International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Review of Dimensionality Reduction Techniques

Bothe Priya V.[1], Rangole Jyoti S.[2]

Department of Electronics and Telecommunication, Vidya Pratishthan's College of Engineering Baramati, India

Abstract: ***High-dimensional data are common in many domains, and dimensionality reduction is the necessary to cope with the curse-of-dimensionality. This phenomenon states that an enormous number of samples is required to perform accurate predictions on problems with high dimensionality. Dimensionality reduction, which extracts a small number of features by removing irrelevant, redundant, and noisy information, can be an effective solution. Different statistical methods for dimensionality reduction have been proposed in recent years and various research groups have reported contradictory results when comparing them. The commonly used dimensionality reduction techniques include supervised approaches such as Linear Discriminant Analysis (LDA), and unsupervised ones such as Principal Component Analysis (PCA), and additional spectral and manifold learning methods. When class labels are available, the supervised approaches such as LDA are generally more effective than the unsupervised ones like PCA for classification. This paper aims at the review of two most widely used dimensionality reduction techniques, PCA and LDA. Based on this a way ahead will be presented to facilitate research and development in sediment classification.***

Keywords: Dimensionality reduction, Linear Discriminant Analysis (LDA), model selection, Principal Component Analysis (PCA)

1. Introduction

Dimensionality reduction has been a key problem in many areas of information processing, such as data mining, information retrieval, and pattern recognition [1]. When data are represented as points in a high-dimensional space, one is often confronted with tasks like nearest neighbor search. Many techniques have been proposed to index the data for fast query response, such as K-D tree, R tree etc. [2]. However, these techniques can only operate with small dimensionality, typically less than 100. The effectiveness and efficiency of these techniques drop exponentially as the dimensionality increases, which is commonly referred to as the curse of dimensionality.

During the last few years, with the advances in computer technologies and the advent of the World Wide Web, there has been an explosion in the amount of digital data being generated, analyzed, stored and accessed. Much of this information is multimedia in nature, including text, video data and image [3]. The multimedia data are typically of very high dimensionality, ranging from several thousands to several hundreds of thousand. Learning in such high dimensionality in many of the cases is almost infeasible. Thus, learnability gives necessity of dimensionality reduction. Once the high-dimensional data is mapped into a lower dimensional space, conventional indexing schemes can then be applied to it.

The goal of this paper is to present an independent study of two most popular dimensionality reduction algorithms in completely equal working conditions. They are: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA [4] finds a set of the most representative projection vectors such that the projected samples retain the most information about original samples. LDA [5] uses the class information and finds a set of vectors that maximize the between-class scatter while minimizing the within-class scatter. PCA performs dimensionality reduction while preserving as much of the variance in the high dimensional space as possible. LDA performs dimensionality reduction while preserving as much of the class discriminatory information as possible.

The Fig. 1 below shows two learning instances, marked by circles and crosses, for each class whose underlying but unknown distribution is shown by the dotted curve shown [6]. Taking all of the data into consideration, PCA will compute a vector that has the largest variance associated with it. It is shown by the vertical line labeled PCA. On the other hand, LDA will compute a vector that best discriminates between the two classes. This vector is indicated by the diagonal line labeled LDA. The decision thresholds yielded by the nearest-neighbor approach for the two cases here are marked as DPCA and DLDA. As can be seen by the manner in which the decision thresholds intersect the ellipses corresponding to the class distributions, PCA will yield superior results in this case.

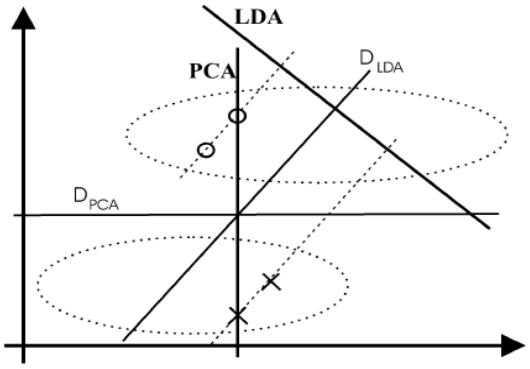

Figure 1: There are two different classes represented by the two different Gaussian-like distributions. However, only two samples per class are supplied to the learning procedure (PCA or LDA).

Of late, there has been a trend to prefer LDA over PCA because, as intuition would suggest, the former deals directly with discrimination between classes, while the latter deals with the data in its entirety for the principal components analysis without paying any particular attention to the underlying class structure. Examples such as the one depicted in Fig. 1 are quite convincing with regard to the fact that LDA is not always superior to PCA.

2. Dimensionality Reduction Techniques

A. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a backbone of modern data analysis. It is a black box that is widely used but poorly understood. Principal component analysis is appropriate when there has been obtained measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) which will account for most of the variance in the observed variables. The principal components can be then used as predictor or criterion variables in subsequent analyses [7][8]. PCA is used widely in all forms of analysis – from neuroscience to computer graphics because it is a simple, non-parametric method of extracting relevant information from confusing data sets. With minimal additional effort PCA provides a roadmap for reducing a complex data set to a lower dimension data set to reveal the sometimes hidden, simplified structure that often underlie it. PCA is useful when an obtained data on a number of variables possibly a large number of variables and believe that there is some redundancy in those variables [9]. In such case, redundancy means that some of the variables are possibly correlated with one another, because they are measuring the same construct. Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of principal components artificial variables that will account for most of the variance in the observed variables. PCA finds a linear projection of high dimensional data into a lower dimensional subspace such that the variance retained is maximized and the least square reconstruction error is minimized [10].

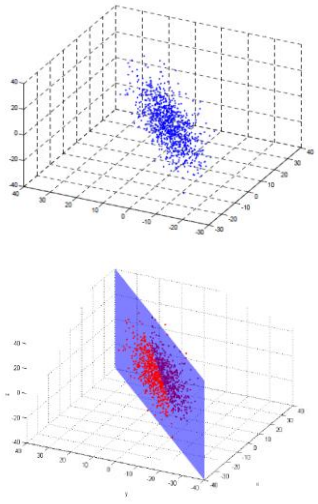

Figure 2: PCA Concept

PCA Steps are as follows [10]:
PCA: transforms an $N \times d$ matrix X into an $N \times m$ matrix Y

1) Subtract the mean from each of the dimensions: This produces a data set whose mean is zero. Subtracting the mean makes variance and covariance calculation easier by simplifying their equations. The co-variance and variance values are not affected by the mean value.
2) Calculate the $d \times d$ covariance matrix:

$$C = \frac{1}{N-1} X^T X$$

3) Calculate the eigen vectors and eigen values of the covariance matrix: If A is a square matrix, a non-zero vector v is an eigenvector of A if there is a scalar λ (eigen value) such that, $Av = \lambda v$.
4) Reduce dimensionality and form feature vector: The eigen vector with the highest eigen value is the principal component of the dataset.
5) Once eigen vectors are found from the covariance matrix, the next step is to order them by eigen value, highest to lowest. This gives the components in order of significance. Feature Vector=(λ1 λ2 λ3… λr)
6) Derive the new data: Final Data=Row Feature Vector x Row Zero Mean Data

Row Feature Vector is the matrix with the eigen vectors in the columns transposed so that the eigen vectors are now in the rows, with the most significant eigen vector at the top [11]. Row Zero Mean Data is the mean adjusted data transposed i.e. the data items are in each column, with each row holding a separate dimension. Final Data is the final data set, with data items in columns and dimensions along rows.

Reconstruction of Original Data:
Remember that,
Final Data=Row Feature Vector x Row Zero Mean Data.
Then,
Row Zero Mean Data=Row Feature Vector^{-1} x Final Data.
And thus,
Row Original Data=(Row Feature Vector^{-1} x Final Data) + Original Mean

If it reduces the dimensionality, obviously, when reconstructing the data then it loses those dimensions that were chosen to discard.

Assumptions and Limits

- Linearity: Linearity takes the problem as a change of basis. Several areas of research have explored how applying a non linearity prior to performing PCA could extend this algorithm which has been termed kernel PCA [12].
- Mean and variance are sufficient statistics [12]: The formalism of sufficient statistics captures the notion that the and variance the mean entirely describe a probability distribution. The only class of probability distributions which are fully described by the first two moments are exponential distributions e.g. Gaussian, Exponential etc.
- Large variances have important dynamics [13]: This assumption also includes the belief that the data has a high SNR. Hence, principal components associated with larger variances represent interesting dynamics, while those with lower variances represent noise.
- The principal components are orthogonal [13]: This assumption provides an intuitive simplification which makes PCA linked with linear algebra decomposition techniques.

Limits and Extensions of PCA
Both the strength and weakness of PCA is that it is a non-parametric analysis [9]. There are no parameters to tweak and no coefficients to adjust based on user experience hence the answer is unique and independent of the user, this is the strength of PCA. This same strength can be viewed as a weakness. If in case one knows a-priori some features of the

structure of a system, then it makes sense to incorporate these assumptions into a parametric algorithm or an algorithm with selected parameters.

Applications of PCA [14]
PCA is used for
a) data compression,
b) Dimensionality Reduction,
c) quality control, and
d) deriving geophysical parameters for The Atmospheric InfraRed Sounder (AIRS).

B. Linear Discriminant Analysis (LDA)

One of the most popular dimensionality reduction technique is the Linear Discriminant Analysis (LDA). LDA searches for the project axes on which the data points of different classes are far from each other while requiring data points of the same class are closer to each other [15]. The optimal transformation or projection of LDA can be computed by applying an eigen decomposition on the scatter matrices of the given training data. LDA has been abundantly used in many applications such as text processing and face recognition. However, the scatter matrices are dense and the eigen decomposition could be expensive in both time and memory for high-dimensional large-scale data. Moreover, to get a stable solution of LDA, the scatter matrices are required to be nonsingular, that is not true when the number of features is larger than the number of samples [16].
LDA is used to estimate a linear combination of features that can better separate two or more classes. The LDA finds such direction 'a' which provide maximum linear separation of classes. An example of a data projection on directions 'a' and 'b' is given in Fig. 3. There are generally many possibilities for finding directions but only some are optimal for data discrimination.

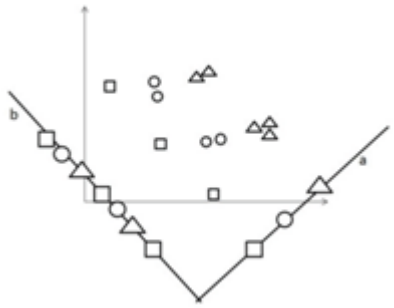

Figure 3: LDA finds a direction 'a' that maximize the separation of data

Steps of LDA [17]:

1) A measure of data separation can be given as the maximum of separation coefficient *F* (1):
$$F = \frac{tr(S_m)}{tr(S_w)}$$
where *Sm* gives the between-class scatter, *Sw* within class scatter. The bigger the value of *F* (1) the grater probability of classes separation.
2) Let us assume that there *C* classes, each containing *N* observations *xi*. The measure of within-class scatter *Sc* for the class *c* can be estimated as:
$$S_c = \sum_{i=1}^{N} (x_i^c - \mu^c)(x_i^c - \mu^c)^T$$
where μ^c indicates mean of the all observations x_i for *c-th* class.
3) Generalization *Sw* of the within class scatter for all *C* classes can be calculated as:
$$S_w = \sum_{i=1}^{c} \frac{n_i}{N} S^i$$
where *ni* is the number of *xi* observations in each class and *N* is a total number of all observations.
4) The value of between class scatter for class c can be calculated as:
$$S_b^c = \sum_{i=1}^{c} (\mu^i - \mu)(\mu^i - \mu)^T$$
where μi indicates the mean of the all observations *xi* for i-th class and μ indicates the mean of the all observations *xi* for all classes.
5) Generalization of between class scatter *Sm for* all *C* classes can be expressed as:
$$S_m = \sum_{i-1}^{c} \frac{n_i}{N} S_b^i$$
where *ni* means the number of *xi* observations in each class and N is a total number of all observations.
6) It can be proved that directions providing the best class separation are eigenvectors with the highest eigen values of matrix:
$$S = S_w^{-1} S_m$$
7) Generally the matrix *S* is not a symmetric matrix and calculation of eigenvectors can be difficult task. This problem can be solved by using generalized eigen value problem. A transformed data set can be obtained by:
$$y = x^T W$$
where *W*=[*w1,w2,..,wM*] is a matrix build with the *M* eigen vectors of matrix *S* connected with the highest eigen values. LDA reduces the original feature space dimension to *M*. A new data set *y* is created here as a linear combination of all input features *x* with weights *W*.

Advantages of LDA

- Multiple dependent variables [18].
- Reduced error rates.
- Easier interpretation of Between-group Differences: each discriminant function measures something unique and different.

Limitations of LDA [19]

- LDA implicitly assumes Gaussian distribution of data.
- LDA implicitly assumes that the mean is the discriminating factor, not variance.
- LDA may overfit the data.

3. Applications of LDA

- Linear Discriminant Analysis techniques are used in statistics, pattern recognition, and machine learning to find a linear combination of features which characterizes or separates two or more classes of objects or events [20].

- In LDA, resulting combination may be used as a linear classifier or more commonly in the dimensionality reduction [20].

4. Conclusion

The two Dimensionality reduction techniques wise Principal Component Analysis, Linear Discriminant Analysis are studied. From the review, it is clear that both of these techniques give good results for dimensionality reduction.

5. Acknowledgment

This is to acknowledge and thank all the individuals who played defining role in shaping this Conference paper. Without their Coordination, guidance and reviewing this task could not be completed alone. I avail this opportunity to express my deep sense of gratitude and whole hearted thanks to my guide Prof. J. S. Rangole madam for giving her valuable guidance, inspiration and encouragement to embark this paper. I also take great pleasure in thanking Dr. A. Das Sir for providing appropriate guidance. I would personally like to thank Prof. V.U. Deshmukh Sir, Head of Electronics and Telecommunication Dept. at Vidya Pratistaisthan's College of Engineering (VPCOE), and our Honble principal Dr. S. B. Deosarkar sir who creates a healthy environment for all of us to learn in best possible way.

References

[1] Shuiwang Ji, Jieping Ye, "Generalized Linear Discriminant Analysis: A Unified Framework and Efficient Model Selection", IEEE Transactions On Neural Networks, Vol. 19, No. 10, October 2008.
[2] K. Fukunaga, "Introduction to Statistical Pattern Recognition", 2nd ed. San Diego, CA: Academic, 1990.
[3] R. E. Bellman, "Adaptive Control Processes: A Guided Tour. Princeton", NJ: Princeton Univ. Press, 1961.
[4] M. Turk, A. Pentland, "Eigenfaces for Recognition", Journal of Cognitive Neuroscience,Vol. 3, No. 1, pp. 71-86, 1991.
[5] W. Zhao, R. Chellappa, A. Krishnaswamy, "Discriminant Analysis of Principal Components for Face Recognition", Proc. of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, pp. 336-341, April 1998.
[6] Aleix M. Martinez, Avinash C. Kak, "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 2, February 2001.
[7] Jonathon Shlens, "A Tutorial on Principal Component Analysis", December 10, 2005.
[8] Jolliffe, I.T., "Principal Component Analysis", Springer- Verlag, New-York, 1986.
[9] Rummel, R. J., " Applied factor analysis", Evanston, IL: Northwestern University Press, 1970.
[10] Aly A. Farag, Shireen Elhabian, "A Tutorial on Principal Component Analysis", University of Louisville, CVIP Lab, Sep. 2009.
[11] Corman, T. H., C. E. Leiserson, R. L. Rivest, "Introduction to algorithms", McGraw-Hill, New York, 1997.
[12] Stevens, J., "Applied multivariate statistics for the social science", Hillsdale, NJ: Lawrence Erlbaum Associates, 1986.
[13] Mark Richardson, "Principal Component Analysis", May 2009.
[14] Huang, H-L and P. Antonelli, "Application of principal component analysis to high-resolution infrared measurement compression and retrieval", J. Appl. Meteor., 40, 365-388, 2001.
[15] Deng Cai, Xiaofei He, and Jiawei Han, "SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis", IEEE Transactions on Knowledge And Data Engineering, Vol. 20, No. 1, January 2008.
[16] P. Howland and H. Park, "Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 26, no. 8, pp. 995-1006, Aug. 2004.
[17] Marcin Kolodziej, Andrzej Majkowski, Remigiusz J. Rak, "Linear Discriminant Analysis As EEG Features Reduction Technique For Brain-Computer Interfaces", Warsaw University Of Technology, PRZEGLĄD ELEKTROTECHNICZNY (Electrical Review), ISSN 0033-2097, R. 88 NR 3a/2012.
[18] K. Etemad and R. Chellapa, "Discriminant Analysis for Recognition of Human Face Images", J. Optics of Am. A, vol. 14, no. 8 pp. 1724-1733, 1997.
[19] Lehrstuhl Sprachsignal Verarbeitung, "Linear discriminant analysis".
[20] Suman Kumar Bhattacharyya, Kumar Rahul, "Face Recognition By Linear Discriminant Analysis", International Journal of Communication Network Security, ISSN: 2231 – 1882, Volume-2, Issue-2, 2013.