Data mining Project

Research Paper

Harshit C. Gurav

2020CT04539

① **Objective:**

Class imbalance occurs in dataset due to which it is difficult to train classifiers on data as they become biased towards a set of classes as a result leading to reduction in classifier performance. This issue can be solved by Synthetic minority Oversampling Technique (SMOTE) which reduces class imbalance. Here we going to use modified version of SMOTE known as Outlier-SMOTE where each data-point is oversampled with respect to its distance from other data-points. Here main aim is to reduce the False negatives

② **Data Pre-processing.**

Dataset consists of Null and NaN values. Datasets were taken from University of California containing imbalance ratios (minority: majority) ranging from 1:9 to 1:40. Features with > 90% of null values were removed. Variables with zero variance were removed as they do not have any significance. After cleaning and seperating, minority samples feed to the algorithm of Outlier-SMOTE. Each sample is assigned probability weight p in range (0,1) for constructing oversampling matrix. Euclidean matrix is generated using minority data samples with n samples and w features. matrix was normalize by indicating apt amount of oversampling required for the samples.

③ **Data Mining**

Binary Logistic regression classification has been used and multiple oversampling rates to expeiment on its performance.

- Cleaned dataset is fetched & divided into majority & minority classes.
- Further classes are split into 90% and 10%. each of majority & minority set.
- Combining 90% of majority & minority & use it as training set, similarly, using 10% of minority & majority as validation set.
- Training the above on classifier.

(4) Metrics.

Confusion matrix was used which indicates the amount of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Precision, Recall and F1-score was used to evaluate the performance.

(5) Visualization :

SHAP (Shapely Additive Explainations) method is used to visualize the readings. It is basically game theoretic approach to explain the output of any machine learning model. This technique was used as an orthogonal validation of their model. SHAP denotes which features are important or which features are dominant while giving the results. It's values are calculated by taking average of marginal contributions of all features. SHAP repeatedly changes each variable's value and sees how it would affect the result and plot the graph. It's goal is to explain prediction of an instance 'x' by computing contribution of each feature of prediction. X-axis contains the impact of feature on result and Y-axis has all features that have been observed.