Kaggle Competition

SML
Priya Rajpurohit 2015073
Kaggle Public Score: 0.37333

❑ Methodology

Started with basic linear classifiers:

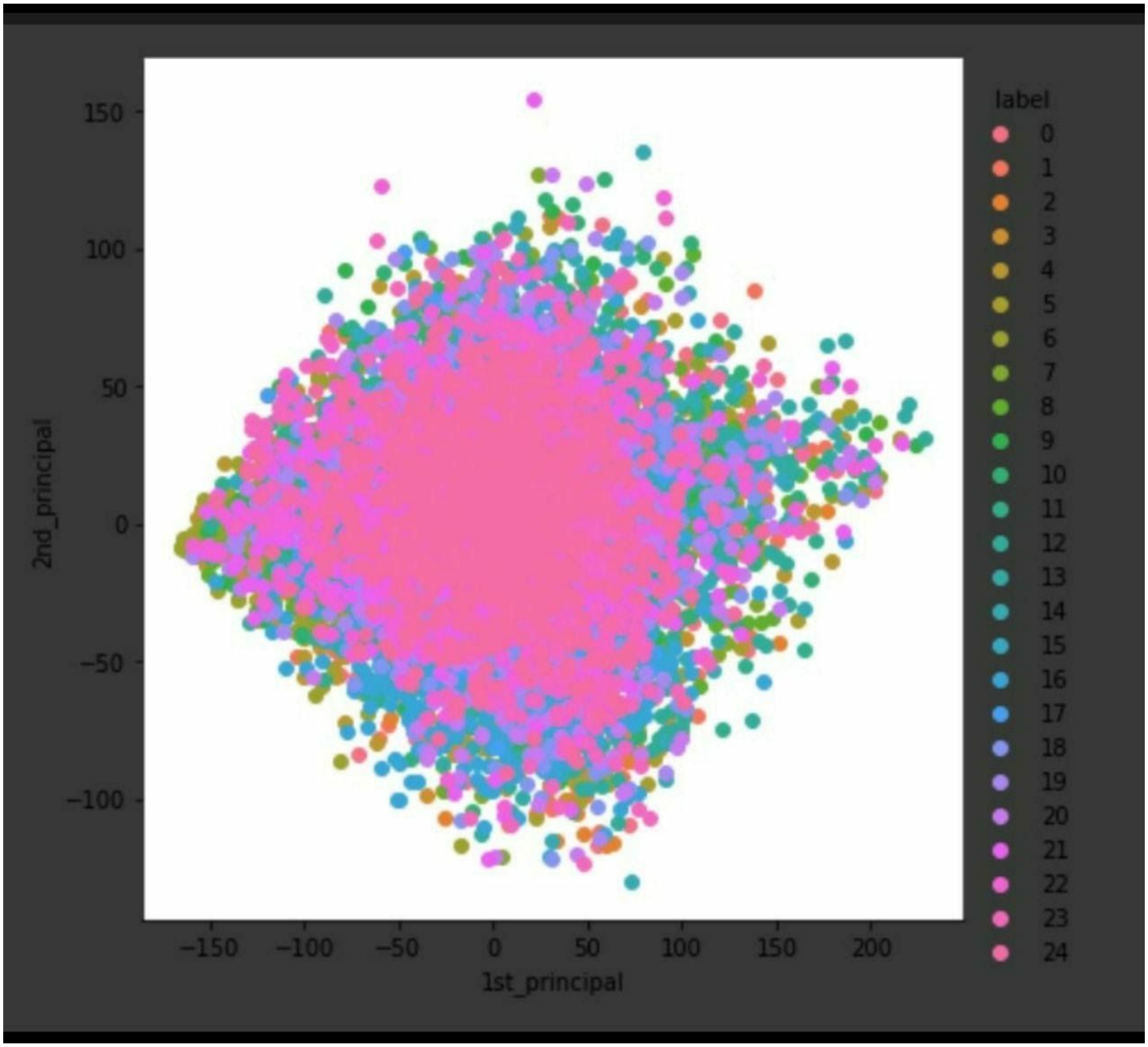
- Linear Discriminant Analysis (LDA)
- K-Neighbor
- Decision Tree
- Gaussian Naive Bayes
- SVM
- Random Forest

Applied them on images or their features using feature extraction methods mentioned below:

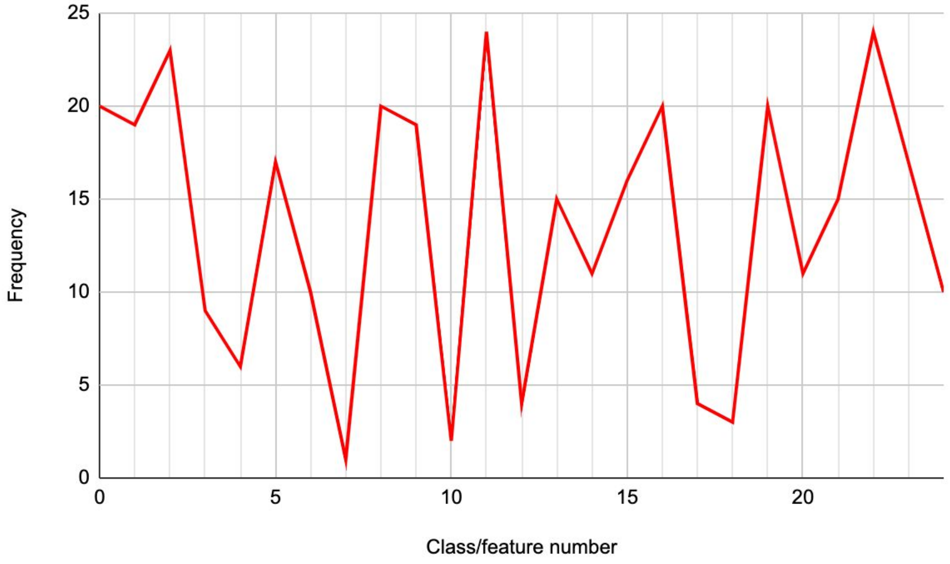
- PCA
- Sobel Filter
- Prewitt Filter
- Histograms of Gradients(HoG)
- HSV+HoG+BGR

Preprocessing of Image included quantile transform and standard scaler to normalize the features or image since that data is imbalanced.

Visualization of Data(PCA(n=2))



Imbalanced Data:



Handled by: Using ensemble classifiers, cross_val_score, StratifiedShuffleSplit, using balanced weights(in case of tree classifiers)

Approach:

1. Basic Classifiers with no preprocessing, feature extraction to get an idea of what minimum accuracy is achieved
2. Few Classifiers having higher accuracy were then run with PCA, HoG, Sobel, Prewitt on Standard scaled features/images depending on the model.
3. Applied cross-validation(splitting train data into train+validation sets) to then classifiers with high accuracy.
   Used: cross_val_score(rf_class, X_train1, y_train1, scoring='accuracy', cv = 5/10), where cv=no.of folds
4. Then the ones with high accuracy were applied with a combination of features(HoG+HSV+BGR) while playing with the hyperparameters of classifier and feature extractors.
   Note: Models have been stored only if a particular threshold is crossed, after applying a classifier on the validation dataset.

❑ Observations and Experiments(Submitted on Kaggle)

Note: The reasons for Failure/Success also contributed to why a certain model was reached and used.

Note: S.No. is corresponding to files uploaded-model/python files.

S.No.	Model used+Processing done	Kaggle Public Score	Reasons for failure or success(F/S)
1	LDA	0.07333	F-Distribution is not normal(gaussian)
2	K-Neighbor	0.11777	F- Features are overlapping
3	Decision Tree	0.11777	F-Overfitting
4	GaussianNB	0.13555	S-works good on high dimensional data F-features are assumed to be independent
5	SVM	0.16666	S-works better with high dimensional data F- Does not work well in case of overlapped classes with unprocessed data

6	Random Forest (n_estimators=100)	0.21333	S-good performance on imbalanced datasets F-Less no of estimators used
7	Random Forest (n_estimators=200,max_depth=30)	0.22444	S-more no. of estimators F- data needs preprocessing/feature selection
8	Random Forest (n_estimators=200,max_depth=30) with cross_validation (80:20 split)	0.22666	S-Cross Validation increases the overall accuracy F-data needs preprocessing/feature selection
9	Random Forest (n_estimators=200,max_depth=30) with cross_validation (80:20 split) on PCA (n_components=50) *Standard Scaled features before PCA*	0.24000	S-Dimensionality reduction improves the training of the model F-feature selection needed
10	SVM(kernel=’linear’) on PCA(n=50) of HoG[orientations=8, pixels_per_cell=(16, 16),cells_per_block=(1, 1),block_norm= 'L2'] with Cross-Validation	0.25111	S-Model trained on HoG F-Dimension reduction decreases the accuracy since PCA components =50)

11	SVM(kernel=’linear’) on PCA(n=25) of HoG[orientations=8, pixels_per_cell=(8,8),cells_per_block=(1, 1),block_norm= 'L2'] with Cross Validation	0.25777	S-Model train on HoG and reduced no. of components of PCA to 25 with pixels per cell (8,8) for HoG F-Dimensionality reduction has negative effect on the model
12	SVM(kernel=’linear’) on HoG[orientations=8, pixels_per_cell=(16, 16),cells_per_block=(1, 1),block_norm= 'L2'] with Cross Validation	0.26444	S-Model trained on HoG features of images with no dimensionality reduction F-SVM gave lower accuracy, target features might still be overlapping
13	LDA on PCA of HoG [orientations=8, pixels_per_cell=(24, 24),cells_per_block=(2, 2),block_norm= 'L2']+HSV+BGR with Cross Validation *Quantile Transformed HoG+HSV+BGR features*	0.31777	S-PCA reduces dimensions of HoG for LDA, low overlapping of features+Quantile Transform used F-pixels per cell high in HoG

14	RF[n_estimators=200,max_depth=30, random_state=42,class_weight='balanced’,max_features='sqrt']on PCA(n=50) on HoG[orientations=8, pixels_per_cell=(16, 16),cells_per_block=(4, 4),block_norm= 'L2'] +HSV+BGR with cross validation *Quantile Transformed HoG+HSV+BGR*	0.32444	S-Random Forests works better on overlapping features +Quantile Transform used F- Under fitting wrt similar model using (20,20) pixel per cell
15	LDA on PCA(n=50) of HoG [orientations=15, pixels_per_cell=(20, 20),cells_per_block=(2, 2),block_norm= 'L2'] +HSV+BGRwith Cross Validation *Quantile Transformed HoG+HSV+BGR*	0.33333	S-Pixel per cell reduction on HoG increased the performance F-Overfitting since pixels per cells high
16	RF[n_estimators=500,max_depth=30, random_state=42,class_weight='balanced]on PCA(n=50) on HoG[orientations=8, pixels_per_cell=(20, 20),cells_per_block=(2,2),block_norm= 'L2'] +HSV+BGR with cross validation *Quantile Transformed HoG+HSV+BGR*	0.34222	S-Increased pixel per cell and decreased cell per block wrt last RF model F-Under fitting wrt similar model using (24,24) pixel per cell

17	LDA on PCA(n=50) of HoG [orientations=15, pixels_per_cell=(8,8),cells_per_block=(2, 2),block_norm= 'L2'] +HSV+BGRwith Cross Validation *Quantile Transformed HoG+HSV+BGR*	0.34888	S-Reduced pixel per cell of HoG Features F-Underfitting ,Since (16,16) pixels per cell shows maximum accuracy
18	RF[n_estimators=600,max_depth=30, random_state=42,class_weight='balanced’,max_features='sqrt']on PCA(n=50) on HoG [orientations=8, pixels_per_cell=(16, 16),cells_per_block=(2,2),block_norm= 'L2'] +HSV+BGR with cross validation *Quantile Transformed HoG+HSV+BGR With Stratified ShuffleSplit*	0.35111	S-Stratified Shuffle Split balanced the data F-Less no. of Pixels per cell used for HoG
19	RF[n_estimators=200,max_depth=30, random_state=42,class_weight='balanced’,max_features='sqrt']on PCA(n=50) on HoG[orientations=15, pixels_per_cell=(24, 24),cells_per_block=(2,2),block_norm= 'L2'] +HSV+BGR with cross validation *Quantile Transformed HoG+HSV+BGR*	0.35333	S-Increased pixel per cell wrt last RF model F-RF is not able to separate features as good as LDA in this case

20	LDA on PCA(n=50) of HoG [orientations=15, pixels_per_cell=(16, 16),cells_per_block=(2, 2),block_norm= 'L2'] +HSV+BGRwith Cross Validation *Quantile Transformed HoG+HSV+BGR*	0.37333	S-Optimal pixel per cell of HoG wrt last similar Model

❑Failed Experiments(Not submitted on Kaggle)

Note: some models were not submitted because the accuracy didn't cross a threshold which may vary, because the failed ones were trained in between the submitted ones

1. RF using Sobel features of Image segmented(background removed)

   Accuracy on validation-10.29

Reason for failure- Too many Features, majority pixels now have RGB value corresponding to function value 0, leading to imbalance.

2. SVM trained on Sobel features of Image

   Accuracy on validation- 21.6

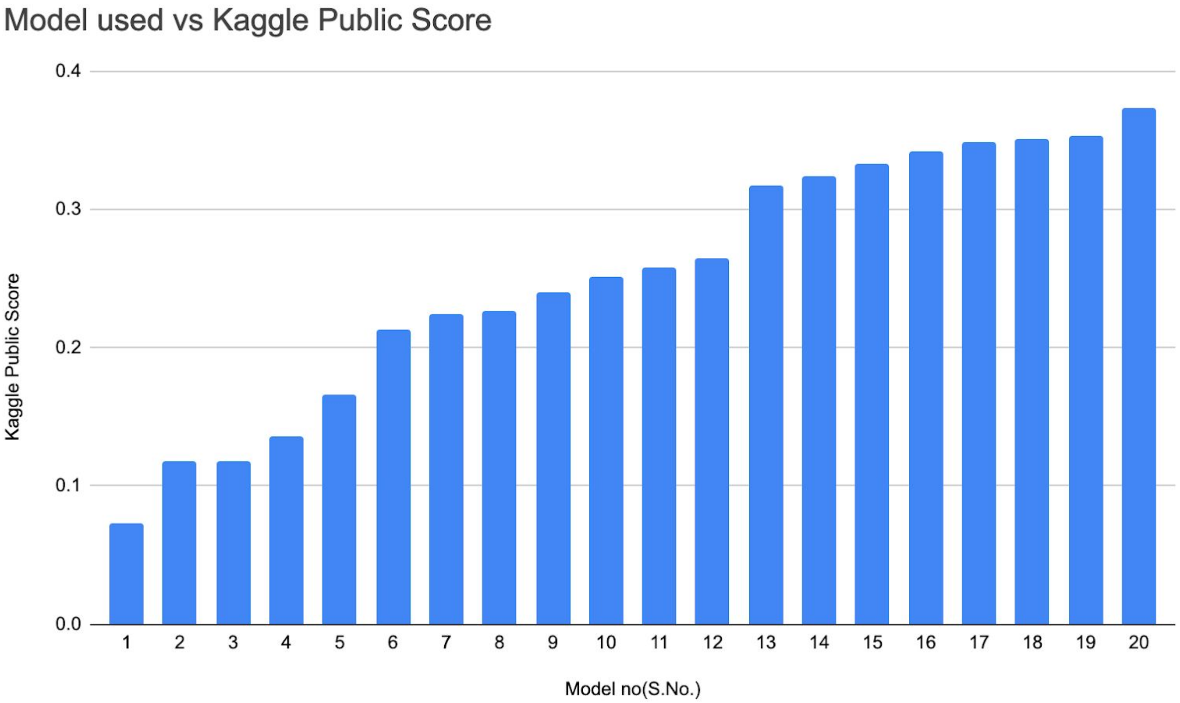
Reason for failure-Sobel filter works on the GRAY image, the gradient of edges is not captured as it is captured by HoG.

3. SVM trained on Prewitt features of Image

   Accuracy on validation- 21.5

Reason for failure-Prewitt filter works on the GRAY image, the gradient of edges is not captured as it is captured by HoG similar to the Sobel filter.

❑ Results



Amongst the models used, the model having LDA as classifier training on PCA(n=50) of HoG [orientations=8, pixels_per_cell=(16, 16),cells_per_block=(2, 2),block_norm= 'L2'] +HSV+BGR features(concatenated) with Cross Validation, where HoG+HSV+BGR features were used after undergoing Quantile Transformation.

❑ Analysis

❑ Dataset:

The dataset consists of 25 classes.
The training set contains 16000 images.
The testing set contains 1500 images.

❑ Data preprocessing:

Methods used on images(few based on prior12 knowledge from Digital Image Processing course):

1. PCA
   Reason to use it: for reducing the dimensions of features to be used for training the model, to reduce the risk of overfitting and reduce the computational complexity happening otherwise.
2. Sobel Filter
   Reason to use it: it gives a 2-D spatial gradient measurement on an image, highlighting regions of high spatial frequency that resemble edges. It gives a smoothing effect as well.
3. Prewitt Filter
   Reason to use it: It is a gradient-based edge detection operator and it has gradient features. Similar to Sobel.
4. Histograms of Gradients(HoG)
   Reason to use it: It gives the orientation histograms of edge intensity in a local region. It helps in detecting shapes present in the images. It is more descriptive than Prewitt or Sobel filters. It is also a good texture descriptor.
5. HSV+HoG+BGR
   Reason to use it: Using only HoG features for training wasn't giving a good accuracy on SVM or Random Forest. Hence, when HSV(Hue, Saturation, and Value), BGR(Blue, Green, Red), and HoG were concatenated to transform the dataset, better accuracy was obtained.

Methods used on features:

1. Normalization: Quantile Transformation
   Reason to use it: The function quantile_transform modifies the features to follow a uniform transformation. It spreads out the most frequent value and reduces the effect of outliers.
2. Standardization: StandardScaler
   Reason to use it: This scaler regulates the features by removing the mean and scaling to unit variance. Hence, if a feature has a variance that is higher than others, it might overshadow the objective function and make the estimator incapable to learn from other features correctly as it is supposed to.
3. Cross-Validation: Stratified Shuffle Split
   Reason to use it: It returns randomized stratified folds, which is a combination of best functions of Shuffle Split and K Stratified Fold.
4. Cross_val_score Calculation
   Reason to use it: Give a score by cross-validation to give a prediction of the model's accuracy.

❑ Classification
Basic Models:

1. Linear Discriminant Analysis (LDA)
   Reason to use it: The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix. It works better when trained on PCA transformed features when no. of classes is high.
2. K-Neighbor
   Reason to use it: The model performs learning based on the k nearest neighbors of each query point. When K neighbor was used, the data wasn't balanced. Hence, it did not perform well.
3. Decision Tree
   Reason to use it: Classifies datasets on the basis of decision rules. But the model is prone to overfitting if the depth of the tree is not limited. If the data isn't balanced, it creates a biased tree.

4. Gaussian Naive Bayes
   Reason to use it: even when it works better on Multiple class datasets, it assumes the features to be independent, which is not the case with the given dataset. Hence, it works poorly in this case.
5. SVM
   Reason to use it: It handles high dimensional data pretty well and uses a subset of features for the decision function. It's predicted accuracy increased when it was trained on HoG features, since it highlights shapes, in low dimensionality.
6. Random Forest
   Reason to use it: It fits a number of decision tree classifiers (Bagging of decision trees) on various sub-samples of the dataset and uses averaging to enhance the predictive accuracy and handle over-fitting. It also works on variance reduction.

Inferences

While trying out various experiments using the mentioned approach, it seemed that Random Forest, SVM will be able to classify better and other techniques. But in the end, the maximum accuracy was achieved by applying LDA classifier on PCA of transformed features used mentioned above. This may have happened because - (i)HoG when tuned and used in combination with HSV and BGR features, was able to describe the shapes, textures of images better, hence helping classifiers to train better, (ii)Because the features extracted from images were quantile transformed to fit Gaussian distribution, and LDA works much better when trained on Gaussian distribution, known from theory, and (iii)PCA helped to filter out features, reducing dimensions, leading to even better training of LDA classifier.

Link to Google Drive Folder having models