**Program 2**

**PR2: Image Classification**

**By Chaithra Lakshmi Sathyanarayana**

**Rank: 5**

**F1-Score:0.8559**

**Approach**

Training data is extracted image data. Since, it has 21186 records and 887 dimensional vectors, feature selection technique has been used to reduce the number of features. The data is highly-imbalanced, oversampling was done to handle the same.

1. Three files- train.dat, train.labels, test.dat for training and testing. There are 11 labels from 1-11 used for labelling.
2. Split the train data into train and test data (80:40)
3. Standardize the data by scaling
4. Use feature selection to remove low-variance data
5. Oversampling to handle the imbalanced data.
6. Train the model and calculate the F1-score. Knn-nearest neighbors classifier, Random forest and Extra-Trees classifiers were used.
7. Using the classifier having best-f1 score, retrain the model. Apply this model on the test data from “test.dat” file and generate the labels.

**Methodology**

1. Data Analysis: When I analyzed the data, I found data to be heavily imbalanced, high dimensionality and dense.
2. Split train-test: Cross-validate the data by splitting training data as training and test data in the ratio 80:20. The split-test-data will be used to calculate f1-score of different classifiers.
3. Preprocessing: To scale and standardize the data, StandardScaler() from sklearn library has been used. Scale the test data as well.
4. Feature Selection: Variance threshold feature selector is used. VarianceThreshold() from sklearn performs feature selection. Variance threshold feature selector removes all low-variance features. This reduces the dimensions and retains the features required to accurately classify.
5. Imbalanced data: Data is highly imbalanced was resampled. Oversampling was done using SMOTE. K-nearest neighbors with k=3 was used to perform the resampling.
6. Use data from previous step to train and test the split-data using different classifiers.
7. Knn-classifier: Knn-classifier was used to train the data, using 8 nearest neighbors. 8 neighbors were used since, 8 gave the best f1-score in the previous trials. KNeighborsClassifier from sklearn library is used to train the model. F1-score is 0.5930785914706561
8. Random Forest Classifier: Random forest classifier gave a f1-score better than knn-classifier. RandomForestClassifier from sklearn library is used to train the model. F1-score is 0.6534632166635422.
9. Extra-trees Classifier: Extra-trees classifier gave the best f1-score of all 3. In this “class\_weight=balanced” is used because this parameter adjusts the weights of different classes in y automatically. This feature also helps in handling the imbalanced data. ExtraTreesClassifier from sklearn library is used to train the model. F1-score is 0.7243319131232646.
10. From the different f1-scores, Extra-trees has better F1-score. So, extra-trees classifier will be used to train the model and for predictions.
11. The training data that was read from the file initially is used to retrain the model using Extra-trees classifier.
12. While retraining, the entire training data from train.dat file is used. Therefore, standardization and over-sampling techniques are applied to the training data once again.
13. Using Extra-trees classifier, the data in trained. The new model is used to predict the classes for the test data from file and uploaded.

In the previous clp-uploads, LCA, PCA have been used to perform dimensionality reduction. I found the f1-score improved when I used variance threshold feature selector.

In previous clp-submissions - SVM classifier was used to train the data. Although, SVM had high f1-score, it was very slow. Compare to SVM, f1-score of extra-trees classifier was comparatively higher and fast. So, SVM was not included in this attempt, to reduce the time taken to train the model.

**Conclusion:**

1. For high-dimensional data, dimensionality reduction and/or feature selection help in improving the prediction.
2. When data is highly imbalanced, oversampling, under-sampling or both increase the prediction accuracy for minority classes.