Mixed-robROSE: a Novel Balancing Technique for Mixed-Type Datasets

Statistics and Stochastic Processes

RASOOL TABAN (rasooltaban@tecnico.ulisboa.pt)

Abstract

Imbalanced data, is a typical problem in classification, which may lead to biased models, enhanced by the fact that in majority of the problems the class of interest is the Minority class with small number of observations. Balancing techniques are accepted as promising and powerful strategies to handle the imbalance of data without being limited to any type of classifier.

Datasets coming from real problems regularly are mixed-type, i.e. dataset that has both Numerical and Nominal variables, which makes the imbalanced data problem complicated. In addition to difficulties of mixed-type datasets, presence of outliers can negatively impact the modeling of Minority class.

In this paper, we propose a robust balancing technique, called Mixed-robROSE - that utilize the method of robROSE and prepare it to oversample a imbalanced dataset with mixed-type variables. In this paper, we employ a novel strategy to oversample nominal variables by keeping the proportion of nominal variables' patterns the same. Simultaneously, for the numerical variables, atypical Minority class observations are down-weighted in such a way that potential outliers from this class are excluded in the resampling step.

The performance of the Mixed-robROSE is evaluated using several simulation scenarios, with different influential factors, and also using a set of imbalanced benchmark datasets with mixed-type variables. The results indicate the superiority of Mixed-robROSE over existing competitors.

Keywords: **Imbalanced Classification, Mixed-type data, Oversampling, robROSE, Robust Mahalanobis Distance.**

Proposed Method

First, the method identifies the potential outliers by using the robust Mahalanobis distance based on numerical part of data. Second, for non-outlying observation, all occurred patterns of nominal variables are found and the ratio of occurrence for each pattern is calculated. Third, for each nominal pattern, the matched observations are selected to perform the robROSE resampling for numerical part with respect to the related ratio. Details of the proposed method are presented in Algorithm 1.

Algorithm 1 Mixed-robROSE ($\mathfrak{X}_1$, h, α, eIR, l, k)

Input: $\mathfrak{X}_1 \in \mathbb{R}^{n_1 \times p}$ (Minority training set); h (MCD parameter); α (false alarm rate); eIR (expected imbalanced ratio); $l \in \{A, B\}$ (strategy to choose parents).

Output: $\mathfrak{X}_1^{\mathrm{B}}$ (Balanced Minority training set).

1. $\mathfrak{X}_1^{\mathrm{B}} \leftarrow \mathfrak{X}_1$.
2. For Continues part, estimate robustly the mean vector, $\hat{\boldsymbol{\mu}}_1$, and covariance matrix, $\hat{\boldsymbol{\Sigma}}_1$ using fast-MCD approach.
3. For Continues part, compute $\mathrm{MD}(\boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_1 | \hat{\boldsymbol{\Sigma}}_1)$.
4. Identify non-outlying Continues Minority observations where $X_{non-outlying} \leftarrow \{i : x_i \in \mathfrak{X}_1 \; and \; \mathrm{MD}^2(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_1 | \hat{\boldsymbol{\Sigma}}_1) > \tau_{1-\alpha}\}$.
5. For Categorical part, find all possible patterns from combination of all Categorical variables.
6. Compute ratio of occurrence for each pattern, and assign the needed frequency of each pattern in oversampled data.
7. **Resampling step:**
   (a) For each pattern, $X_{matched} \leftarrow$ find the matched observations from $X_{non-outlying}$.
   (b) Based on the strategy l, **Parent** observation $\leftarrow A$: either get the average of $X_{matched}$ or B : select an observation randomly from $X_{matched}$.
   (c) Generate a random sample from multivariate normal distribution with center of **Parent** observation and trimmed scatter matrix similar to *robROSE* algorithm.
8. Repeat the Resampling step until $\#\mathfrak{X}_1^{\mathrm{B}} < \lceil n_0/\mathrm{eIR} \rceil$.

Simulation Study

The simulation study is designed to investigate the effectiveness of balancing technique on mixed-type imbalanced data and compare it to the competitors under influence of 20% outliers toward Minority class.

Dataset

The dataset is generated as follows: i) numerical part, observations are generated according to a 3-dimensional Normal distribution; ii) for the nominal part, nominal variables are generated with 4 different levels of nominal labels.

Classifiers

Average result of 100 iterations over Four classifiers:
Decision Tree, **Adaboost**, **Random Forest**, **KNN** with Huang distance

Competitors

Four existing balancing technique: **SMOTE-NC**, **SMOTE+Gower**, **SMOTE+huang**, **ROSE**, and **Original** imbalanced data.

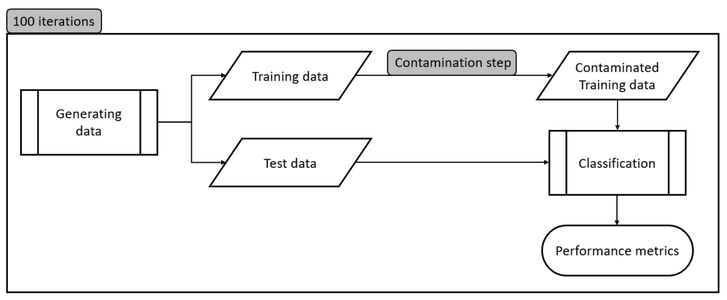

Figure 1: Flowchart of the experimental setup

Results

To perform a fair comparison in case of Imbalanced data, we used ***F1-score*** of Minority class that represents harmonic mean of ***Precision*** and ***Recall***.
Solid lines represent the case without outliers, and dashed lines represent the case with the influence of 20% outliers.

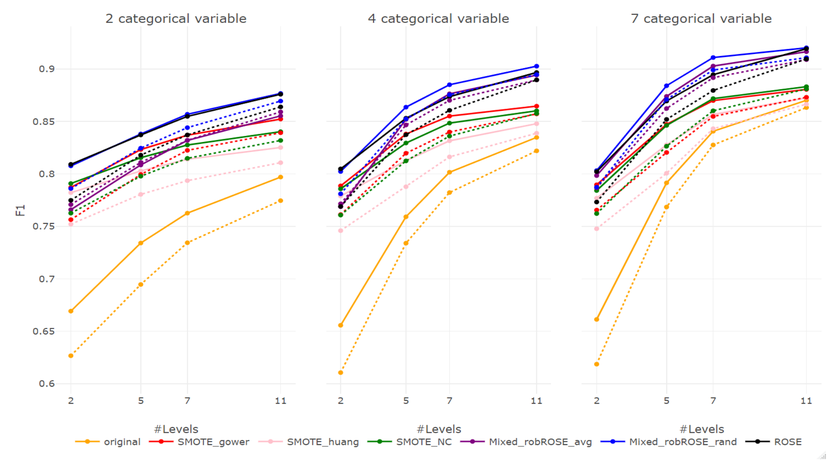

Figure 2: Average F1-score of Minority class

Conclusion

Our proposed method shows strong performance to handle the imbalanced data problem with/without presence of outliers. It outperforms all the competitors and shows the best results in all cases.

In addition, the statistical comparisons between the methods verify the superiority of Mixed-robROSE method (with random selection strategy) over its competitors.