geometric-SMOTE wants to define a safe area to synthesize new points. This is to ensure that no synthetic points to be generated to be noisy samples and increase variety of samples to prevent generation of same subclass in minority sample(intra cluster skewness).

- Shuffle the minority elements to create new instances (N times)
- Pick any point as center and find its K neighbors (usually 3), pick any point from K as surface and calculate radius
- Create a point inside the sphere generated

4. KDE-based SMOTE(2014):

It uses a Kernel density estimator to find the probability distribution for the minority samples to create new data points..

- For an unknown probability density function f, the kernel is defined as

$$\tilde{f}(x) = \frac{1}{n}\sum_{j=1}^{n} K_h\,(x - x_j)$$ Where, K = kernel function, h = bandwidth

- Generate synthetic samples based on density of the instance. Higher density instances are likely to be picked for sample generation.
- By tuning the bandwidth parameter, you can control the trade-off between sample diversity and fidelity to the underlying data distribution

5. GAN based oversampling:

Using Generative Adversarial Networks (GANs) to generate synthetic samples for oversampling minority classes.

- GANs consist of two neural networks, a generator and a discriminator, that are trained simultaneously in a competitive manner
- The generator tries to create realistic data points
- The discriminator tries to figure out if a point is real or synthetic
- Both networks are trained together similar to a game to create a balanced data set with realistic points.

6. Gaussian SMOTE(2017):

To prevent synthetic samples from being generated on the same line between the frequently selected samples, a gaussian distribution is used to sample the random number r, thereby expanding the region of synthetic data generation.

- Compute distance between a minority point and its random neighbor
- Pick a random number between zero and distance value for rough position of synthetic point
- Heuristically select a number as a parameter from a gaussian distribution
- Generate synthetic data point using the parameter

Other Variants:

SMOTE-TOMEK(2004), SMOTE-ENN(2004), Borderline SMOTE(2005), ADASYN(2008), Kmeans SMOTE(2018), Polynom fit SMOTE etc

References:

- GitHub SMOTE Variants
- 7 SMOTE Variations for Oversampling
- Lee H, Kim J, , Kim S. **Gaussian-Based SMOTE Algorithm for Solving Skewed Class Distributions**. IJFIS 2017;17:229-
- Barua, Sukarna & Islam, Md & Murase, Kazuyuki. (2013). ProWSyn: Proximity Weighted Synthetic Oversampling Technique for Imbalanced Data Set Learning. 317-328. 10.1007/978-3-642-37456-2_27.
- Handling imbalanced dta using Geometric SMOTE
- Kamalov F, Moussa S, Avante Reyes J. KDE-Based Ensemble Learning for Imbalanced Data. *Electronics*. 2022; 11(17):2703 https://doi.org/10.3390/electronics11172703
- Bing Zhu, Xin Pan, Seppe vanden Broucke, Jin Xiao,A GAN-based hybrid sampling method for imbalanced customer classification,Information Sciences,Volume 609,2022,,ISSN 0020-0255,