International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Handwritten Digit Recognition using Different Dimensionality Reduction Techniques

Ruksar Sheikh, Mayank Patel

Abstract: ***The objective of this paper is to introduce to Technologies of linear dimension reduction popularly known as Principal Component Analysis and Linear Discriminant Analysis. PCA reduces the size of data and conserve maximum variance in the form of new variable called principal components where LDA works with minimum class distance and maximizing difference between the classes. Axis of maximum variance is found by PCA while axis of class separability is found by LDA. This method is experimented over and MNIST handwritten digit data set. Our conclusion explains PCA can outperform LDA when training data set a small and recalls values with lesser computational complexity. The present in linear techniques in this paper presents clear understanding and methods in comparative manner.***

Keywords: **PCA, LDA, Dimensionality Reduction, MNIST Handwritten Digit dataset, Covariance**

I. INTRODUCTION

We are living in digital Universe Era where the size of data is getting almost double every year or two. The major problem we are facing is increasing in complexity and volume of data (Cheng, You 2016). Due to these problems increase dimensionality, various dimension reduction techniques comes into play (Snehal K, Machchhar 2014). As these techniques reduce the dimensions of data and hold much of the important variables. It reduces the storage problem as well as computational time with same analytical result. It also resolve is a machine learning problem to get better features for regression. Data reconstructed in dimensionality reduction is somewhat similar to original data (Varghese et al, 2012). Here fetching of handwritten digits are in image form which comprises of steps like getting the image preprocessing, segmenting, representation and description and finally recognition and interpretation. Four types of digital images are present that are in binary, grayscale, true color or RGB and index form(Rafael 2002).Here we are using binary representation of images which include text, fingerprints or architectural plans where each pixel is black or white. Similarly grayscale images comprises of X-rays, in RGB images each pixel is explained by amount of color red, green and blue in it. Indexing images are associated with color map. Over years there has been rapid growth in digital information. This pattern has motivated research in image database which was ignored by earlier system as tremendous amount of data was required to represent an image and was difficult to automatically analyze the image. Presently large storage capacities are available at low cost, so storage is not an issue. Huge image database is used by many applications now days likewise biometric, crime prevention, fashion, medical diagnosis etc (H H Pavan 2012). Thus an important issue is handwritten digit database recognition from large database. These dimensionality reduction techniques undergo certain steps like collecting data images, preprocessing, feature extraction, matching and recognition data. Filtration, normalization, sedimentation and object identification comes under pre-processing stage. In feature extraction visual information are extracted from image and store it in feature vector database. Here the problem arises as large number of features are extracted which will be requiring larger storage plus more computational time, so here dimensionality reduction comes into play. The widely used algorithms for dimensionality reduction are Principal Component Analysis and Linear Discriminant Analysis. PCA defines principal component extracted from original features, sometime it is not sufficient to consider only few principal component for many images then LDA comes into play for dimensionality reduction (S. Balakrishnama). LDA guarantees maximal separation within the classes as it maximizes the ratio between the class variance to within class variance. This paper properly depicts comparative study of implementation of both the algorithms to given data set. Experimental results proving considerably reduced dimensions without much degradation in performance. The rest of the paper is organized as follows section 2 literature review section 3 material and methods section for methodologies of PCA and LDA Section 5 comparative study of PCA and LDA and experimental results section 6 comprises of Conclusion.

Revised Manuscript Received on 30 July 2019.

* Correspondence Author

Ruksar Sheikh*, Dept. of Computer Science Engineering, Geetanjali Institute of Technical Studies, Udaipur , India.

Mayank Patel, Dept. of Computer Science Engineering, Geetanjali Institute of Technical Studies, Udaipur , India.

II. LITERATURE REVIEW

When PCA and LDA are comparatively analyzing PCA outperforms LDA when lesser amount of data is analyzed. If the sample data set is larger than LDA may perform better than PCA (Borade, Adgaonkar 2011).

Linear Discriminant Analysis bit by bit paper by Sebastian Raschka explained PCA as unsupervised algorithms which ignore class labels and focus on finding principal components that maximize the variance. In contrast he also stated that LDA is supervised and will focus on maximum separation between the classes. In Literature handwritten characteristic reorganization is extensively investigated (C. Liu. K Nakashima 2003). Previous work on handwritten digit recognition by LeChun Y Gradient based learning Applied to document recognition Proceedings of IEEE 1998.

III. MATERIALS & METHODS

A. Dataset

1) MNIST data-

In computer science, mathematics and engineering field of image recognition is growing very fast. There are various applications in which image recognition techniques are providing solutions for many of the problems like face detection and recognition, surveillance, national security, artificial intelligence etc. The problem of identifying digit from images of handwritten digits comes under MNIST digit recognition. Here the image is determined correctly or digit displayed is predicted. MNIST is a Modified National Institute of Standards and Technology database is a sub subset of large data set by NIST National Institute of Standards and Technology US. It consists of binary images from 0 to 9 and of 28 x 28 sizes. Images are divided into two sets 60,000 training and 10000 testing sets.

Digits	0	1	2	3	4	5	6	7	8	9
Count	5923	6742	5958	6131	5842	5421	5918	6265	5851	5949

Table 1: Distribution of 10 digits

Figure 1: A small subset of the MNIST data set

B. Algorithms and Methodology-

1) Principal Component Analysis

Karl Pearson in 1901 developed PCA as a data reduction technique later H. Hotelling in 1933 independently develop and named it PCA. The technique by which data is reduced in a way that maximum of original data is retained (Nick at 2015). As PCA is statistical approach in orthogonal transformation are used to transform set of correlated data into set of linearly uncorrelated variables (Elavarason, Mani 2015) commonly called as principal components and principal components can be less or equal to actual number of variables. It is variance based transformation where maximum variance is always represented by first principle component (Paul at 2013).PCA is used in Image Compression biometrics, face detection etc (Ramadevi, Usharani 2013)

Let x_1, x_2…..x_n be original data set in D space the smaller subset W with W<D.

Let y_i be defined in equation with i=1.....n Linear combination of variables

$$yi = A^T (x - m_x) \ldots \ldots \ldots \ldots \ldots \ldots (1)$$

where $A = [\alpha 1| \ldots |\alpha_n]$ is matrix with eigenvector of $\sum$, covariance of original derived data and m_x = mean original data.

Algorithm

i. Representing a image in one dimensional vector of size N x N. supposing M vectors of N size (row of image x column of image).represented as sample images. Now training set is Γ1,Γ2,Γ3….ΓM.

ii. Mean pixel intensities of each image is calculated and subtracted from corresponding image. And it carries out for all images.

iii. N^2 x N^2 covariance matrix is calculated as $C = A A^T$.

iv. The eigenvalues of covariance matrix C is calculated by solving $(C\lambda - I) = 0$. To find X eigenvector repeat where Xi is corresponding eigenvalue.

v. In accordance to descending order of eigenvalues the eigenvector are stored.

Choosing the first 'K' eigenvalues and eigenvectors.

2) Linear Discriminant Analysis

Linear Discriminant Analysis [G. Sasikala 2010] is also most commonly used technique of dimensionality reduction in which preprocessing for pattern classification and machine learning is done [Arunasakthi 2014]. Maximize the difference between the classes and minimizing distance within classes is done by LDA. LDA create the largest mean difference between outcome classes data is described using independent features. It formulates a projection A which maximize is the S_b and S_w ratio. S_b is between the classes and S_w is within classes scatter [Yu H, Yang 2001].

$$\underset{A}{\arg\max} \frac{A S_b A^T}{A S_w A^T} \ldots \ldots \ldots \ldots \ldots \ldots (2)$$

Algorithm

i. Calculate the d dimensional mean vector from dataset for different classes.

ii. Calculate scatter Matrix between the classes and within the classes.

iii. Calculate eigenvectors (e_1, e_2, e_3...... e_d) and eigenvalues (λ_1, λ_2, λ_3.....λ_d) for scattered Matrix.

iv. Sort eigenvector by decreasing eigenvalue and eigenvector with largest eigenvalue is chosen to form d x k matrices W.

v. Use W Matrix to change samples on to new subspace. Summarizing by

$$y = W^T \times X \quad \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where X= d x 1D vector of 1 sample

y= transform k x 1D sample in new subspace. With Complexity BIG O N (approx.)

IV. EXPERIMENTAL RESULTS

We loaded the data in machine learning program Python and PCA and LDA are performed. The MNIST dataset comprises of 60,000 training and 10000 testing data set. The principal components selects by PCA in dimensionality reduction depends on variance of data and maximum variance is retained by PCA. We plot original normalize data and recover data after applying PCA. As we have bulk of data entries so we selected a few of digits to be displayed as output for analyzing result very well. Digits image are selected randomly. Analyzing effect on recovered images can be changed by number of principal components. The quality of image we want to recover depends on number of chosen eigenvectors which represents the database. Analyzing the result after varying the number of eigenvectors M the observation is

M = 90 recovered images is blur and can't be recognized

M = 300 image obtained is less blood and can be identified

M = 800 a better images recovered.

No significant impact on images by further increasing the value of M.

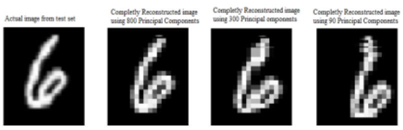

Figure 2 :- Digits after Dimensionality Reduction using PCA

With LDA three sets of experiments are conducted we randomly used 8 images per digit for training and for testing another 8 images. We repeated every experiment 10 times to reduce variations. Before dimensionality reduction the average recognition accuracy is 90.8% after applying dimensionality reduction. Null spacing is discarded and average recognition accuracy becomes 86.6 %. This proves null space are important as discriminative information does exist outside of it.

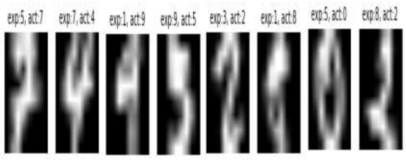

Figure 3: - Digits after Dimensionality Reduction using LDA

Comparative study of PCA and LDAnTo extract the best feature of the image PCA and LDA is used. By PCA and LDA, 75% of feature set is reduced. Then the reduced image is compared to reduce data feature set. With different feature extraction schemes quality of the image is retrieve. First the images are plotted without dimensionality reduction in figure 1, then after dimensionality reduction image and plotted in figure 2 and 3. Here the image of PCA is compared with LDA and showed PCA is better tool for dimensionality reduction when used on small data set when compared to LDA.

V. CONCLUSION

This paper represents the two most common dimensionality reduction techniques namely LDA (Linear Discriminant Analysis) and PCA (Principal Component Analysis). Revealing meaningful structures and unexpected relationships in multivariate data is the main aim of dimensionality reduction. Much of class discriminatory information is saved by LDA while first few principal components to reduce the dimension of data significantly is used by PCA. Eigenvectors are computed by LDA and stored in scatter matrices, the between classes scatter matrix and within classes scatter matrix. Many a times LDA outperforms PCA when data is linearly sorted and large in amount but when data is not linearly sampled PCA can perform in a better way. Here we conclude that PCA outperform LDA in many of the cases and PCA can be adopted as effective method for dimensionality reduction only when dataset is small or when input data is non -uniformly sample the distribution.

REFERENCES

1. Cheng Long; You Cheny (2016): Hybrid Non-Linear Dimensionality Reduction Method Framework Based on Random Projections, IEEE International Conference on Cloud Computing and Big Data Analysis.
2. Snehal K Joshi.; MachchharSahista (2014): An Evolution and Evaluation of Dimensionality Reduction Techniques- A Comparative Study. IEEE International Conference on Computational Intelligence and Computing Research.
3. Varghese Nebu; Verghese Vinay; Prof. P. Gayathri and Dr. Jaisankar N. (2012): A Survey of Dimensionality Reduction and Classification Methods. International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3.
4. Rafael Gonzalez and Richard E. Woods, 'Digital Image processing', Addison Wesley, 2nd Edn., 2002. H.H. Pavan Kumar Bhuravarjula And V.N.S Vijaya Kumar(2012), 'A Novel Content Based Image Retrieval Using Variance Color Moment', Int. J. Of Computer And Electronics Research,Vol. 1, Issue 3, ISSN: 2278-5795, pp.93-99.
5. S. Balakrishnama and A. Ganapathiraju, 'Linear Discriminant Analysis - A Brief Tutorial'. Institute for Signal and Information Processing, Department of Electrical and Computer Engineering Mississippi State University.
6. BoradeSushmaNiket; Dr. Adgaonkar Ramesh P. (2011); Comparative analysis of PCA and LDA. ICBEIA.
7. Sebastian Raschka (2014), 'Linear Discriminant Analysis bit by bit'.

8. C. Liu. K. Nakashima, H. Sako, and H. Fujisawa, "Handwritten digit recognition: benchmarking of state-of-the-art technigues,"Pattern Recognition, vol.36, no. 10, pp. 2271-2285, 2003.
9. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE. 1998 Nov;86(11):2278-324.
10. Nick William; Shelton Joseph; Bullock Gina; Esterline Albert; AsameneKassahun (2015): Comparing Dimensionality Reduction Techniques, Proceedings of the IEEE Southeast Con 2015, Florida.
11. Elavarasan N.; Dr. Mani K. (2015): A Survey on Feature Extraction Techniques ,IJIRCCE, Vol. 3, Issue 1.
12. Ramadevi G. N.; Usharani K. (2013): Study on Dimensionality Reduction Techniques and Applications. IJPAPER, Vol 04, Special Issue01.
13. G. Sasikala , R. Kowsalya and Dr. M. Punithavalli (2010), 'A Comparative Study Of Dimension Reduction Techniques For Content-Based Image Retrieval', The Int. J. of Multimedia & Its Applications, Vol.2, No.3, Doi : 10.5121/Ijma.2010.2303 40.
14. Arunasakthi. K and Kamatchipriya. L(2014), 'A Review On Linear And Non-Linear Dimensionality Reduction Techniques', Machine Learning And Applications: An Int. J. (Mlaij), Vol.1, No.1, Pp.65-76.
15. Yu H, Yang J. A direct LDA algorithm for highdimensional data with application to face recognition. Pattern recognition. 2001 Oct 1;34(10):2067-70.

AUTHORS PROFILE

Ruksar Sheikh completed her B.tech in Computer Science in year 2014 form Rajasthan Technical University, India, pursuing her M.tech in Computer Science from Rajasthan Technical University, India. Her research work include Machine learning, Python , Dimensionality Reduction, Image processing etc.

Mayank Patel completed his Ph.D. in the domain of Multimedia Services over Wireless LAN from College of Technology and Engineering (MPUAT, Udaipur). He is working as an Associate Professor in the Department of Computer Science and Engineering at Geetanjali Institute of Technical Studies, Udaipur, Rajasthan. His area of interest includes Programming in C, Data Structure and Algorithms, OOPS, Programming in JAVA, Web Application Development through J2EE, Web Application Development through .Net Framework, Computer Networks and Wireless Networks, Principles of Programming Languages and many more. I also publish various textbooks in the programming domain