# Incremental Dimensionality Reduction for Global Feature Set Extraction

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

Under guidance of

Dr. Sonali Agarwal
IIIT Allahabad

# Candidate's Declaration

We hereby declare that the work presented in this project report entitled “Incremental Dimensionality Reduction for Global Feature Set Extraction”, submitted as 6th semester B-Tech IT mini project is an authenticated record of our original work carried out from January 2018 to April 2018 under the guidance of Dr. Sonali Agarwal. Due acknowledgements have been made in the text to all the resources and frameworks used.

Signature:

Date: 26th April, 2018

IIT2015032 Rohan MR

IIT2015039 Nishant Verma

IIT2015042 Raghav Saboo

IIT2015045 Harsh Vardhan

# Supervisor's Certificate

This is to certify that the project work "Incremental Dimensionality Reduction for Global Feature Set Extraction" is a bonafide work of Rohan MR (IIT2015032), Nishant Verma (IIT2015039), Raghav Saboo (IIT2015042), and Harsh Vardhan (IIT2015045) who carried out the project work under my supervision.

Signature
Dr. Sonali Agarwal

Date: 26th April , 2018

# Acknowledgment

We would like to thank all the people who have helped us in this endeavour. Their support means a lot to us and we wish to thank them individually.

We begin by thanking Dr. Sonali Agarwal for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project.

We would also like to express our special thanks to the panel under Prof. OP Vyas, Dr. Abhishek Vaish, Dr. Sonali Agarwal and Dr. Bibhas Ghoshal for their never ending support.

# Table of Contents

Candidates’ Declaration 2

Supervisor’s Certificate 3

Acknowledgement 4

Abstract 6

Introduction 7

Motivation 8

Problem statement 9

Literary Review Report 10-20

Plan of Action 21

Software Requirements 22

Implementation Details 23-26

Comparisons 27-30

Results 31-33

References 33

Suggestions and Remarks 34

# ABSTRACT

Many systems of handwritten digit recognition built using the complete set of features in order to enhance the accuracy. However, these systems lagged in terms of time and memory. These two issues are very critical issues especially for real time applications. Therefore, using Feature Selection (FS) with suitable machine learning technique for digit recognition contributes to facilitate solving the issues of time and memory by minimizing the number of features used to train the model.

# INTRODUCTION

In many supervised learning problems feature selection is important for a variety of reasons: generalization performance, running time requirements, and constraints and interpretational issues imposed by the problem itself.

With ever growing data complexity and the advent of big data, machine learning techniques have become indispensable in order to extract meaningful data from big dataset. In this regard feature selection becomes the prime focus of research for dealing with datasets having tens of thousands of variables.

In classification problems we are given *f* data points Xi E ~n labelled *Y* E ±1 drawn from a probability distribution *P(x, y).* We would like to select a subset of features while preserving or improving the discriminative ability of a classifier. As a brute force search of all possible features is a combinatorial problem one needs to take into account both the quality of solution and the computational expense of any given algorithm.

Traditionally, feature selection methods are used for centralized computing environment. But because of big data explosion and non-scalability of current feature selection algorithm there is shift towards data distribution and merging output of selection algorithms on each subset of datasets.

# MOTIVATION

With ever growing data complexity and the advent of big data, machine learning techniques have become indispensable in order to extract meaningful data from big dataset. In this regard feature selection becomes the prime focus of research for dealing with datasets having tens of thousands of variables

Traditionally, feature selection methods are used for centralized computing environment. But because of big data explosion and non-scalability of current feature selection algorithm there is shift towards data distribution and merging output of selection algorithms on each subset of datasets

Given that data is being generated at a faster pace than ever, the necessity to reduce the features regularly in order to maintain the system performance and to extract useful knowledge for decision making is becoming more evident. In such a scenario, an incremental approach to reduce the feature space, to accumulate the knowledge without looking back at the previous data, will be a remarkable one.

# Problem Statement

To obtain global feature set for handwritten digit images using incremental dimensionality reduction.

# Literary Review Report

## Introduction to Feature Selection for Handwritten Digit Recognition [1]

Many systems of handwritten digit recognition built using the complete set of features in order to enhance the accuracy. However, these systems lagged in terms of time and memory. These two issues are very critical issues especially for real time applications. Therefore, using Feature Selection (FS) with suitable machine learning technique for digit recognition contributes to facilitate solving the issues of time and memory by minimizing the number of features used to train the model.

## Support vector machines (SVMs) [2]

Support vector machines (SVMs) have been extensively used as a classification tool with a great deal of success from object recognition to classification of cancer morphologies and a variety of other areas, see e.g. In this article we introduce feature selection algorithms for SVMs. The methods are based on minimizing generalization bounds via gradient descent and are feasible to compute. This allows several new possibilities: one can speed up time critical applications (e.g. object recognition) and one can perform feature discovery (e.g. cancer diagnosis). We also show how SVMs can perform badly in the situation of many irrelevant features, a problem which is remedied by using our feature selection.

# One Class Classification (OCC)[3]

Conventional multi-class classification algorithms aim to classify an unknown object into one of several pre-defined categories. A problem arises when the unknown object does not belong to any of those categories. In one-class classification, one of the classes (referred to as the positive class or target class) is well characterized by instances in the training data. For the other class (non-target), it has either no instances at all, very few of them, or they do not form a statistically-representative sample of the negative concept.

To motivate the importance of one-class classification, let us consider some scenarios. One-class classification can be relevant in detecting machine faults, for instance. A classifier should detect when the machine is showing abnormal/faulty behaviour. Measurements on the normal operation of the machine (positive class training data) are easy to obtain. On the other hand, most faults will not have occurred so one will have little or no training data for the negative class.

In a conventional multi-class classification problem, data from two (or more) classes are available and the decision boundary is supported by the presence of example samples from each class originate the term One-Class Classification in their research work. Different researchers have used other terms to present similar concepts such as Outlier Detection, Novelty Detection or Concept Learning. These terms originate as a result of different applications to which OCC has been applied.

## Support vector domain description[4]

This paper showed the use of data description method which is used for novelty detection and outlier detection. In many applications we need to decide whether a new sample belongs to a group of existing training samples ie it is an inlier or it is something different ie it is an outlier. A spherical boundary is constructed around the target data with the support vectors describing the boundary of the sphere. The data domain description problem is also called one class classification problem. We need to find a sphere with minimum volume containing all the target data points. It is sensible to outlier so we introduce slack variables to penalize large distances.

$$F(R, a, \xi_i) = R^2 + C\sum_i \xi_i, \qquad (1)$$

Sphere has center at a, radius R and C is parameter that specifies the tradeoff between volume of sphere and no of errors. We want to minimize the above equation considering the constraints –

$$(x_i - a)^{\mathrm{T}}(x_i - a) \leqslant R^2 + \xi_i \quad \forall_i, \xi_i \geqslant 0. \qquad (2)$$

We convert the primal problem into a dual problem and the Lagrangian built is –

$$L(R, a, \alpha_i, \xi_i) = R^2 + C\sum_i \xi_i - \sum_i \alpha_i\{R^2 + \xi_i - (x_i^2 - 2ax_i + a^2)\} - \sum_i \gamma_i \xi_i, \qquad (3)$$

The Lagrange multipliers are alpha I's, epsilons are >=0 and after setting the partial derivatives equal to zero we get –

$$\frac{\partial L}{\partial R} = 0: \quad \sum_i \alpha_i = 1$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0: \quad \mathbf{a} = \frac{\sum_i \alpha_i \mathbf{x}_i}{\sum_i \alpha_i} = \sum_i \alpha_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial \xi_i} = 0: \quad C - \alpha_i - \gamma_i = 0$$

(4)

Thus the center of the sphere is the linear combination of inputs multiplied by factors alpha I's which we obtain by maximizing –

$$L = \sum_i \alpha_i (x_i \cdot x_i) - \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j),$$

with constraints $0 \leqslant \alpha_i \leqslant C,\ \sum_i \alpha_i = 1.$ (5)

Thus when a new test point z comes we classify it as an inlier when the following condition is satisfied –

$$(z \cdot z) - 2\sum_i \alpha_i (z \cdot x_i) + \sum_{i,j} \alpha_i \alpha_j (x_i \cdot x_j) \leqslant R^2.$$

(6)

Radius is obtaining by calculating the distance between the center and any support vector. All the points on boundary have their corresponding alpha I's >0 and < C and points inside the sphere have their alpha I's equal to 0. Outliers have alpha I's equal to C.

$$\|\mathbf{x}_i - \mathbf{a}\|^2 < R^2 \rightarrow \alpha_i = 0, \gamma_i = 0$$

$$\|\mathbf{x}_i - \mathbf{a}\|^2 = R^2 \rightarrow 0 < \alpha_i < C, \gamma_i = 0$$

$$\|\mathbf{x}_i - \mathbf{a}\|^2 > R^2 \rightarrow \alpha_i = C, \gamma_i > 0$$

(7)

Thus radius is -

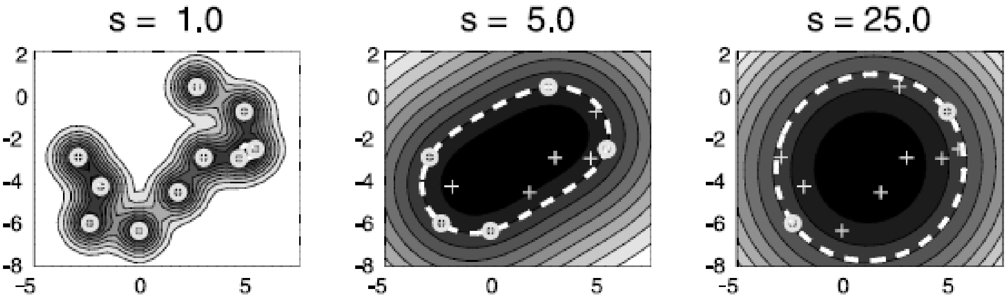
$$R^2 = (\mathbf{x}_k \cdot \mathbf{x}_k) - 2\sum_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}_k) + \sum_{i,j} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

for any $\mathbf{x}_k \in SV_{<C}$, the set of support vectors which have $\alpha_k < C$. (8)

When the data is not spherically distributed the inner products can be replaced by valid Kernels ie the Kernels that satisfy the Mercer's theorem. Thus kernels map the input into a higher dimensional feature space. For example a polynomial kernel with d=2 converts (x1,x2) into (x1,x2,x1*x2,x1^2,x2^2). Gaussian Kernel is generally used which has the form-

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/s^2) \quad (9)$$

S is the standard deviation. When we increase S the no of support vectors decreases and description starts looking like a sphere.



The circles are the support vectors.

For C<1/N there is no solution and for C>1 we can always find a solution thus the specified range of C that concerns us is between 1/N and 1 ie 1/N<=C<=1.

Thus this model was tested alongside several other techniques like KNN, Parzen, Gaussian, Instab on several UCI Machine learning data sets and the results showed that SVDD worked better than most of the other techniques. Thus domain description is an important class of problem in which we need to test whether a new sample belongs to given existing data and SVDD does this task efficiently.

# Support Vector Data Description by Tax and Duin[5]

This paper talked about Data description problem which is also known as One class classification problem in which we make a description with training data of target class and try to predict any new sample that whether it belongs to the target class or not. Normal classifiers generate good results for data close to training data. Data description can be used when we have huge data of one class and other class has very less data. An example is machine monitoring system in which we examine the current condition of a machine and detect if any fault has occurred or not. Normal working conditions are known to us but there can be a numerous no of possibilities in which the machine can become faulty. Thus SVDD works best in this case and in all cases in which we have data of only one class and we don't care about rest of the data as there may be infinite and we just want to make a prediction for a new sample that whether it belong to the given class or not. In usual classifiers the no of classes are finite so we cannot detect outliers that does not belong to any class and we cannot solve a problem in which we have data of only one class and we don't care about rest of the data as there may be infinite and we just want to make a prediction for a new sample that whether it belong to the given class or not. Two methods were proposed for one class classification using SVM one by Scholkopf that finds a hyperplane to separate the target data from origin with maximum margin and the other by Tax and Duin which finds a minimum volume hypersphere that contains all the points of the target class either inside or on the boundary of the sphere. Kernels can be used to map the input data into higher dimensional space if they are not separable in the current

space. A submersible water pump was analysed for it’s normal working conditions and any fault if it occurs should be detected.
The experiment showed that SVDD together with Gaussian Kernel gave the best results. Using polynomial Kernel did not gave satisfactory results due to the influence of norms of input vectors. The method by Scholkopf that finds a hyperplane to separate the target data from origin with maximum margin also gave results comparable to the SVDD together with Gaussian Kernel.

## Feature scaling in support vector data description[6]

This paper talked about importance of feature scaling in SVDD. When data of only one class is available then the problem is called one class classification problem. SVDD are quite efficient in solving one class classification problems. It is tougher than the normal two class classification problem as data of both classes are available but in this case only data of one class is available so it becomes difficult to identify the boundary to separate the classes. SVDD , K means , KNN are classifiers that are affected by scaling of features present in data. Three different types of scaling were applied namely –
1 Scaling by variance – Here the features in each direction was divided by it’s variance.
2 Scaling by domain – All features are scaled in the range [0,1].
3 Scaling to minimax – minimum of the maximum value of features in all directions is assigned as radius of sphere R and all values are scaled in the range [0,R].

All types of scaling was applied on two datasets one artificial ie the Hingelyman dataset and the other a real world dataset namely the handwritten digits from NIST16 database. Results showed that feature scaling reduced the error on the outlier data and a better boundary was obtained around the training data for one class classification.

## Incremental Dimensionality Reduction [7]

Traditionally, dimensionality reduction is used when the additional features start to degrade the performance. However it can be used to effectively maintain accuracy at the same time maintaining the performance of the system. This process however can be done as and when new data with different dimensions are encountered. And this is called Incremental dimensionality reduction. This process takes the approach of extracting, carrying forward and building the knowledge without reusing the previously explored data. This process of dimensionality reduction consists of feature subsetting and feature transformation. Feature subsetting is the process of representing the entire feature set by retaining independent features and discarding the dependent ones. The knowledge obtained from a data set is characterized by the variance of the feature set, classification accuracy, and predictor performance.

In this paper hyper spectral data of Indiana pines is used. The incremental flow of features is considered in two ways. One is sequence compulsive arrival of features and the other from a distributed environment. In our study, we will consider only the distributed environment technique. The framework for the incremental feature subsetting consists of two variables namely, Pearson's correlation coefficient represented by `r' and the threshold factor `t'. For PCC values less than 0.6 both the features are added into the optimal subset and for ones with greater than 0.6, the one with higher variance is kept and the other rejected.

In this paper, the distributed environment was modelled by considering the dataset divided into three batches and incrementally adding them in different order. The experiments carried out showed that knowledge was progressively maintained and hence establish its validity. For feature transformation PCA was used. The paper then went on to find the optimum merging sequence using Prim's and Kruskal's techniques. We would be using these techniques to try and achieve dimensionality reduction for our own dataset.

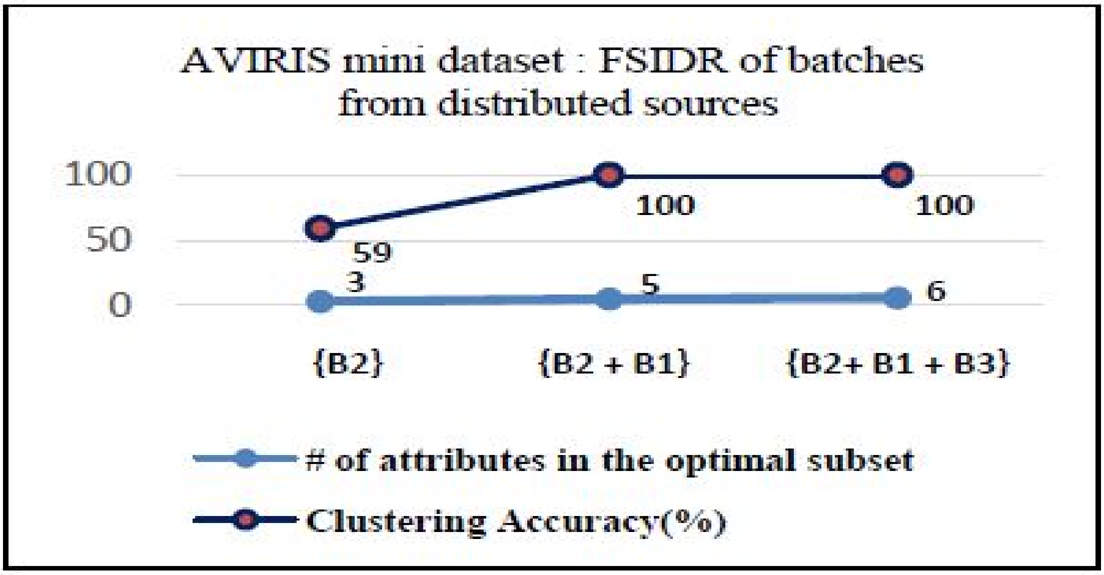


Fig 2: Distributed FSIDR of distinct class dataset

## Feature Selection [8]

With ever growing data complexity and the advent of big data, machine learning techniques have become indispensable in order to extract meaningful data from big dataset. In this regard feature selection becomes the prime focus of research for dealing with datasets having tens of thousands of variables.

Ultra-high dimensionality implies requirement of huge memory along with high computational cost for training. High dimensionality also results in curse of dimensionality which means various phenomena occurs when we organise and analyse data in high dimensional space that do not occur in low-dimensional settings. Usually data set is represented as matrix where rows implies samples and columns as features. So there is need to find narrower matrices having smaller features but same accuracy and performance. Process of finding these is dimensionality reduction. For this we use Feature extraction.

Feature extraction is decomposed as feature construction and feature selection. We used feature construction procedure normalization as pre-processing step to reduce some features after which feature selection methods are employed.

Feature selection is data processing method for selecting relevant features and discarding irrelevant or redundant features of problem with a minimum degradation of performance.

The problem of feature selection can be stated as : given set of original features F, find subset of features F' such that F' has same accuracy and effect on performance as F has.

It has two key elements namely criterion function and subset searching method with given criterion function. Function is used for predicting score of feature subsets while searching method explore feature subset space and uses this score to find best subset of features.

In this paper we used SVDD based feature selection method as they are even applicable to normal (generated) dataset.

## Feature Selection and OCSVM[9]

In this paper, we used feature selection as one class SVM problem. One class SVM represents underlying data by finding hyperplane which maximally separates point from origin. Support vectors in OCSVM are basically outliers. We intend to separate the set of features from the set of non-features in space where features are data points and samples are dimension. Support vectors that describe hyperplane boundary are required set of feature that represent underlying set of features. Hence task reduces to finding these support vectors for which SVDD is used.

# PLAN OF ACTION

- Dataset Generation
  - Create image data
  - Normalization
  - Digit Extraction
  - Normalization

- Feature Extraction
  - Scan line Method
  - Zone based Method
  - Run length Method

- Incremental Feature Transformation and selection using Principal Component Analysis (PCA) and Pearson's Correlation Coefficient (PCC).

- One Class Classification on the obtained global feature set.

# Software Requirements

Software's used –

- OS used - Ubuntu 16.10,Windows
- Python 3.5.2
- Pycharm Community Edition 2017
- MATLAB
- Libraries used-
    - Pandas
    - Tensor Flow
    - Numpy
    - Cvxpy
    - Cvxopt
    - Scikit learn

# Implementation Details

## ➢ Dataset Generation

- Around 600 Images of each handwritten digit 0-9 was taken.
- Then we resized each image to 50 x 50 pixels image.
- After that we cropped the image and removed the outer empty portions.
- The cropped image was again rescaled to 50 x 50 pixels image.

## ➢ Feature Extraction

- Important features were extracted using 3 different methods :-
    - Scan line method
        - We selected 2*(m + n) features of m x n image.
        - For 50x50 image we selected 200 features using horizontal and vertical scan line.
    - Zone based method
        - We divided our image into zones and for each zone we performed scan line approach.
        - For 50x50 image we divided our image into zones of 10x10. So, in total we have 25 zones each having 40 features. Therefore, we have 1000 features.

- Run length method
    - We selected 5*(m + n) - 6 features of m x n image.
    - For 50x50 image we selected 494 features.

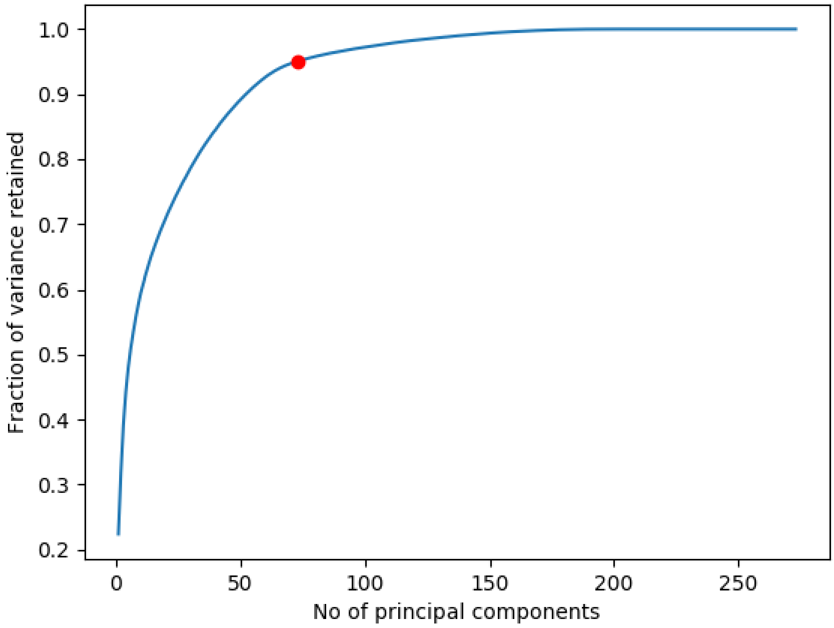
- On the basis of performance (accuracy prediction) we selected Run length method and applied it on 6000 samples of each digit from Mnist dataset. So, we obtained 274 features on 28 x 28 image.

## Incremental Feature Transformation

Two approaches –

- We divided 6000 samples of each digit into batch of size 10 containing 600 samples of each digit. On the first batch we applied PCA to reduce the no of features so as to obtain the principal components that capture 95 % of Variance of the data and then it was merged incrementally with the next batch containing 274 features and 600 samples to obtain reduced feature set using PCC with threshold set as 0.7.This process was repeated until all the batches are processed.

- We divided 6000 samples of each digit into batch of size 10 containing 600 samples of each digit. On the first batch we applied PCA to reduce the no of features so as to obtain the principal components that capture 95 % of Variance of the data and then it was merged incrementally with the next batch after applying PCA to the next batch containing 274 features and 600 samples and thus finally obtaining the reduced feature after merging using PCC with threshold set as 0.7.This process was repeated until all the batches are processed.



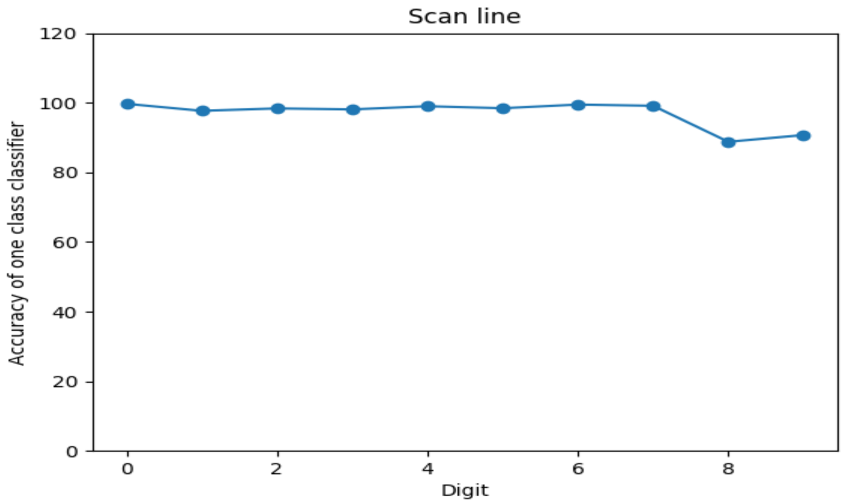
We selected our number of principal components on the basis of fraction of variance retained as shown by above plot.

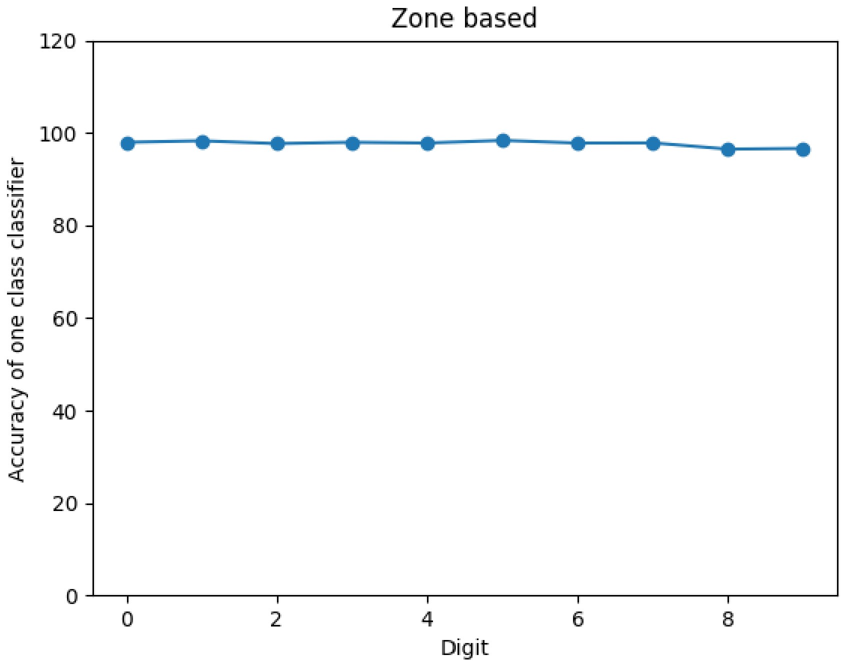
## ➢ One class classification

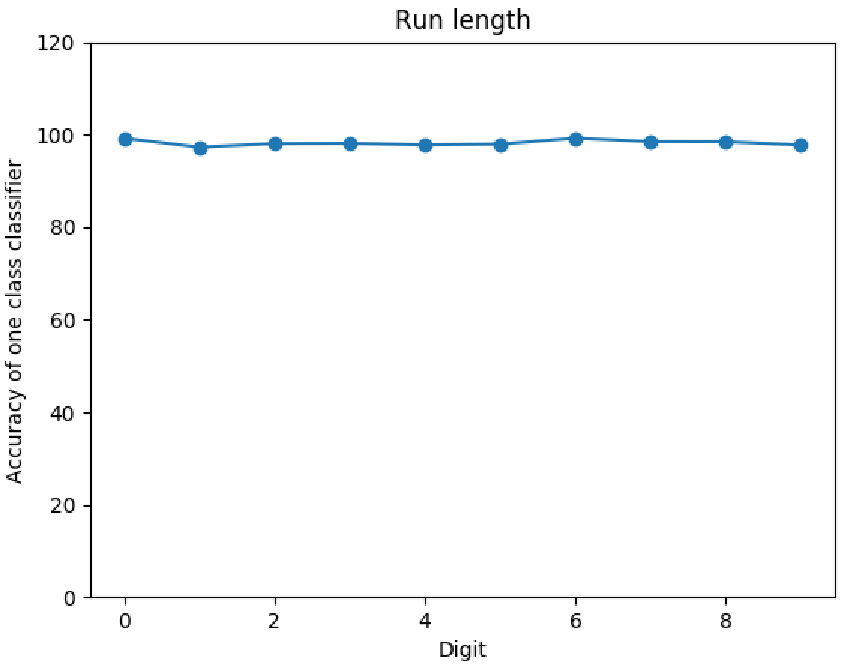
- Then we did one class classification using sklearn's one class SVM and for testing we applied K fold cross validation for testing accuracy of our trained model.
- 10 different models were trained and tested one for each of the 10 digits

# COMPARISONS

## ➢Feature Extraction



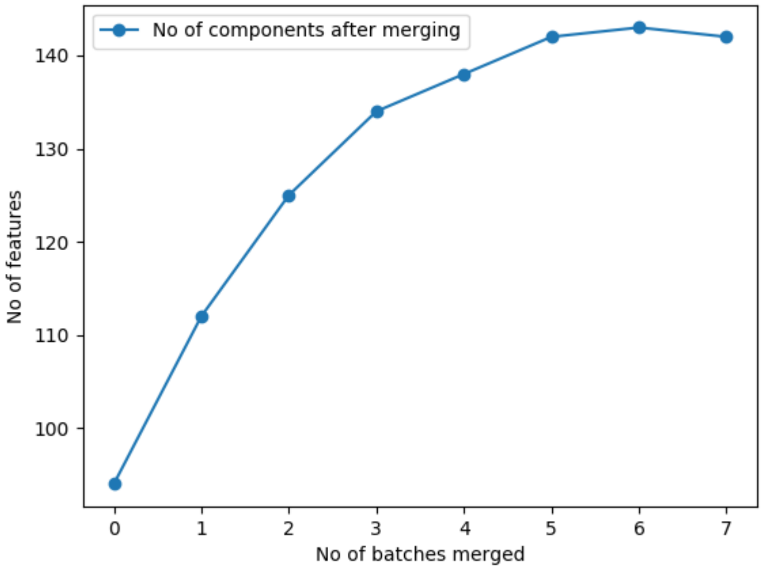




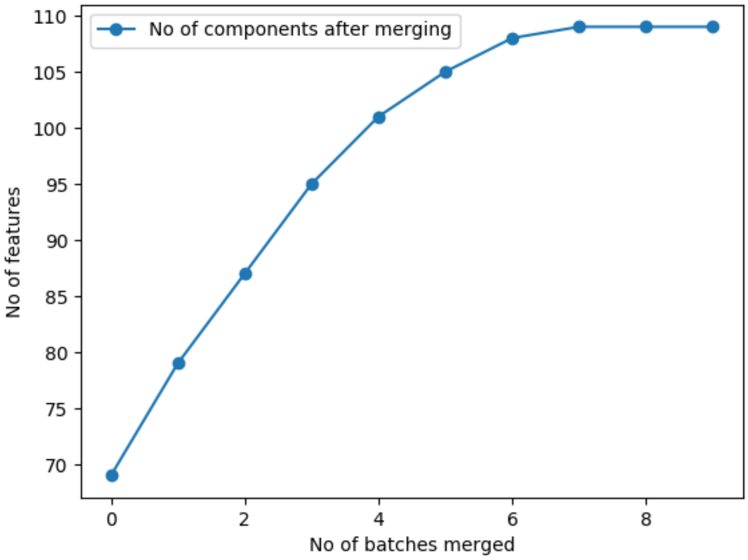
- On the basis of above plots it can be concluded that on average run length method will give better accuracy.

- ## Number of features

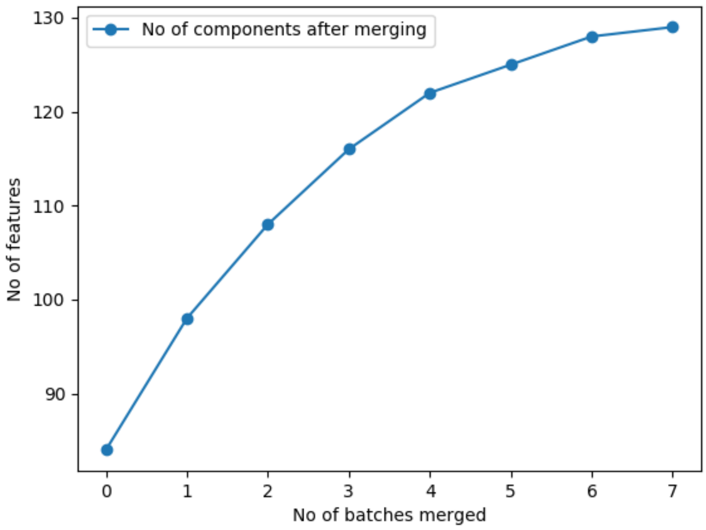
- For digit 0



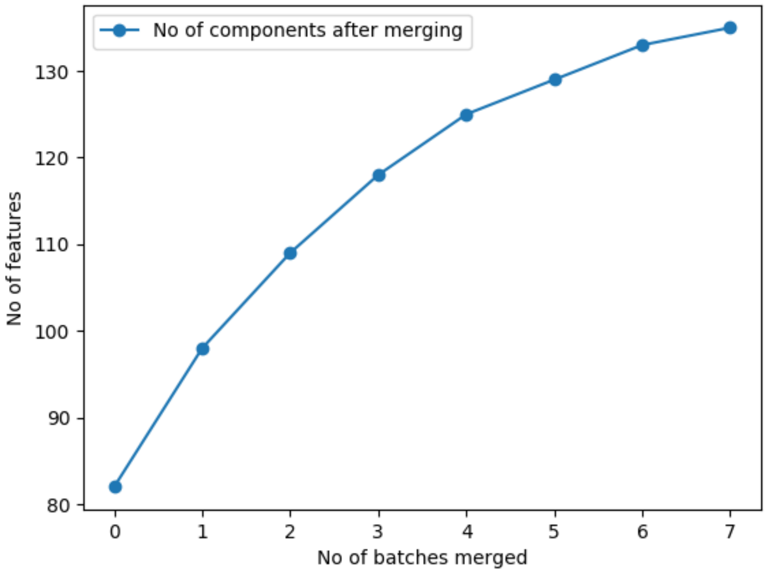
- For digit 1



o For digit  6



o For digit 9

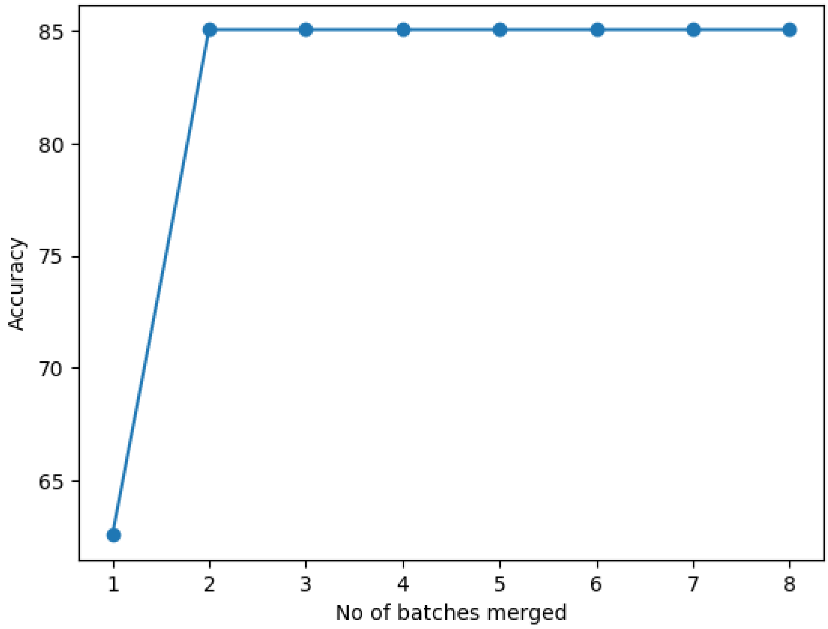


- From above plots we can conclude that number of components decreases after merging.

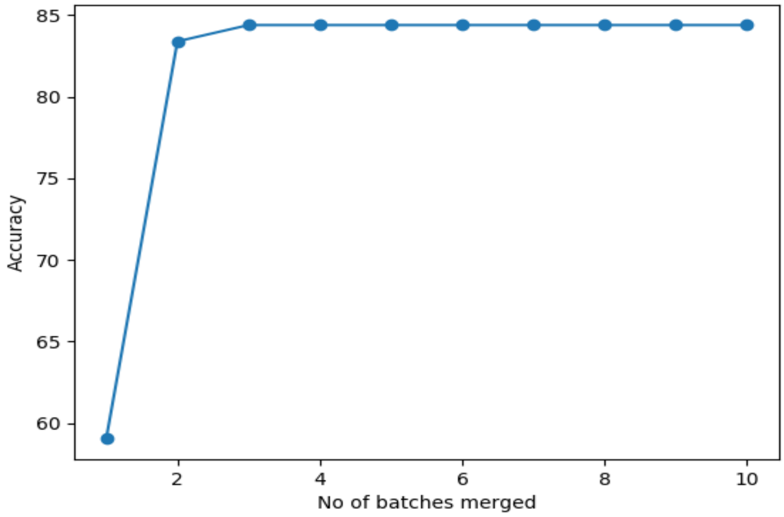
- Rest of digits will have same graph pattern as above.

# RESULTS

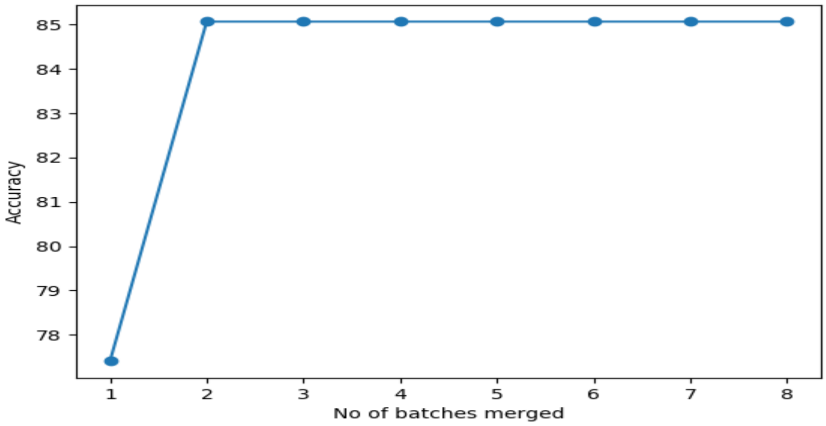
- For digit 0



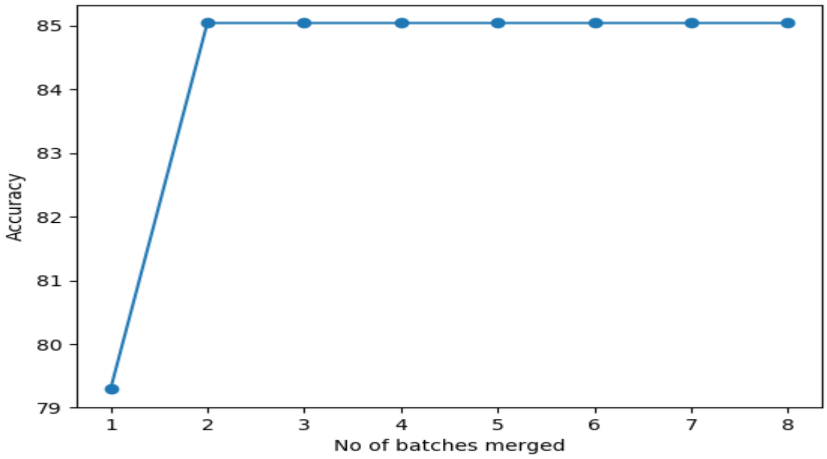
- For Digit 1



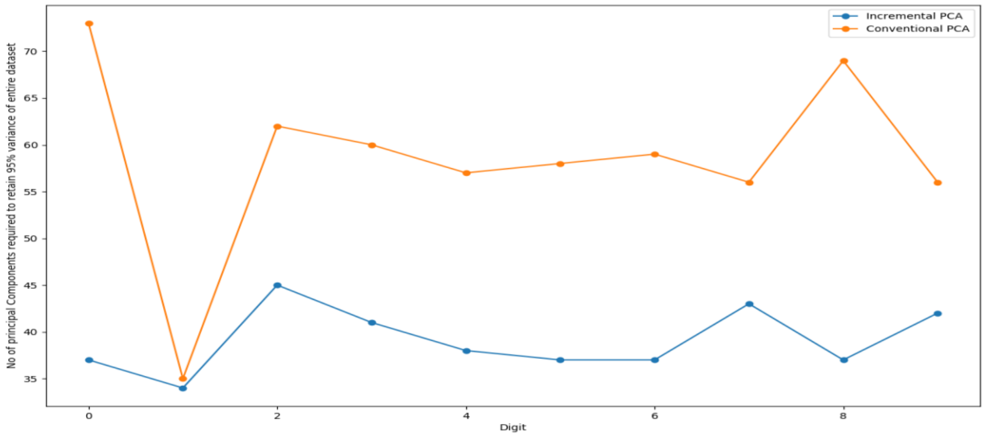
- For digit 6



- For digit 9



- So using incremental approach and with help of reduced number of features we can achieve a significant level of accuracy.



- Thus, the incremental PCA uses approximately less components to assimilate the same amount of information that the conventional PCA would require to use.

## References

[1] A. E. Areej Alsaafin, “A Minimal Subset of Features Using Feature Selection for Handwritten Digit Recognition,” *Intelligent Learning Systems and Applications,* vol. 9, no. 4, pp. 55-68, 2017.

[2] S. M. O. C. M. P. P. V. V. J. Westont, “Feature Selection for SVMs,” 2000.

[3] M. G. M. Shehroz S Khan, “A Survey of Recent Trends in One Class Classification,” Springer Verlag-LNAI, Ireland, 2009.

[4] R. P. D. David M.J. Tax, “Support vector domain description,” *Pattern Recognition Letters,* vol. 20, pp. 1191-1199, 1999.

[5] R. P. D. DAVID M.J. TAX, “Support Vector Data Description,” Kluwer Academic, Netherlands, 2004.

[6] D. T. R. D. P. Juszczak, “Feature scaling in support vector data description,” Netherlands, 1999.

[7] P. Nagabhushan, “Incremental Dimensionality Reduction in Hyperspectral Data,” *International Journal of Computer Applications,* vol. 163, no. 7, p. 0975 – 8887, 2017.

[8] N. S.-M. A. A.-B. V. Bolón-Canedo, “Recent advances and emerging challenges of feature selection in the context of big data,” *Knowledge-Based Systems,* vol. 86, pp. 33-45, 2015.

[9] B. K. K. a. S. P. Prasad Y., “Feature Selection using One Class SVM: A New Perspective,” in *NIPS 2013 (MLCB)*, Nevada, 2013.

Suggestions And Remarks: