SOUP: A new approach in handling Imbalanced Multi-class problems

ML2 Individual Presentation

JAZEL R. JAYME

November 18, 2021

Imbalanced data

BINARY CLASS VS MULTI-CLASS

- Most of the current research concerns binary classification problem
- Multiclass imbalanced classification problems are more difficult than their binary counterparts.
- Mainstream technique for multi-class:
  - SMOTE

Problems with existing approaches:

Do not consider the mutual relations between classes that

are different for majority

What is SOUP?

It stands for Similarity Oversampling and Undersampling Preprocessing. (Lango et al. (2017))

Theory| Safe level.

$$\text{safe}(x_{C_i}) = \frac{1}{n}\sum_{j=1}^{l} n_{C_j}\mu_{ij} \qquad \text{Eq.1}$$

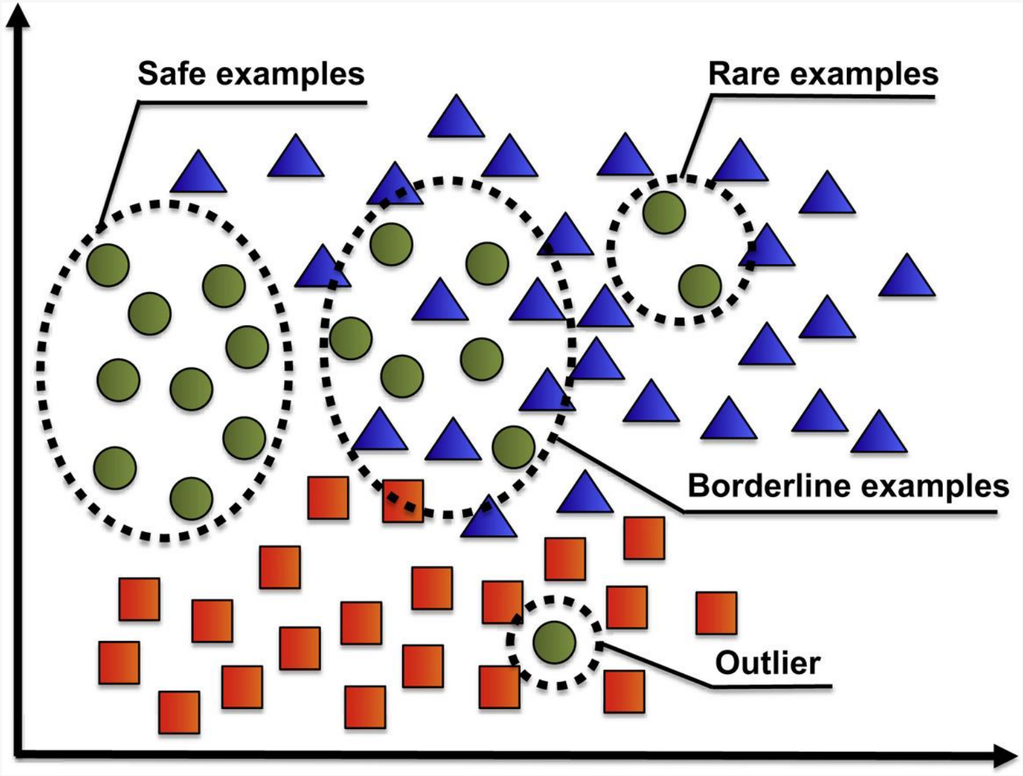
$$\mu_{ij} = \frac{\min(|C_i|, |C_j|)}{max(|C_i|, |C_j|)}$$

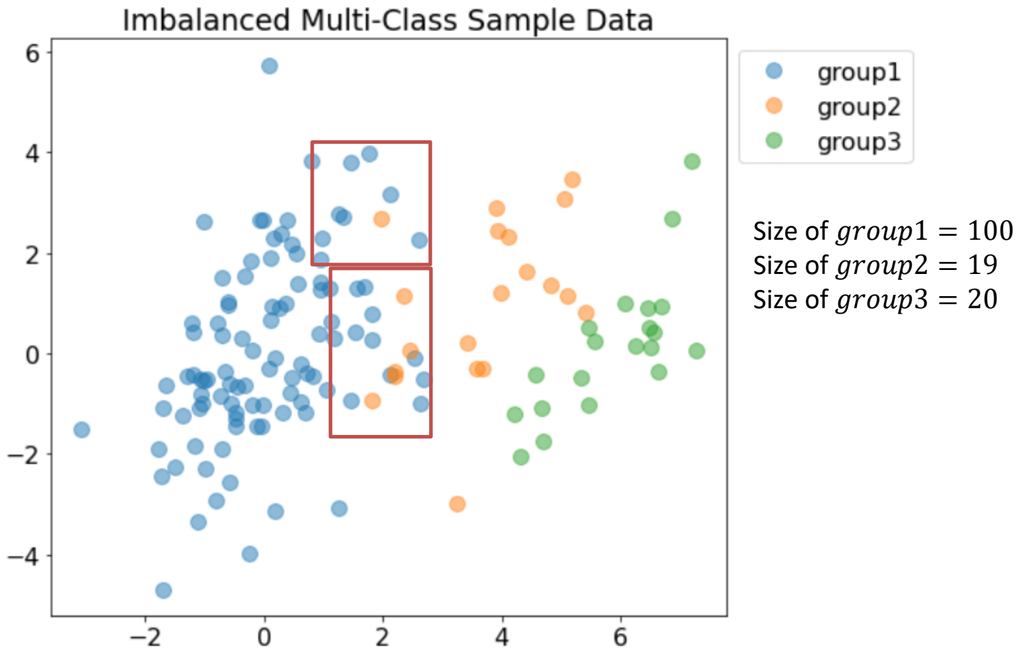
μ_{ij} : the degree of similarity

n_{C_j} : the number of examples from class C_j inside the considered neighborhood of x

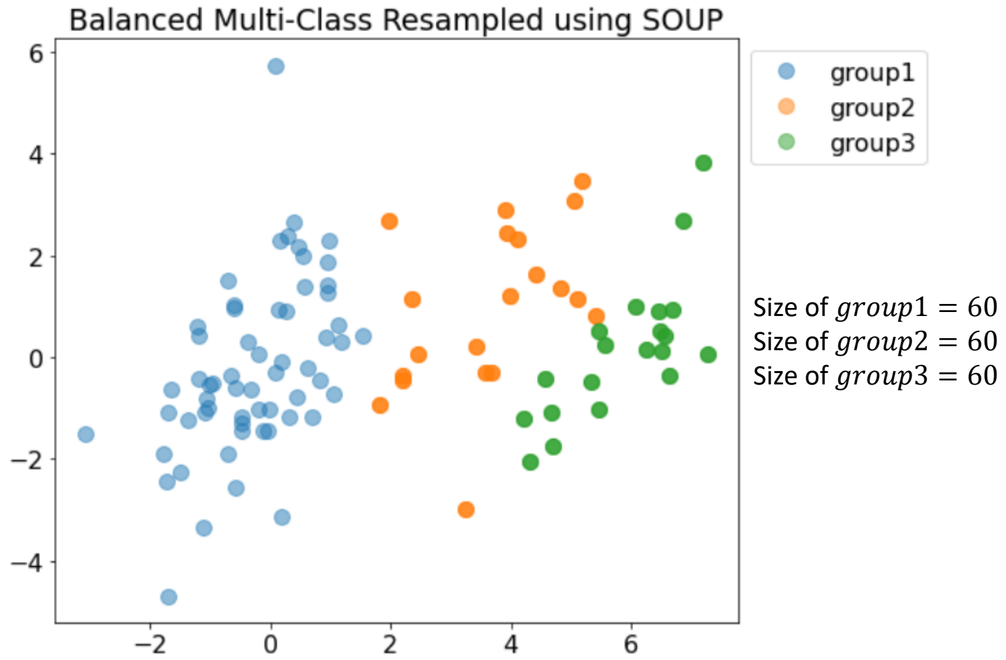
n : a total number of neighbors

$|C_i|$: the number of samples belonging to class i





Size of $group1 = 100$
Size of $group2 = 19$
Size of $group3 = 20$



Size of $group1 = 60$
Size of $group2 = 60$
Size of $group3 = 60$

Process

$$m = \text{mean}(\max\{N_{i_{MIN}}\},\ \min\{M_{i_{MAJ}}\}) \qquad \text{Eq.2}$$

where m = final size of the all classes

$N_{i_{MIN}}$ = the size of minority class

$M_{i_{MAJ}}$ = the size of the majority class

MEASURE the Safe level of samples within class

UNDERSAMPLE Majority classes with size m

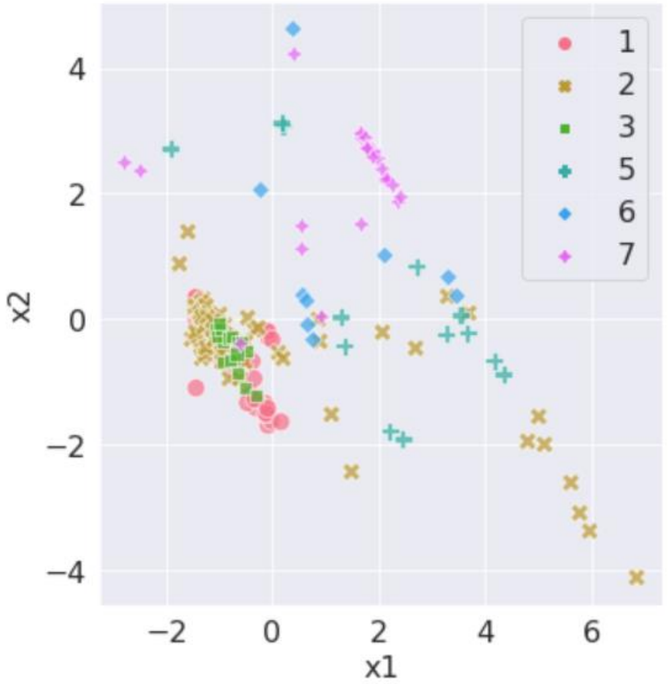
(remove unsafe samples)

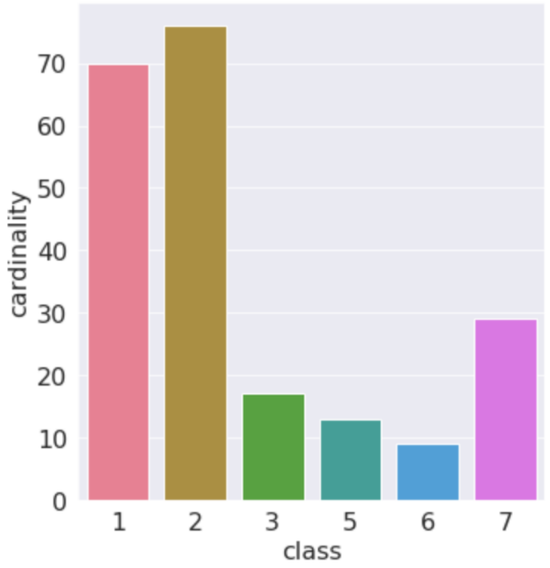
OVERSAMPLE minority classes with size m

(duplicate safe samples within safe levels)

UCI Glass

Imbalanced Data





6 Types of glass
2 majority classes; 4 minority classes

01

Data Cleaning and Scaling.

Dataset has **9 numeric attributes** which corresponds to different mineral contents. The goal is to **predict the type of glass**.

02

Data Resampling using SOUP and SMOTE.

Examine plot and play with different *k_neighbors*

03

Auto-ML implementation.

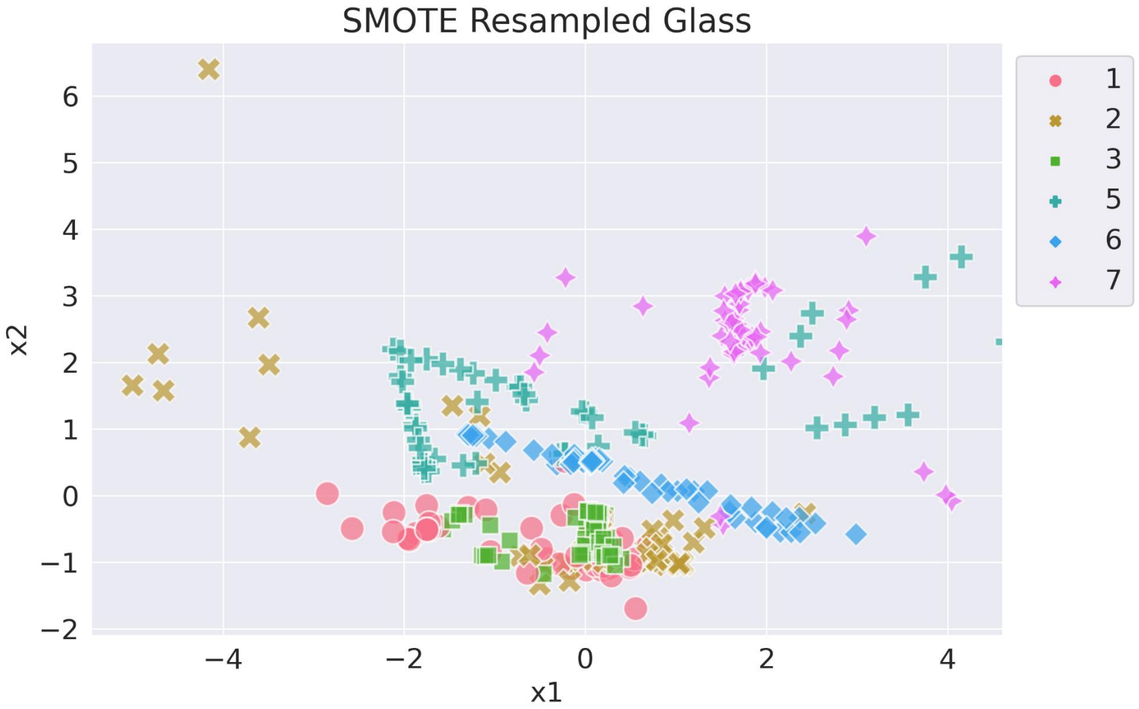
Get the G-Mean score as the performance metric.

$$gmean = (\prod_{i=1}^{n} sensitivity_i)^{\frac{1}{n}}$$

SOUP

All glass types have 34 samples.

SMOTE



All glass types have 53 samples.

Comparing SOUP
With SMOTE

	TEST SET G-mean (%) SCORES		
MODEL	NO RESAMPLING	SOUP (k=7)	SMOTE(k=7)
Logistic Regression	**21.1**	**79.2**	**70.2**
Linear SVM	**19.8**	**70.7**	**65.5**
Decision Trees	57.0	47.9	53.5
Random Forest	57.8	62.9	64.3
Gradient Boosting	**54.1**	**60.0**	**50.4**

UCI Glass Dataset
Test size = 30%

Best model: Logistic Regression trained from resampled dataset using SOUP

INSIGHTS

- Safety coefficients can be efficiently exploited in resampling techniques to improve classifiers.
- SOUP looks at the complex relations and similarities between classes and proposed a dataset with reasonable size.
- It can also work significantly better than SMOTE.

SALAMAT ☺

Questions?