Incremental Dimensionality Reduction for Global Feature Set Extraction

IIT2015032 Rohan MR
IIT2015039 Nishant Verma
IIT2015042 Raghav Saboo
IIT2015045 Harsh Vardhan

Problem Statement

To generate global feature set of handwritten digits by incremental dimensionality reduction for fast one class classification.

Steps followed in the Project

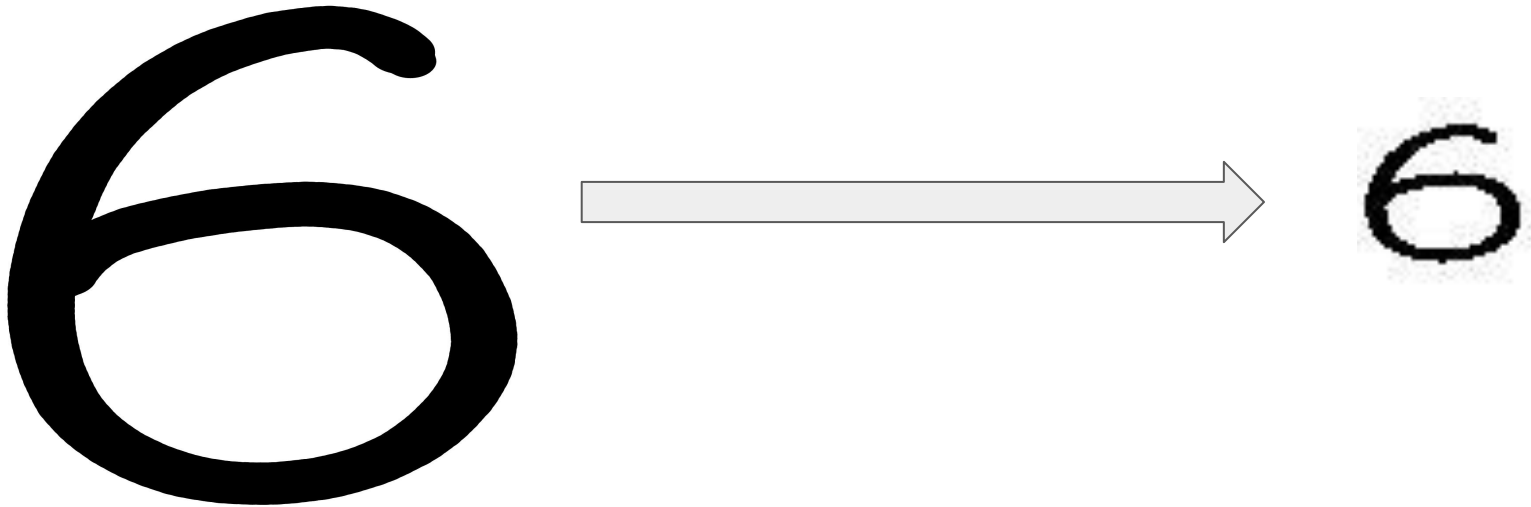
1. Generate Dataset of handwritten digits
2. Feature Extraction
3. One class classification
4. Global Feature set generation

Steps used in Dataset Generation

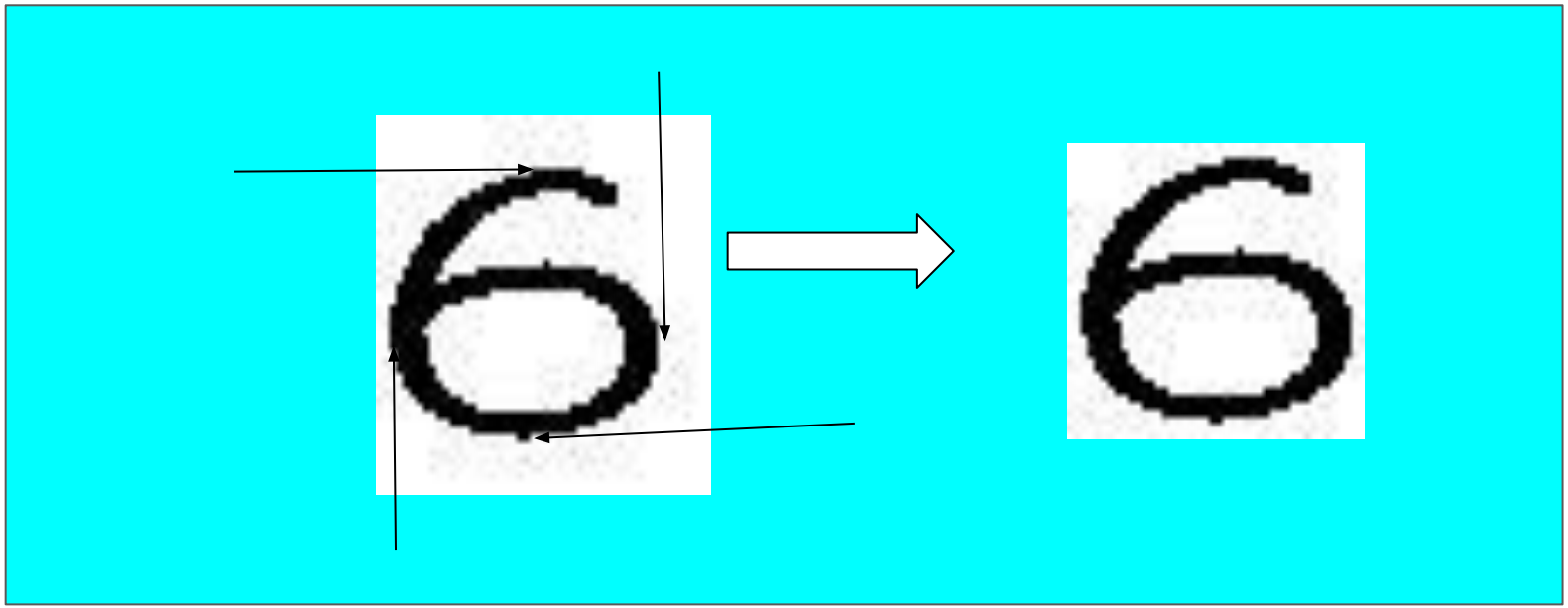
1 Create image data

We have generated data of handwritten digits 0-9.

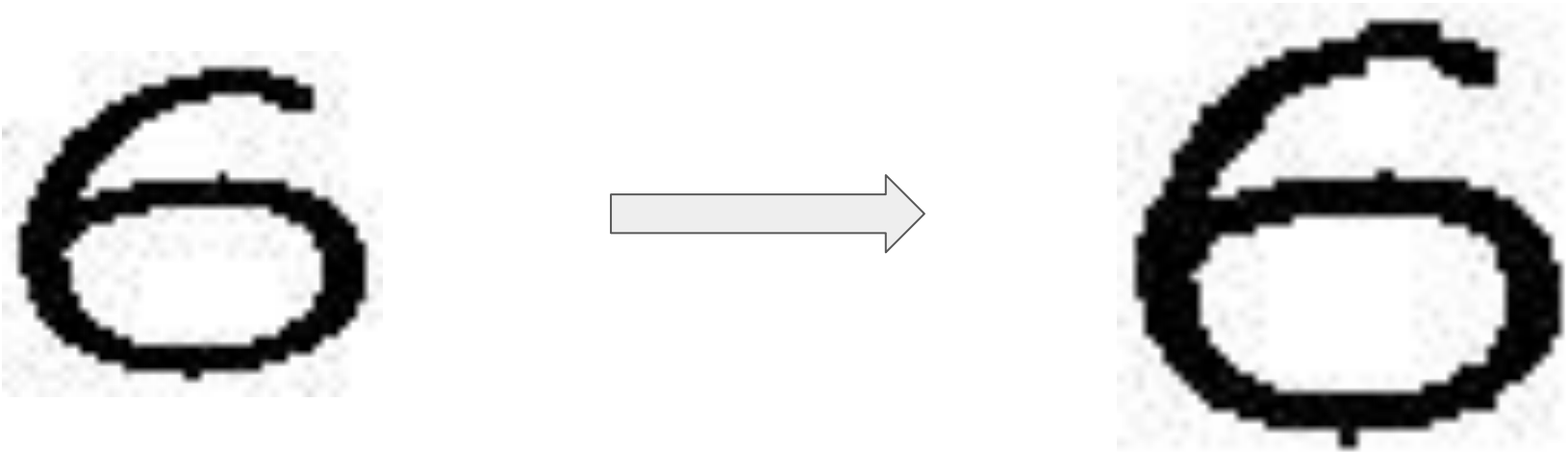
2 Resize the original image to 50X50 pixel image.



3 Digit Extraction from the image.



4 The cropped image was again upscaled to 50x50 pixel image



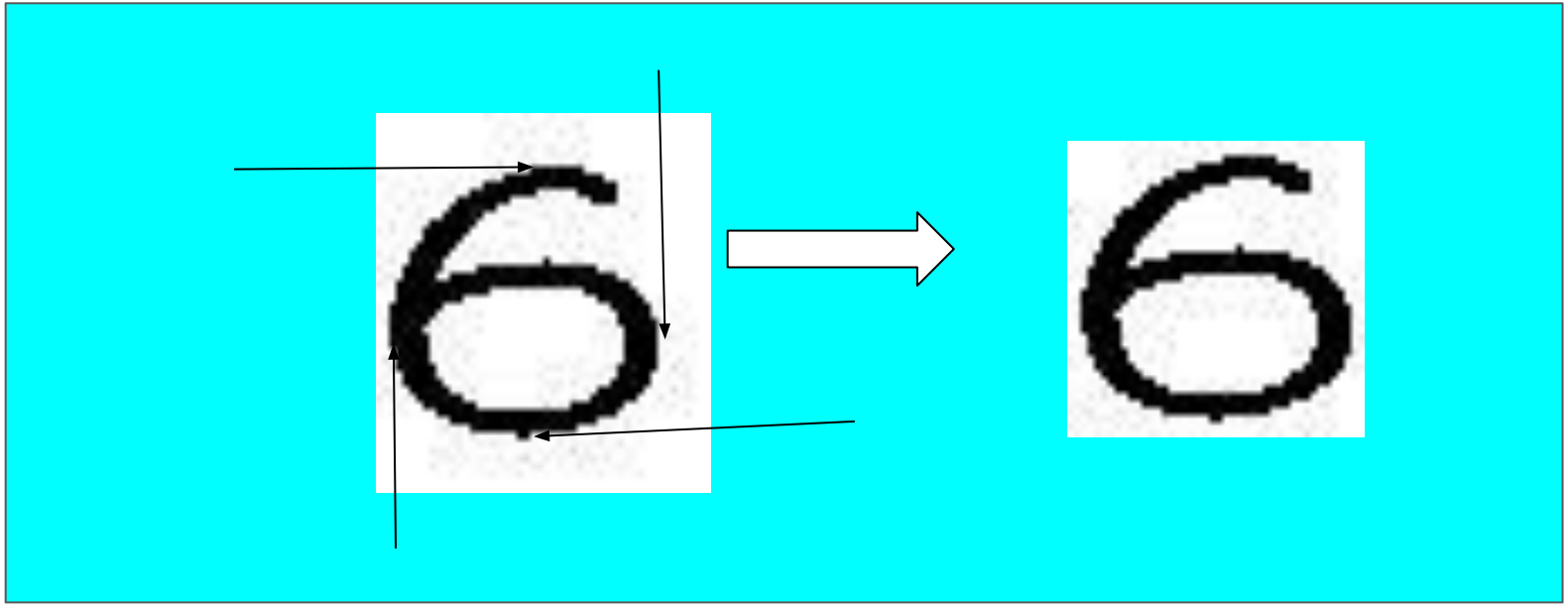
Feature Extraction from the generated images

Three approaches used -

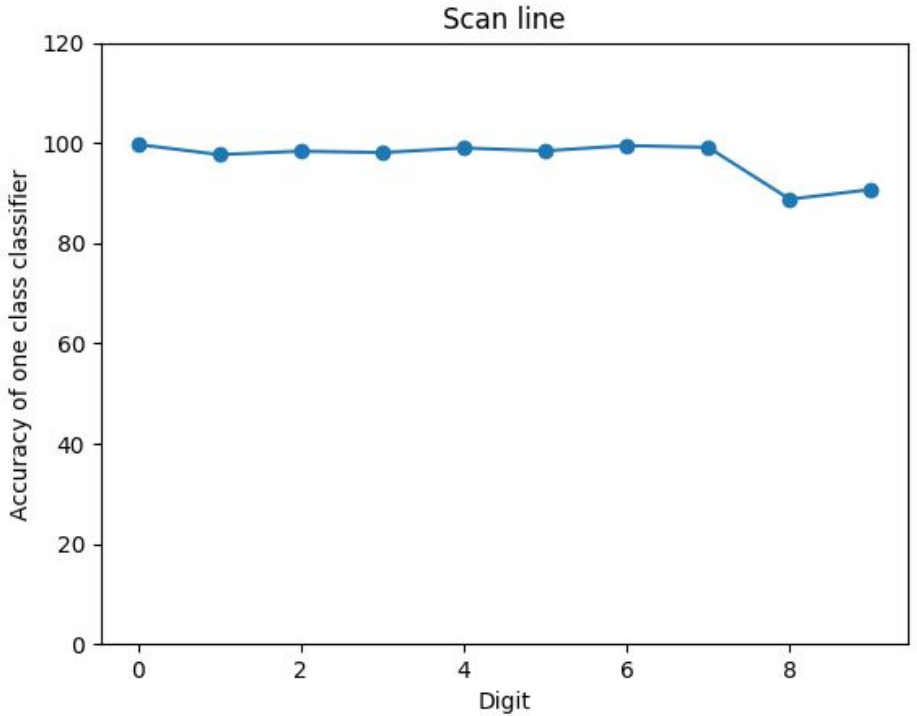
1. Scan line
2. Run length
3. Zone based feature extraction

Scan line

No of features - 2*(m+n) for an mxn image

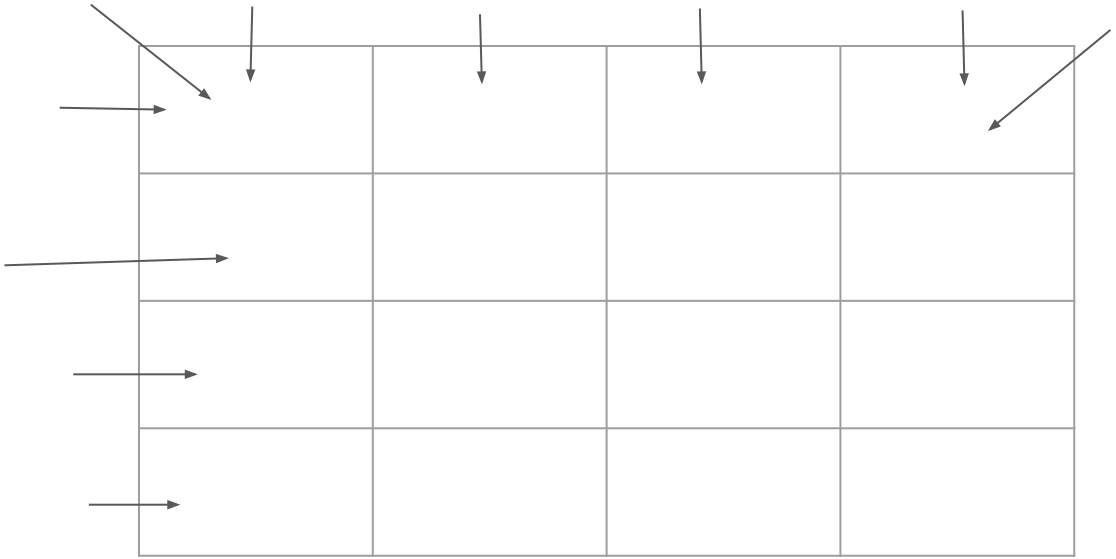


Accuracy obtained

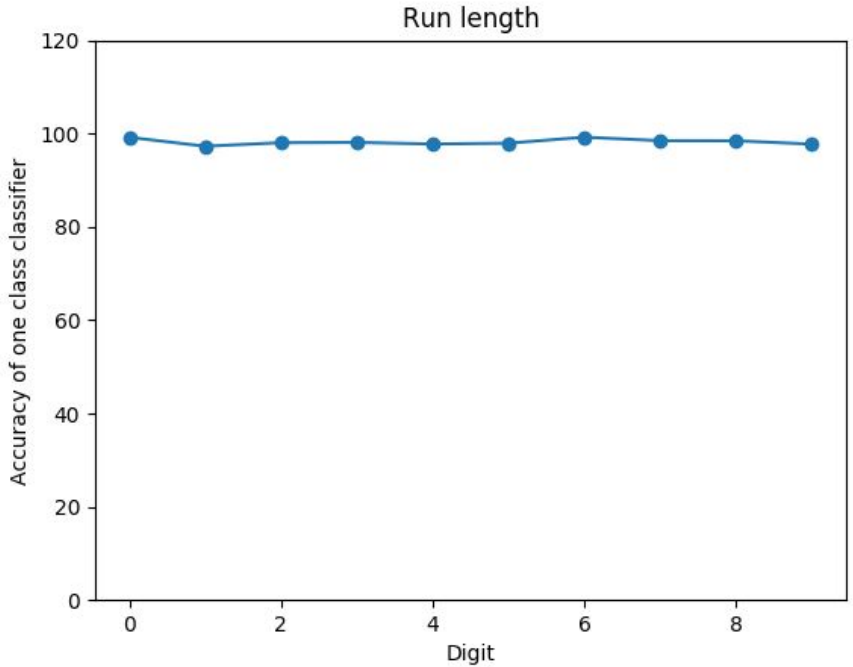


Run length

No of features - 5*(n+m) - 6 for an mxn image

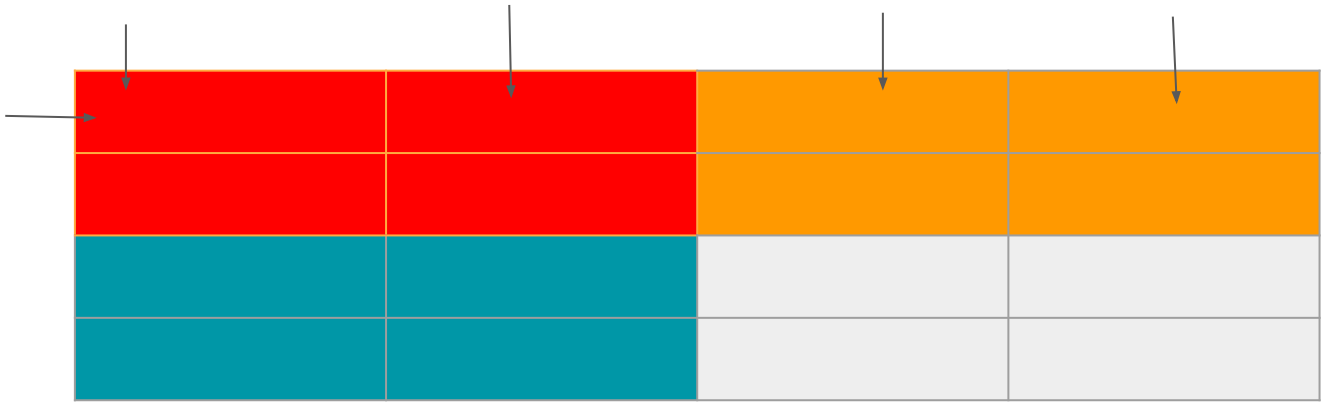


Accuracy Obtained

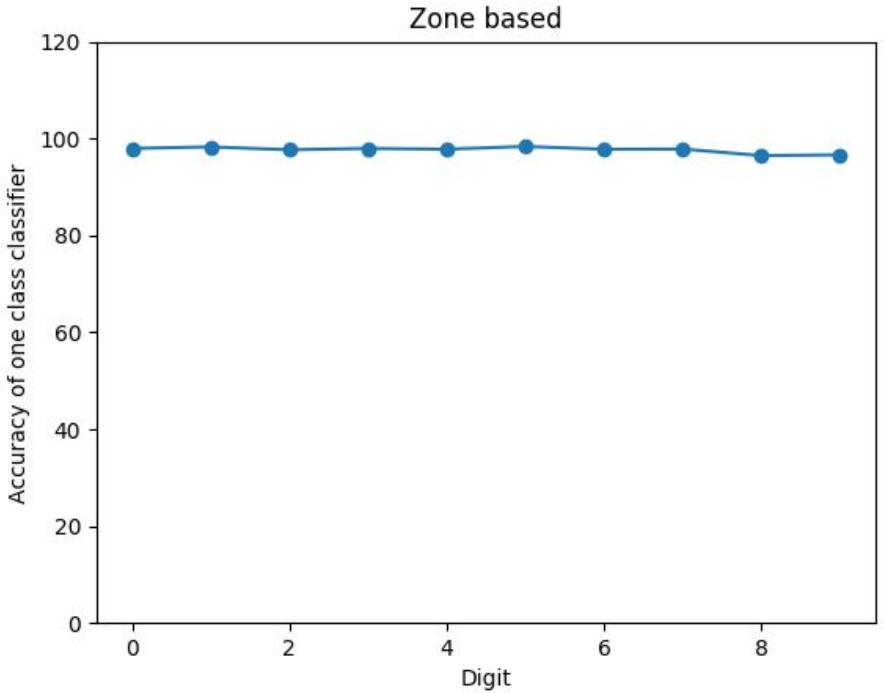


Zone Based

No of features was 1000 for 50x50 image with 10x10 sized zones each zone having 40 features thus 25 zones x 40 features in each zone



Accuracy Obtained

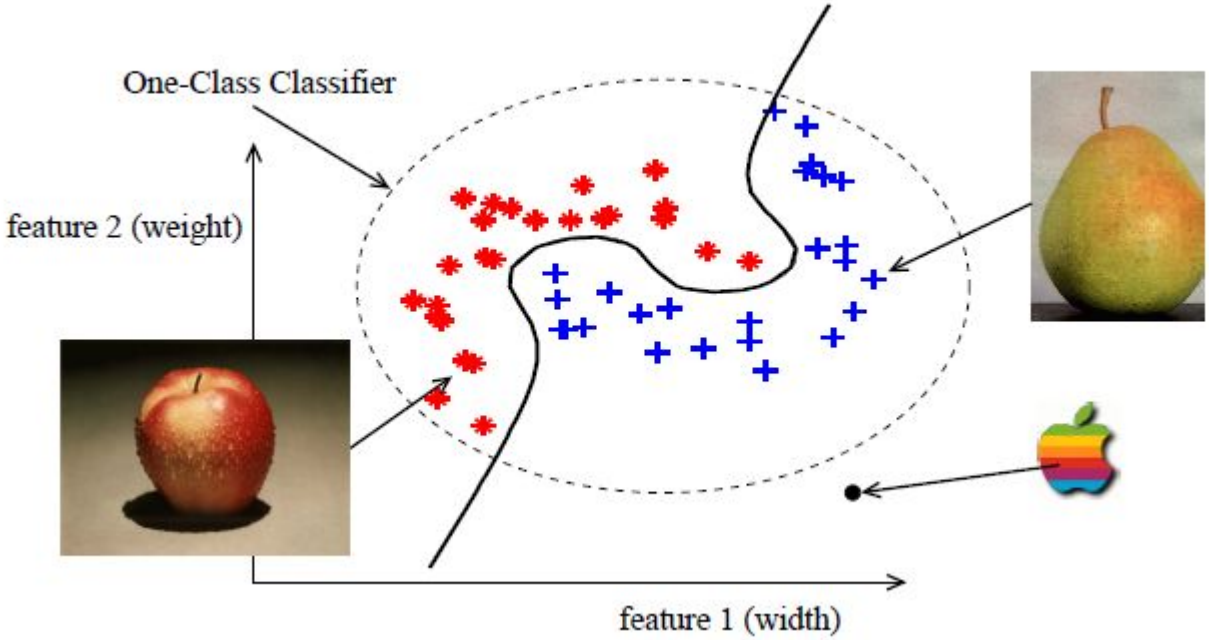


One class Classification

In machine learning one class classification tries to find objects of a specific class by learning from only objects of that class.

Many applications can be found for example novelty detection,outlier detection.

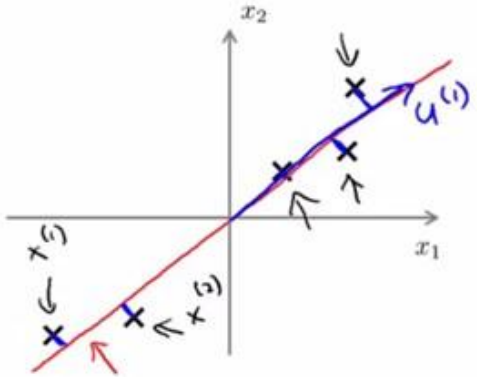
Ex - Machine Monitoring System ,etc.


One-Class Classifier
feature 2 (weight)
feature 1 (width)

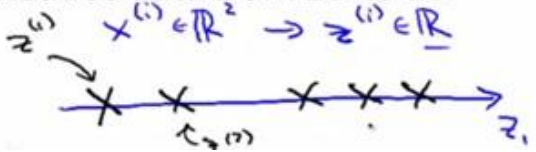
Principal Component Analysis

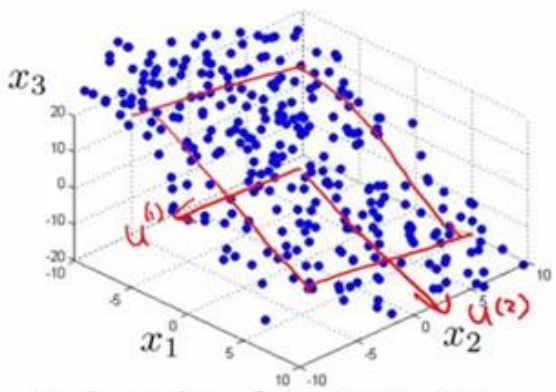
The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent.

Principal Component Analysis (PCA) algorithm



Reduce data from 2D to 1D





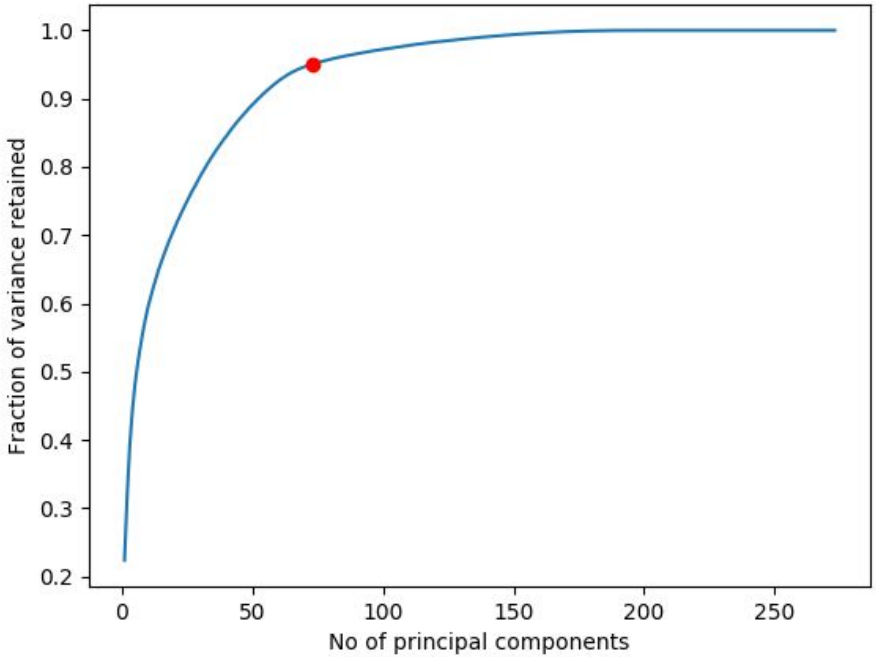
Reduce data from 3D to 2D

How to choose the no of Principal Components??

1. Pick a threshold that gives the fraction of variance retained in the original data.

2. Choose the minimum no of components that satisfy the threshold.

3. This is done by iteratively running the loop from 1 to no of features in the original dataset.


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Fraction of variance retained
0
50
100
150
200
250
No of principal components

Pearson Correlation Coefficient PCC

In statistics, the Pearson correlation coefficient , is a measure of the linear correlation between two variables X and Y. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

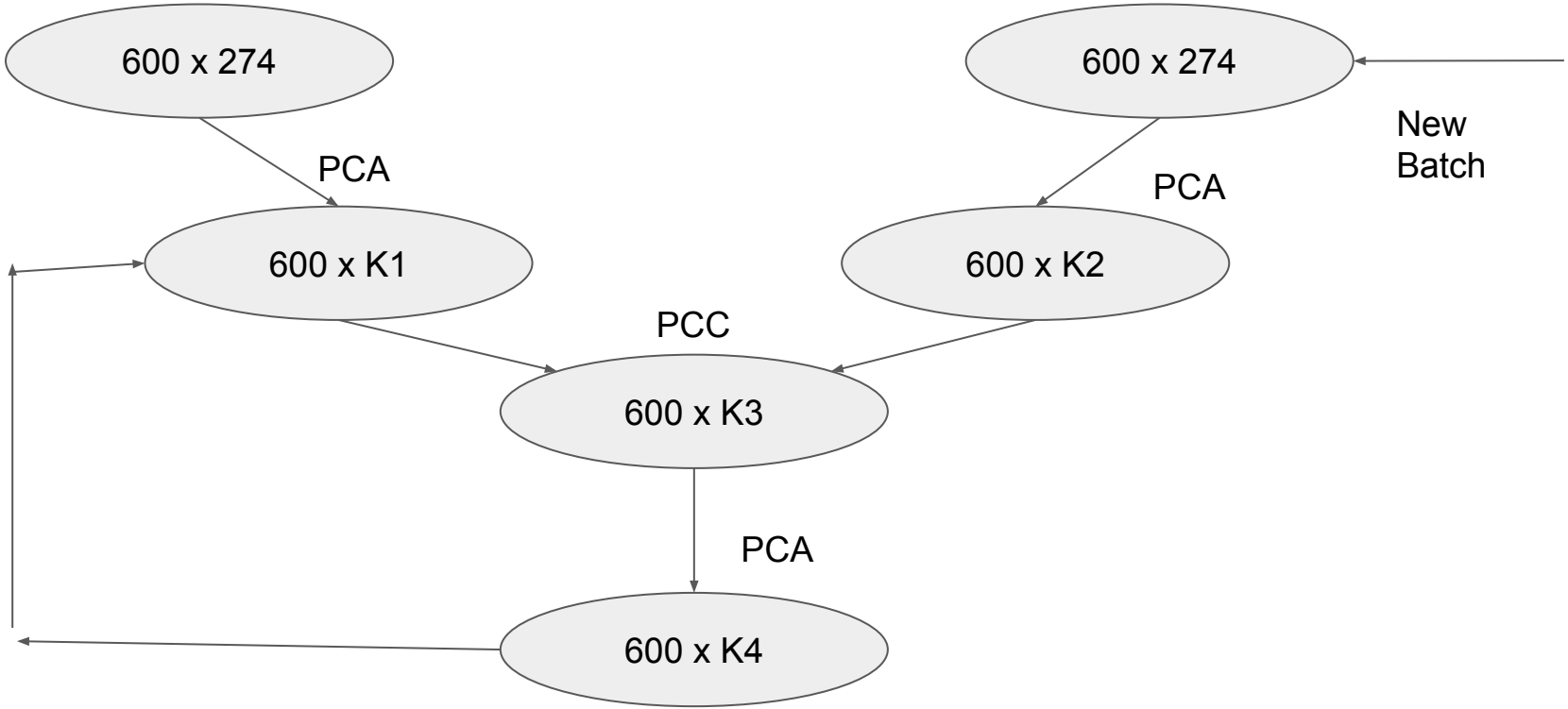
where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

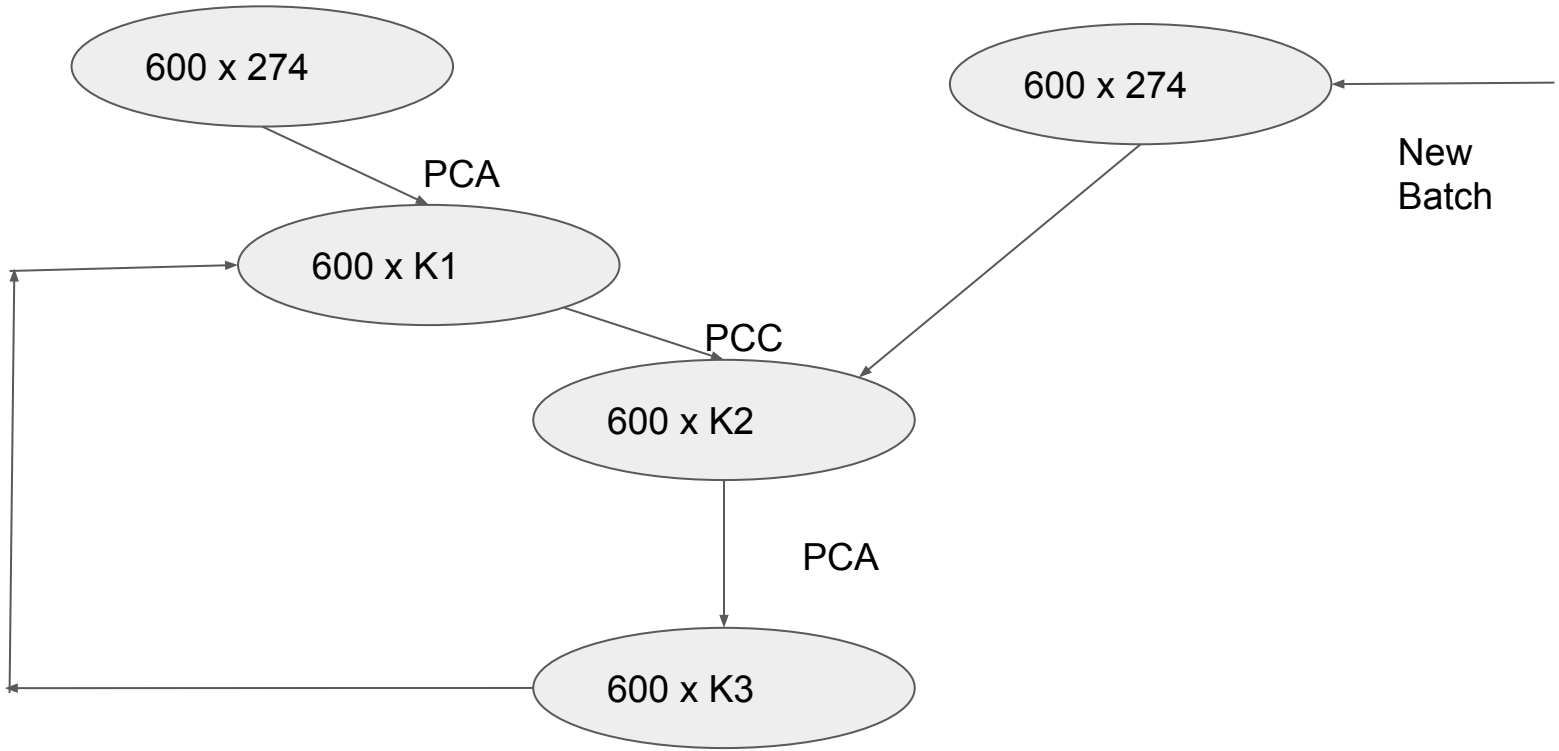
Techniques for generation of Global Feature Set

- Incremental double PCA with PCC
- Incremental single PCA with PCC
- Single PCA

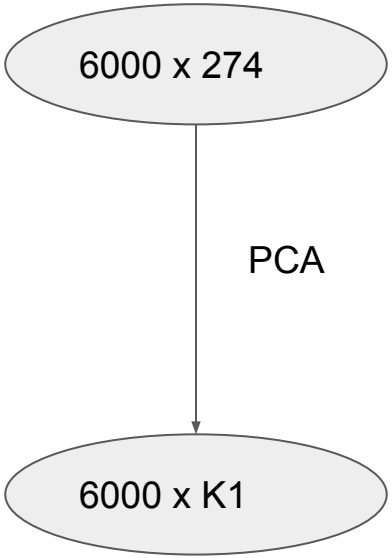
Incremental Double PCA with PCC



Incremental Single PCA with PCC



Regular Single PCA



Comparison

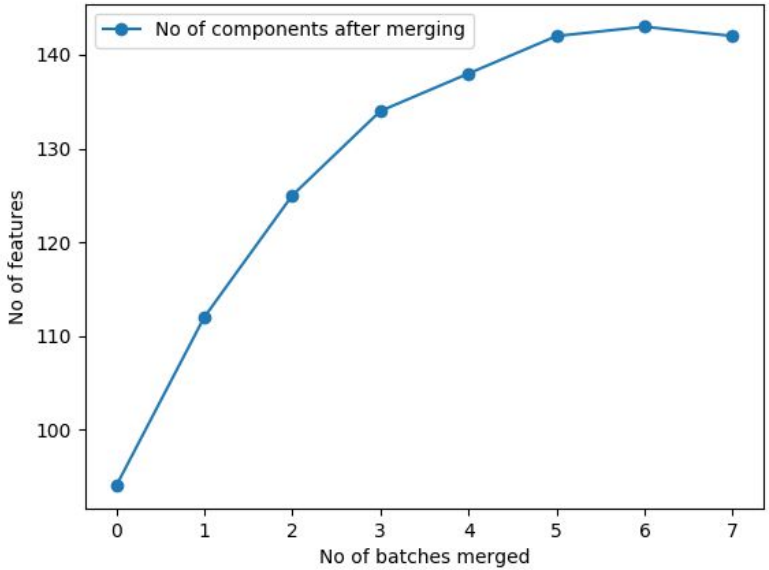
- Incremental Double PCA with PCC (K1)
- Incremental Single PCA with PCC (K2)
- Regular single PCA (K3)

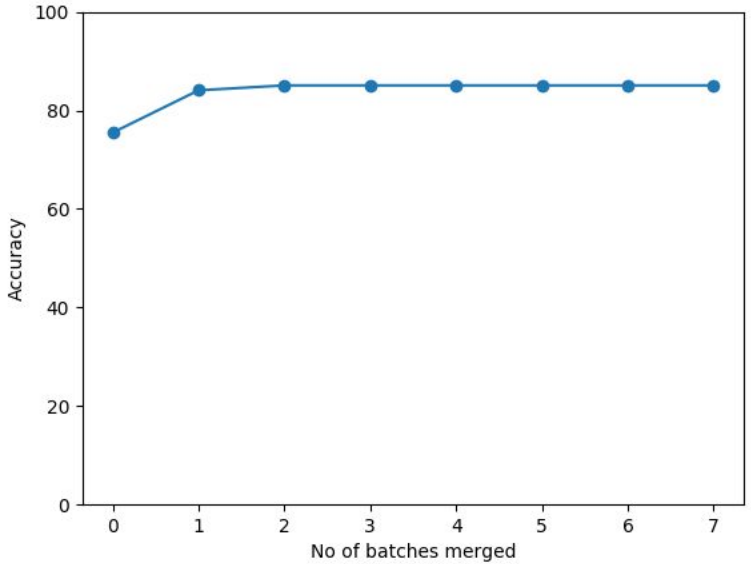
K1 , K2 , K3 number of components required to represent same amount of variance in data

$$K1 < K2 < K3$$

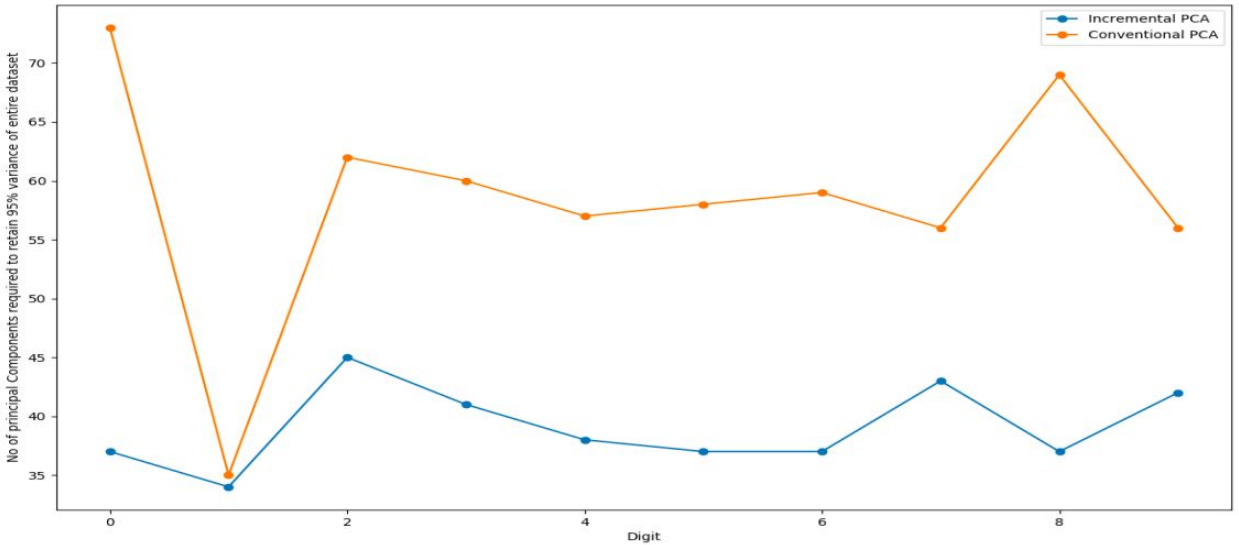
Method 2 is faster than method 1 and all have comparable accuracy levels

Graph for Double PCA





Conventional vs Incremental PCA



Future scopes

- Improve accuracy to that of State-of-art models
- Find a method to classify a single sample

Thank You