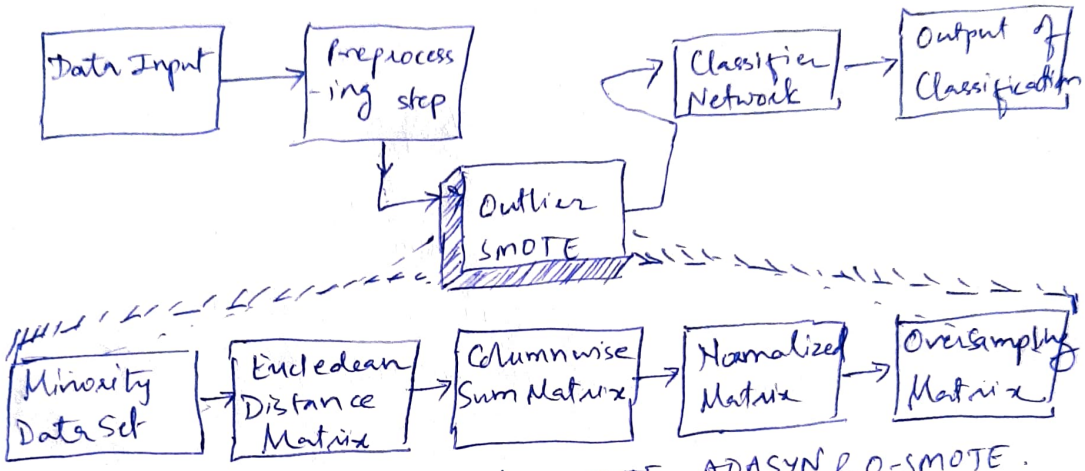
Data Mining : Project :
Research Paper

Ketan Mamtora
2020 CT 04524

(1) **Objective:** Improved over sampling method to create new and near accurate synthetic data based on existing imbalanced data set. It adapts existing SMOTE (Synthetic Minority Oversampling Technique). The proposed method is called Outlier-SMOTE. It will give more importance to outliying & remotely placed samples.



→ Compare the results against SMOTE, ADASYN & O-SMOTE.

→ Overall goal is to reduce the chance of false Negative by using over-sampling technique for each data point. Using this technique on covid-19 data set to achieve following benefits:

(a) Eliminate tests of people, less likely affected by covid-19

(b) Use less resources & optimal use of available resources

(2) **Data Preprocessing :-** Use of publicly available data set with benchmark of imbalance ratio of 1:9 to 1:40 (University of California), then use covid-19 data set available from hospital in brazil (covid tests & Lab test results).

Covid data set has total 5644 records, of which only 553 records are covid positive ⇒ 1:9 minority to majority ratio.

(a) The data with Null & NaN with 90% will be removed

(b) '0' variance data will be removed

(c) '19' variables will be processed.

(3) Data Mining & Metrics :-

(a) Perform data mining activities in two phases :
   (i) Phase 1 : Sample data set → Training → Validation [K rounds]
   (ii) Phase 2 : Cmid data set → Training → Validation

(b) Use oversampling (OS) rate between 100% to 500%

(c) Use k-fold cross validation method (k=10), where k-1 (=9) data set for training & kth data set for validation.

(d) Each data set of all (k) split into 90% + 10% of majority & minority sets,

(e) Combine 90% of majority & 90% of minority set for training & 10% majority & 10% of minority for validation.

(f) Use logistic regression & Random forest classifier on the oversampled data set.

(g) Calculate Recall, Precision & F1-score for all OS rate.

(h) Compare these outcomes for SMOTE, ADASYN & O-SMOTE algorithm.

(4) Visualization :

(a) Use barchart to verify majority & minority samples on the original data set.

(b) Tabular comparison of Recall, Precision, F1 score for all the different over sampling data set. (100% to 500%).

(c) Use of SHAP library to visualize impact of variation in the output model with change in values of individual features.

(d) Summarize the interesting observations through correlation matrix & SHAP output.