Outlier SMOTE Algorithm Research

Name: Meenavilli Sri Sai Nitin Raj Vamsi
BITSID: 2020CT04508
Group: 12

Objective:-

- Class imbalance is a common issue in every Data set. Due to this data becomes biased to specific classes, training classifiers is becoming difficult, hence effecting its performance.
  - => this issue is tackled by using over(or) Under Sampling Techniques, the one that which stood above all is the SMOTE algorithm, which generates synthetic Samples of each minority class by oversampling ~~of~~ each minority data ~~Samples~~ point by Combination of existing minority class ~~Samples~~ neighbours.
  - => this may lead to overlapping of minority data Samples. To overcome this issue authors come up with a modified version of SMOTE known as Outlier SMOTE.
  - => Where in each data point is oversampled with respect to its distance from its other data points
  - => the data point which is farthest from the data point is given high priority and is oversampled more than its Counterparts.

Data Pre-Processing:-

- => Cleaning the dataset by removing features with >90% null values, variables with Same ~~zero~~ variance, and Combing Columns by taking row wise sum identify the minority Samples, and pass them to OUTLIER SMOTE.
- => this will return an euclidean distance matrix of nxn size, for n Samples and m features.

for constructing Oversampling Matrix each Sample is assigned probabili weight P(0 to 1).

=> the ~~given~~ the Sum of Columns gives the distance of Samples from minority classes

=> Now the oversampling matrix may contain all numericals. So normalizing it

=> Rounding off values of normalized matrix upto 2 decimal points.

Data Mining:-

=> Using Binary logistic Regression classification algorithm and gauging its performance by using different Sampling rates.

Metrics:-

=> the main goal of outlier SMOTE algorithm is to reduce false possitives. So used Confussion Matrix, Precision, recall, F-measure

=> Precision = Correctly classified Possitive / Total possitive Samples.

=> Recall = No. of " " " / No. of Observations in possitive class.

=> F-measure = Harmonic mean of Precision & recall.

Visualization:-

=> the Authors used ~~a~~ SHAP algorithm which is a Game theoretic approach model.

=> Denoting which features are important or which features dominant

=> Prediction of an instance X by Computing the Contribution of each feature to the prediction

=> the X-axis contains the impact of features on result (-ve means less impact, +ve is high

=> the Y-axis has all features that have been observed.