PR2: Image Classification

Name
Varun Shah

Student ID
010823657

Rank
23

F1-Score
0.8168

Feature Selection
Performed following steps for data preprocessing:

1. Read data from “train.dat” and stored each document in the list lines. Same way read data from “test.dat” and stored each document in the list lines_test.

2. Classes are read from train.labels file and stored in list classes where as train data is read from train.dat and each document of image data is stored in a lines_train list

3. Train data and test data are standardized using StandardScaler functionality of sklearn.preprocessing library

4. Dimensionality Reduction is performed on the standardized train and test data sets using SVD

5. As current dataset it imbalanced, Oversampling is performed on the train data and respective target class using SMOTE technique

Classifier Model Development
Performed following steps for Classifier Model Development:

1. Classification model developed with K Nearest Neighbors Classifier used with 4 n-neighbors (KNeighborsClassifier from sklearn.neighbors)

2. Classification model implemented using KNeighborsClassifier with 4 object to classify method

3. Model trained by calling fit method of KNeighborsClassifier and passing X_train and X_test as parameters

4. As the model trained, pred method of KNeighborsClassifier called to predict classes of y_train (test data set)

Methodology:

- Fetched train, test and train.labels data from file and stored them into list_train, list_test and classes respectively. Classes represent target list and correspondent list represents that class. As per training data there are 11 classes.

- Using StandardScaler standardized train and test data.

- As the given data set's dimensionality is very high, it is highly recommended to reduce the dimensionality, which would provide best result and les execution time training and testing using classifier. In this program, Linear Method SVD (Singular Value Decomposition) is used with following parameters:
  - n_components: 48 (Reducing Dimensionality to 48 gives most efficient performance)
  - n_iter: 10
  - random_state: 42

- As the dataset is imbalanced, training model could be difficult. So, oversampling is performed with following parameters:
  - ratio: "minority" as it is necessary to resample data such that model can be trained on balanced dataset
  - k_neighbors: 2

- K Nearest Neighbor Classification Model used for prediction. Following are the parameters set for training model and prediction.
  - n-neighbors: 4 as K Nearest Neighbor algorithm gives best performance for current dataset

- Classifier trains model with using resampled data:
  - X_resampled: Resampled Train data
  - y_resampled: Resampled target classes

- Trained model is applied to test data X_test_svd and classes predicted for respective image data