ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

Comparative Analysis of Dimensionality Reduction Techniques

Dr. S.Vijayarani, S. Maria Sylviaa

Assistant Professor, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, India[1]

M.Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, India[2]

ABSTRACT: Datasets are most important for performing all the type of data mining tasks. Every dataset has many numbers of attributes and instances. Dimensionality reduction (DR) is one of the preprocessing steps which is used to reduce the dimensions (attributes or features) without losing the data. There are two divisions of reduction they are feature extraction and feature reduction. Feature extraction is the process of decomposition of attributes of the original data (i.e.) merging the attributes of the data Feature selection is the process of selecting the subset of attributes by eliminating features with little or no predictive information. Feature extraction techniques are more adequate than the feature selection. Reduction is done to the larger dataset to decrease the curse of dimensionality. The main objective of this paper is to provide a systematic comparative analysis on feature reduction algorithms such as PCA, LDA and FA to medical dataset (Thyroid, Oesophagal).The performance factor considered are number of attributes reduced and time is observed.

KEYWORDS: Dimensionality reduction, Feature extraction, PCA, LDA, FA

I. INTRODUCTION

Number of attributes required for implementing data mining techniques is differs from application to application. All the attributes in the dataset is not utilized by the data mining algorithm. To perform the data mining tasks for example: clustering, classification, association rule, etc datasets plays a significant role. Every data set is having many numbers of attributes and instances. The number of attributes required for performing the data mining tasks is differed from application to application. All the attributes are not used for implementing the data mining algorithms. If we consider all the attributes, it will increase the execution time and occupies the program memory space unnecessarily. In order to avoid this, before performing any data mining task, we have to use dimensionality reduction techniques. Dimensionality reduction is one of the preprocessing steps used in number of application to reduce the dimensions of high dimensional data to increase the efficiency of the data analysis. The divisions of dimensionality reduction are Feature extraction (FE) or reduction and Feature selection (FS). Feature extraction is the process of decomposition of attributes of the original data (i.e.) merging the attributes of the data. Some applications of feature extraction are semantic analysis [1], data compression, data decomposition, projection and pattern recognition and it enhances the speed and effectiveness of supervised learning [1]. Feature selection is the process of selecting the subset of attributes by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points[2].The problem of high dimensional data is that, it needs many number of features for performing data mining techniques. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [3]. Before indexing the data it should be reduced which in turn performs the task efficiently. Dimensionality reduction is important in many domains, since it mitigates the curse of dimensionality and other undesired properties of high-dimensional spaces [4].There are numerous algorithms used in DR to reduce the attributes they are Principal Component Analysis, Linear Discriminant Analysis and Factor Analysis,etc.

DOI: 10.15680/IJIRCCE.2016. 0401006

The remaining section of this paper is structured as follows, Section2 illustrates about the review of literature, Section3 describes how the feature extraction algorithms (Principal Component Analysis, Linear Discriminant Analysis and Factor Analysis) are used to reduce the features Section4 discusses the experimental results and conclusions are given in Section5

II. RELATED WORK

Lauren van den et.al.[6] has author presented the 12 non-linear dimensionality reduction techniques, they are kernel principal component analysis (Kpca), Isomap, Maximum Variance Unfolding, diffusion maps, Locally Linear Embedding(LLE), Laplacian Eigen maps, Hessian LLE,Local Tangent Space Analysis, Sammon mapping, multilayer auto encoders, Locally Linear Coordination and manifold charting. These techniques are applied to both artificial and natural datasets. There are five different artificial dataset (Swiss roll dataset, helix dataset, twin peaks dataset, broken Swiss roll dataset and the high-dimensional dataset) used to investigate how the data lies on lower dimensional manifolds. The five natural datasets are(MNIST dataset contains 60000 handwritten digits in this 5000 digits were selected, COIL20 dataset contains 784 dimensions with 28x28 pixel images of 20 different objects , NiSIS dataset consist of 3,675 grayscale images, ORL dataset contains 400 grayscale images(faces), and the HIVA dataset contains 3,845 data point. The experimental results have explained the weakness, continuity and performance of the techniques.
Rekha Aswathi et.al [7] has performed normalization which is used to standardize all the features in the dataset and dimensionality reduction to perform clustering. Here, the diabetic’s dataset which contains 768 instances and 8 plus class attributes has been taken and PCA algorithm is used to reduce the dimensions. Out of 8 features, 4 features are selected without the loss of information.WEKA3.7 tool is used to investigate the diabetes data. After performing dimensionality reduction density based clustering algorithm is used to find the maximal set of density. Dimensionality reduction is used to increase the accuracy of the clustered data.

Liliana Ferrier et.al. [8] has compared two different feature extraction algorithms. The features of the products based on the review of the customer, is considered as the dataset. In the first algorithm the candidate features are identified and they are pruned. In the second approach association rule mining is used to find the frequent pattern.Here,the dataset is based on the customer review which are collected from the social website such as amazon,cnet and it is based on five different products(two digital cameras, a DVD player, an MP3 player and a cell phone). Likelihood Ratio Test is the method used to extract the features of the product.

Edwige Fangseu Badjio et.al. [14] has presented the methods for visual data mining in order to mine the data and to make cognitive. Here, the author has performed the attribute selection method i.e. wrapper method and filter method. Here, seven types of dataset (lung cancer, promoter, sonar, Arrhythmia, Colon Tumor, and CentralNervSyst) have been used and the accuracy has been calculated before reduction and after reduction. In this reduction framework, the numbers of attributes have been reduced. The data visualization is represented in order to determine the relationship between the data. The algorithms like (LDA, QDA, and KNN) have been used in this work and found that LDA have performed efficiently and reduces the attributes effectively.

III. METHODOLOGY

Dimensionality reduction is the important factor which is used to reduce the features of the original data without the loss of information. The main objective of this analysis work is to compare the three different feature reduction algorithms namely PCA, LDA and FA. The architecture is as follows\

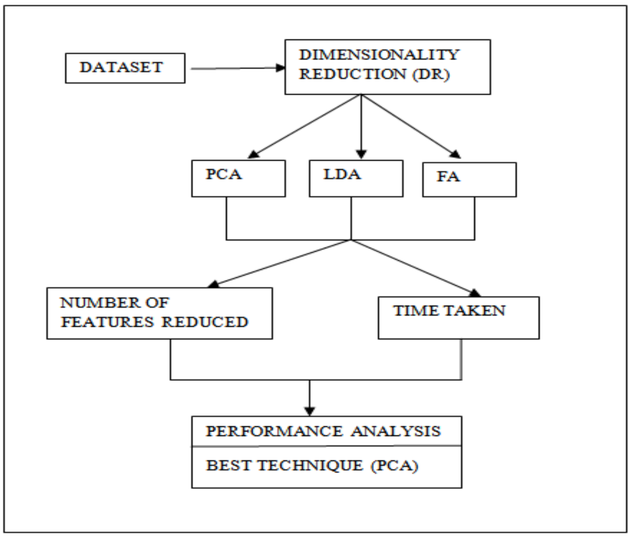


Figure.1. Architecture of dimensionality reduction

A. DATASET

Data set used in this work are-Thyroid dataset, Oesophagal dataset. Thyroid dataset is collected from KEEL Data Repository. It contains 7200 instances and 22 attributes whereas, Oesophagal data set is collected from SMCR repository which contains 979 instances and 13 attributes.

B. FEATURE EXTRACTION

Dimensionality reduction is of two types. They are feature extraction (FE), feature selection (FS).Here; FE is used to reduce the features. Feature extraction is the process of decomposition of attributes of the original data (i.e.) merging the attributes of the data. High dimensional data normally requires lot of memory and power consumption. In FE technique the large numbers of attributes are merged together based on the algorithms used and they are converted into lower dimensional space.

C.PCA

Principal Component Analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components [9]. PCA algorithm is used in number of applications to reduce the features and transforms the higher dimensional space into lower space. PCA is used for analyzing the data for prediction. First the data matrix is taken as input. Then the mean x is subtracted from the original data X, then the covariance is calculated for matrix X, then Eigen vector and Eigen values are calculated. From the values the largest Eigen value is found to be the principal component. The PCA algorithm is given in Table1.

Table.1. PCA Algorithm

```
STEP 1: X ⟵ Create N×d data matrix with one row vector $x_n$ per data point
STEP 2: X subtract mean from each row vector $x_n$ in X
STEP 3: $\Sigma$ ⟵ Covariance matrix of X
STEP 4: Find Eigen vectors and Eigen values of $\Sigma$
STEP 5: Principal Component ⟵ the M Eigen vector with largest Eigen values.
```

The advantages of PCA are the PC's are uncorrelated and the first component explains about the largest percentage in the n dimensional dataset and the second component with next largest percentage and so on.

D.LDA

Linear discriminant Analysis(LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events[10].LDA is used to reduce the dimensions of the data set. In discriminant analysis, the dependent variable Y is the group and independent variables X are the features of the group. Where W^T represents the class

$$\boxed{Y = W^T.X}$$

The goal is to project a dataset into a lower-dimensional space with good class-separability in order to avoid over fitting ("curse of dimensionality")[11] and also reduce computational costs [11].The advantages of LDA is that it reduces the error rate and they are interpreted easily between data groups. The LDA algorithm is given in Table 2.

Table.2. LDA Algorithm

1. Compute the *d*-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\mathbf{e_1}, \mathbf{e_2} ... \mathbf{e_d}$) and corresponding eigenvalues ($\lambda_1, \lambda_2 ... \lambda_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing Eigen values and choose **k** eigenvectors with the largest eigenvalues to form a $d \times k$-dimensional matrix **W** (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the equation $\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$ (where $\boldsymbol{X}$ is an $n \times d$-dimensional matrix; the i^{th} row represents the j^{th} sample, and $\boldsymbol{Y}$ is the transformed $n \times k$-dimensional matrix with the n samples projected into the new subspace).

First the dataset is taken as input then the scatter matrix is computed between the classes. Then the Eigen vectors and Eigen values are computed based on the scatter matrix, and it forms d x k dimensional matrix, number of Eigen vectors is chosen by decreasing the largest Eigen values. d x k Eigen vector matrix then transforms sample space into new subspace.

A.FACTOR ANALYSIS

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors [12].Factor analysis simplifies the data.FDA algorithm is given in Table 3.

Table.3. Factor Analysis Algorithm

1. The correlation matrix of the input raster maps is computed.
2. The orientation of the factors is computed based on the eigenvectors and eigenvalues of the correlation matrix.
3. The vectors in each column are normalized. The normalized values are the factor analysis coefficients.
4. A map list is created which contains an expression with which the transformed and raster maps (factors) can be defined and calculated.
5. When the map list is opened, the pixel values of the input maps are transformed into the new raster maps (factors).

First, the matrix is taken as input then the correlation matrix is calculated. The orientation factor is computed based on the Eigen vectors and Eigen values. Every column's vector is normalized inorder to equalize the data and the normalized values are denoted as factor analysis coefficient. The factor list is created with expression and they are transformed into new factors. The advantage of factor analysis is that there is no distinction between dependent and independent variables rather all variables are analyzed together to identify factors [13].

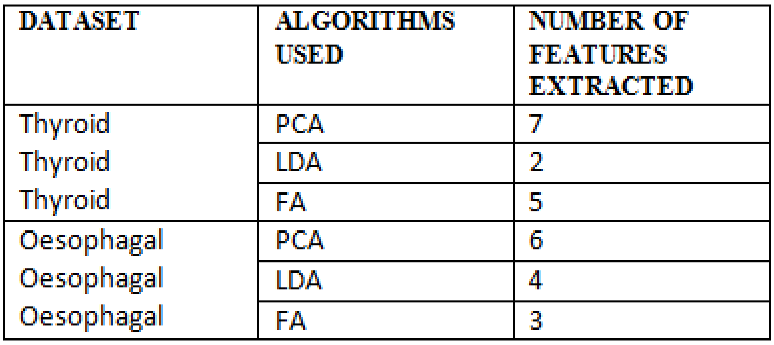
IV. EXPERIMENTAL RESULTS

The implementation has been done in MatlabToolboxv2.5 (2010b).In order to evaluate the accuracy of algorithm the factors like Tic toc time and number of features are considered.

Features Extracted

The feature extraction is illustrated in Table.4. It describes about the number of features extracted by Principal Component Analysis, Linear Discriminant Analysis and Factor Analysis for thyroid and Oesophagal dataset. The feature extracted is based on number of input features.

Table.4. Number of features extracted



DATASET	ALGORITHMS USED	NUMBER OF FEATURES EXTRACTED
Thyroid	PCA	7
Thyroid	LDA	2
Thyroid	FA	5
Oesophagal	PCA	6
Oesophagal	LDA	4
Oesophagal	FA	3

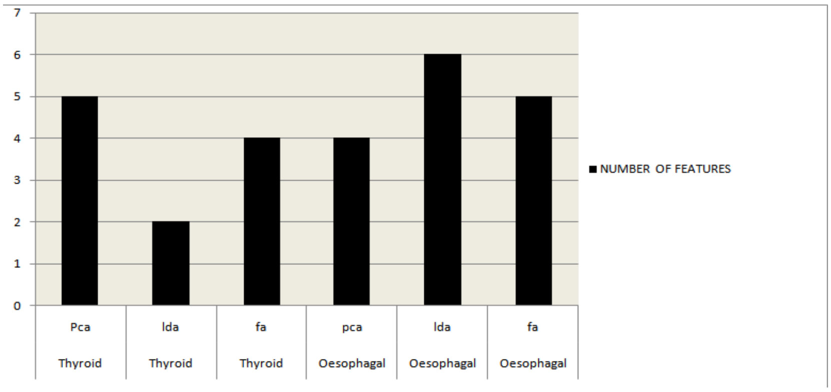


Figure.2. Features extracted

Execution Time

The two factors such as Tic toc time and clock time are taken. Tic toc time is defined to be the elapsed time in Matlab. Clock is the display time

Table.5. Time taken to extract the features

DATASET	ALGORITHMS USED	TICTOC(sec)	CLOCK(sec)
Thyroid	PCA	1.00665	1.01
Thyroid	LDA	1.001172	1.02
Thyroid	FA	1.00917	1.01
Oesophagal	PCA	1.00928	1.01
Oesophagal	LDA	1.01031	1.01
Oesophagal	FA	1.0136	1.01

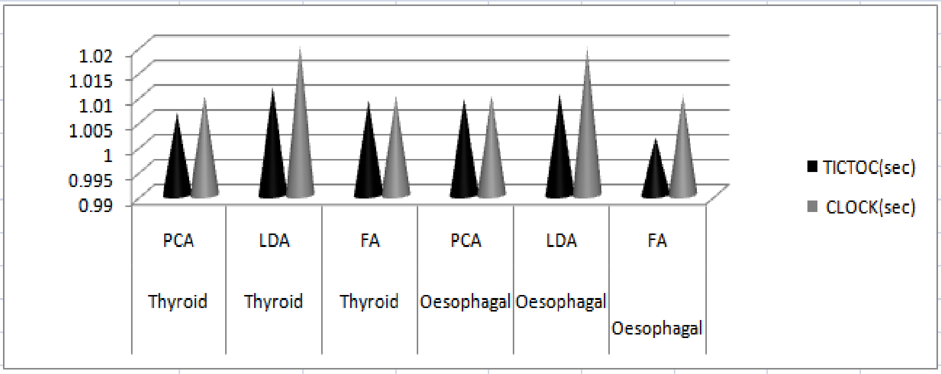


Figure.3. Time taken to extract the feature

V. CONCLUSION

Feature extraction is the concept of combining the related features together and eliminating other features. Reduction of features is useful to increase the accuracy of data mining techniques. In this work the performance of three different feature extraction algorithms PCA, LDA, and FA have been analyzed. Time taken and number of features reduced are considered as the performance measures. From the analysis, it is observed that PCA performance is more efficient than other algorithms.

REFERENCES

1. http://www.comp.dit.ie/btierney/Oracle11gDoc/datamine.111/b28129/feature_extr.html
2. YongSeog Kim, W. Nick Street, and Filippo Menczer, University of Iowa, USA- Feature Selection in Data Mining
3. K. Fukunaga. Introduction to Statistical Pattern Recognition. Academic Press Professional, Inc., San Diego, CA, USA, 1990.
4. R. Kharal. Semidefinite embedding for the dimensionality reduction of DNA microarray data. Master's thesis, University of Waterloo, 2006.
5. Dr. S. Vijayarani1, Ms.P.Jothi2-partitioning clustering algorithms for data stream outlier Detection-International journal of innovative research in computer and communication Engineering-vol. 2, issue 4, April 2014
6. Laurens van der Maaten Eric Postma-Dimensionality Reduction: A Comparative Review- Tilburg centre for Creative Computing, Tilburg University 5000 LE Tilburg, http://www.uvt.nl/ticc
7. Rekha Awasthi, Anil Kumar Tiwari and Seema Pathak-An Analysis Of Density Based Clustering Technique With Dimensionality Reduction For Diabetic Patient-International Journal of Computer Engineering and Applications, Volume IX, Issue IV, April 15 www.ijcea.com ISSN 2321-3469.
8. Liliana Ferreir, aNiklas Jakob and Iryna Gurevych-A Comparative Study of Feature Extraction Algorithms in Customer Reviews-https://www.informatik.tudarmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2008/AComparativeStudyOfFeatureExtraction.pdf
9. https://en.wikipedia.org/wiki/Principal_component_analysis
10. https://en.wikipedia.org/wiki/Linear_discriminant_analysis
11. http://sebastianraschka.com/Articles/2014_python_lda.html
12. https://en.wikipedia.org/wiki/Factor_analysis
13. http://www.slideshare.net/kompellark/t19-factor-analysis
14. Edwige Fangseu Badjio, Francois Poulet-Dimension Reduction for Visual Data Mining-ESIEA Recherche Parc Universitaire de Laval-Chang´e

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 1, January 2016

BIOGRAPHY

Dr. S. Vijayarani, MCA, M.Phil, Ph.D., working as Assistant Professor in the Department of Computer Science, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. Maria Sylviaa.S has completed Master of Computer.Applications. She is currently pursuing her M.Phil in Computer Science in the Department of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining and Network Security.

DOI: 10.15680/IJIRCCE.2016. 0401006