Image Classification

Name: Sneha Vadakkemadathil **SJSU ID:** 011810721
Leaderboard Rank: 1
F1-Score: 0.8442

PROBLEM STATEMENT:
Develop a predictive model that can determine, given an image, which one of 11 classes it is. The objective is to implement feature selection/reduction technique and experiment with different classification models.

APPROACH:
I experimented with various dimensionality reduction techniques, classification algorithms, and resampling methods as the object classes were heavily imbalanced. The training data which consists of 21186 image records, after applying various pre-processing methods, was fed to the model to evaluate the performance, and the best model was chosen. This model was then used to classify the images in the test dataset into 11 different classes. The best model was arrived at by performing feature selection using ANOVA (Analysis of variance) F-value and then applying Extra Trees Classifier on the low dimensional dataset. Although KNN and Random Forest algorithms, after parameter tuning, performed well with respect to the overall f1-score, the individual recall/precision for each of the 11 classes was better while using Extra Trees Classifier.

METHODOLOGY OF CHOOSING THE APPROACH AND PARAMETERS:
Data pre-processing: I experimented with the following data pre-processing methods.

1. Scaled the feature values using the standard scalar to feed the data to a linear dimensionality reduction method like PCA.
2. Scaled the feature values using the min-max scalar, providing different feature ranges.
3. Converted the data to a sparse matrix format to perform dimensionality reduction using Truncated SVD.
4. Removed constant features from the data by finding Variance Threshold for experimenting with different feature selection methods.

I chose method 4 in my final solution as it worked the best with my feature selection method.

Resampling: I tried the below resampling methods to handle the class imbalance.

1. **Random over sampling** of minority classes with and without specifying class ratios.
2. **Random under sampling** of majority classes with and without specifying class ratios.
3. **SMOTE**

I noticed that the F1-Score reduced after resampling, and hence did not use it in my final solution.

Dimensionality Reduction: I experimented with the following DR techniques.

1. PCA after feature scaling with parameters ranging from 10 – 60.
2. Truncated SVD after converting data into a CSR matrix format.
3. Linear Discriminant Analysis with SVD solver.
4. Feature Selection using Random Forest feature importance.
5. Feature Selection using ANOVA F-value with parameters values 10 -70.

I used ANOVA F-value to select the 60 best scoring features to achieve dimensionality reduction in my final solution. I found this method and feature number to increase the recall of the rare classes.

Classification: The experimented with the following algorithms to classify the images.

Classification Algorithm	Dimensionality Reduction	F1- Score
KNN	PCA, Components=50	0.81
Random Forest	Feature Selection using Random Forest	0.829
GaussianNB	PCA, Components=48	0.51
VotingClassifier (Extra Tree + KNN)	PCA, Components=48	0.826
KNN	Truncated SVD, components = 50	0.78
Extra Trees Classifier	Truncated SVD, components = 50	0.81
Gradient Boosting Classifier	PCA, Components=50	0.71
KNN	LDA, Solver SVD	0.67
Extra Trees Classifier	Feature selection using ANOVA F value	0.844

For each of these classifiers, I tuned the hyper parameter values and compared the F1-scores. For example, for KNN, I used K in range (1,12). For random forest I used different class weights and number of estimators. As both extra trees classifier and KNN were performing well, I tried out a voting classifier with both these algorithms. I also performed 5- fold cross validation to test the models.

I chose Extra Trees Classifier in my final solution as it outperformed other algorithms with respect to class separation and F1-Score*. I experimented with different hyper parameter values and the below values gave the best results with my test data.*

- Random Seed:1
- The number of trees in the forest (n_estimators): 200
- The function to measure the quality of a split (criterion): Gini
- The minimum number of samples required to split an internal node (min_samples_split): 5
- Weights associated with classes (class_weight): Computed based on the bootstrap sample for every tree grown. (*balanced_subsample*)
- The number of features to consider for the best split (max_features): All

SUMMARY:
The best F1-Score of 0.8442 was achieved with ANOVA F-value feature selection and Extra Trees Classifier, without resampling.

Leaderboard:

Rank	F1 (on 50%)	User ID	Submission Count
1	0.8442	11810721	8
2	0.8422	11502985	14
3	0.8416	12424126	8