IMPACT OF THE USAGE OF DIMENSIONALITY REDUCTION TECHNIQUES IN MEDICAL DATASETS: AN ANALYSIS

S. MARIA SYLVIAA, DR. S. VIJAYARANI

Abstract: Dimensionality reduction (DR) is used to reduce the number of attributes to make the task to be performed efficiently without any data loss. Dimensionality reduction is one of the preprocessing steps used in data mining to improve the efficiency of the task. Dimensionality reduction is performed in two different ways; they are feature extraction and feature selection. The primary goal of this research work is to compare the efficiency of four dimensionality reduction algorithms. They are MDS (Multi-Dimensional Scaling), Auto Encoder, ProbPCA (Probabilistic Principal Component Analysis) and CFA (Common Factor Analysis). Dimensionality reduction is not only useful to speed up algorithm execution, but actually might help with the final classification/clustering accuracy as well [13]. Three different medical datasets, Thyroid, Oesophagal and Heart disease datasets are used for experimentation. The performance factors used to analyze the efficiency of these algorithms are number of attributes reduced and execution time. From the results, it is proven that MDS algorithm performance is better than other algorithms. In addition to this, PAM clustering accuracy and time required to perform the clustering process is also identified. For the original data set, considering all the attributes (i.e. before performing the dimensionality reduction (PRE-DR)) the PAM clustering accuracy and execution time is calculated and this is compared with the PAM clustering accuracy and execution time for the new data set (i.e. after performing the dimensionality reduction (POST-DR)). From this, we have observed that the clustering accuracy is same for PRE-DR and POST-DR and POST-DR needs minimum time to perform the clustering process.

Keywords: Dimensionality Reduction, MDS, auto encoder, ProbPCA, CFA, PAM clustering.

Introduction: Datasets play an important role to perform various data mining tasks like clustering, classification, association rule, neural networks, etc. Every data set is having many numbers of attributes and instances and the number of attributes required for performing the data mining tasks is differed from application to application [1].To implement various data mining techniques number of attributes are enforced. In most of the applications, not all the attributes are required to perform the task [15]. If all attributes are used to perform the task it will increase the time of execution and occupies more memory space .Hence, it is necessary to select the attributes or reducing the attributes and then it can be used for data mining tasks. Dimensionality reduction is used to reduce the number of attributes to make the task efficiently without losing the data. Dimensionality Reduction is of two types they are feature extraction and feature selection. Feature Extraction (FE) is also known as Feature Reduction and it extracts the necessary attributes, from this it creates a subset of original attributes. Feature Selection (FS) is the process of selecting the required subset of attributes by eliminating unwanted attributes by using the original features. Some of the applications of feature extraction are semantic analysis [2], data compression, data decomposition, projection and pattern recognition [2].In order to improve the efficiency of the data mining tasks it is necessary to select the required attributes for processing or to reduce the unnecessary attributes. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points [3].

The primary objective of this work is to compare the efficiency of the dimensionality reduction techniques using medical datasets. The algorithms considered for analysis are MDS (Multi-Dimensional Scaling), Auto Encoder, ProbPCA (Probabilistic Principal Component Analysis) and CFA (Common Factor Analysis). Three different medical datasets, Thyroid, Oesophagal and Heart disease datasets are used for experimentation. The commonly used DRM methods are Principal Component Regression (PCR), Canonical Correlation Regression (CCR), RRR and Partial Least Squares (PLS World, 1966) [14].

The remaining section of this paper is organized as follows. Section2 illustrates the review of literature; Section3 describes Multidimensional Scaling, Auto Encoder, Probabilistic Principal Component Analysis and Common Factor Analysis algorithms.Section4 discusses the experimental results and conclusions are given in Section5

Review of Literature: Lauren van den et.al.[4] has presented the 12 non-linear dimensionality reduction techniques, they are kernel principal component analysis, Isomap, maximum variance unfolding, diffusion maps, locally linear embedding, 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 datasets (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, which 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 3675 grayscale images, ORL dataset contains 400 grayscale images(faces) and HIVA dataset contains 3845 data points. The experimental results have explained the weakness, continuity and performances of these techniques.

Michael et.al.[5]has presented four different nonlinear dimensionality reduction algorithms to eight membered rings to demonstrate the importance of nonlinear correlations in molecular motion. For high-dimensional encodings ranging from 8 to 276 dimensions, these algorithms are able to provide low-error embeddings within the theoretical limit of 5 dimensions. Smaller molecular motions have been chosen in order to test the samples by avoiding sampling issues. The performance of LLE and Isomap was very similar for the case presented. LLE is attractive from a theoretical standpoint in that only local euclidean distances are considered. The auto encoder performed best at low dimensionalities, generates fast explicit forward and reverse maps, and considers reconstruction error explicitly in the objective function.

Mojie Duan et.al.[6] has presented three dimensionality reduction algorithms PCA, LLE, Isomap to Protein conformation space (C Space).This work has given the puzzling dimensionally reduction results of β-hairpin has been considered and the linear method PCA is performed better than nonlinear methods, i.e. Isomap and LLE. RMSD Isomap were computed between every pair of atom coordinate vectors and has chosen 20 closest neighbors for each conformation to build a graph, with RMSD values as edge weights PCA essentially uses an explicit section to represent the protein C Space, while the RMSD-based Isomap and LLE essentially use an implicit representation.

Methodology: Dimensionality reduction is the important in data mining which is used to reduce the features of the original data without any loss of information. The main objective of this research work is to compare the four different attribute reduction algorithms namely MDS, Auto Encoder, PPCA and CFA. Figure 1 shows the proposed system architecture of this work.

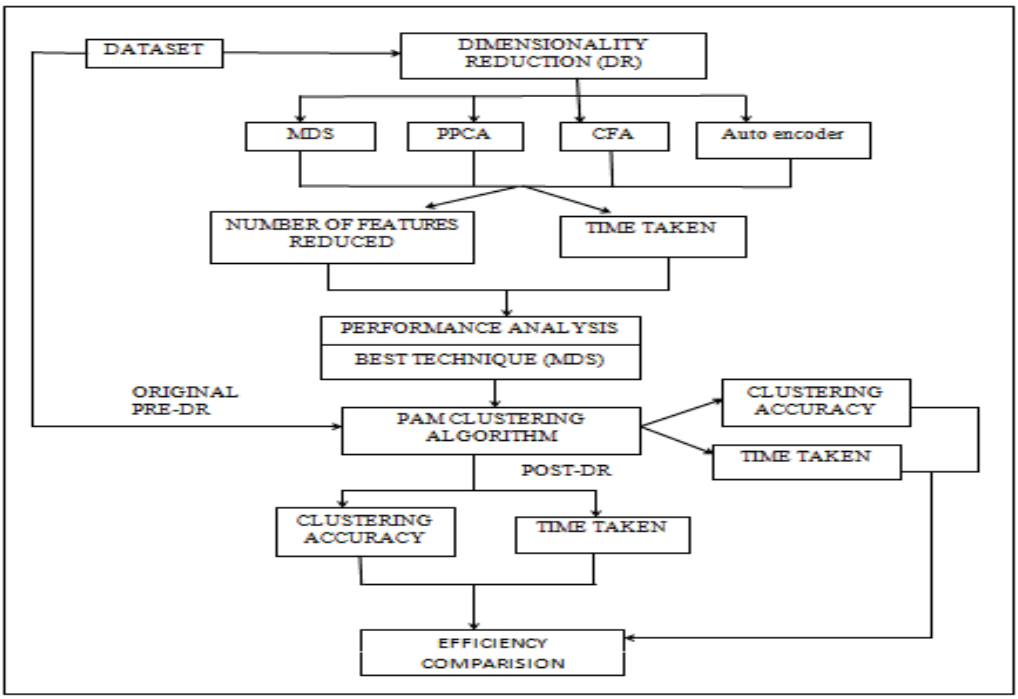

Figure.1. System Architecture

Dataset: Medical data sets used in this work are, Thyroid dataset, Oesophagal dataset and Heart disease dataset. Thyroid dataset is collected from KEEL Data Repository. It contains 7200 instances and 21 attributes whereas, Oesophagal data set is collected from SMCR repository which contains 979 instances and 13 attributes. Heart disease dataset is collected from UCI repository. It contains 366 instances and 12 attributes.

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

MDS [7]: Multi-Dimensional Scaling algorithm is the procedure of viewing the level of coincidence among the dataset.MDS algorithm aims to place each object in N-dimensional space such that the between-object distances are preserved as well as possible and each object is then assigned coordinates in each of the N dimensions [7].The matrix d value is taken as output and squared matrix is calculated. J value is calculated in order to apply the double centering and the largest Eigen value is extracted and then the diagonal matrix is calculated to extract the resultant. The pseudo code of MDS algorithm is given in Table1.

MDS Algorithm;

STEP 1: Set up the squared proximity matrix $D^2 = [d^2_{ij}]$

STEP2: Apply double centering: $B=-1/2\ J\ D^{(2)}\ J$ using the centering matrix $J=I-1/n\ 11'$, where n is the number of objects.

STEP3: Extract the m largest Eigenvalue $\lambda_1,\ \lambda 2 \ldots \lambda_m$ of Band Corresponding eigenvectors $e_1, e_2 \ldots e_m$.

STEP4: Now $X=E_m \Lambda_m^{1/2}$ where E_m is the matrix of m eigenvectors and Λ_m is the diagonal matrix of m Eigen values of B. Classical MDS assumes Euclidean distances

PPCA [8]: Probabilistic Principal Component Analysis is equivalent to General Principal Component analysis which is a statistical procedure. It 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 [8]. Maximum likelihood is the estimator and it is calculated to find the probabilistic value. Normal PCA is a limiting case of probabilistic PCA, taken as the limit as the covariance of the noise becomes infinite similarly small ($\psi = \lim_{\sigma 2 \to 0} \sigma^2 I$) [9]. It captures dominant correlations with few parameters and multiple PCA models can be combined as a probabilistic mixture. Here y is taken as input data and the latent space is calculated by the relation R. Using the Gaussian distribution the covariance β is calculated.

The PPCA algorithm is given in

PPCA Algorithm:

STEP 1: Consider a set of centered data of n observations and d dimensions: $Y = [y.\ y_n]^T$

STEP 2: We assume this data has a linear relationship with some embedded latent space data x_n Where $Y \epsilon R^{Nxd}$ and $x \epsilon R^{Nxq}$

STEP 3: $Y_n = W\ x_{n+n}$, where x_n is the q-dimensional latent variable associated with each observation and ϵR^{dxq} is the transformation matrix relating the observed and latent space.

STEP 4: We assume a spherical Gaussian distribution for the noise with a mean of zero and a covariance of β-1

STEP 5: Likelihood for an observation y_n is: $P\ (y_n \mid x_n\ W, \beta)$

CFA [10]: CFA is nothing but Common Factor Analysis or Coordinated Factor Analysis. It is based on the fundamental assumption that some underlying factors, which are smaller in number than the observed variables, are responsible for the co-variation among them. Factor analysis is a statistical approach that can be used to analyze interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions [10].CFA algorithm is same like FA. The CFA algorithm is given in Table3.First, the data is collected and 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. By calculating the factors the rotation and interpretation of the data is done. The factor list is created with expression and they are transformed into new factors.

CFA Algorithm

STEP 1: Data collection and generation of the correlation matrix

STEP 2: Extraction of initial factor solution by computing the orientation of the factor based on the Eigen vectors and Eigen values of the correlation matrix

STEP 3: Rotation and interpretation (also validation)is done by calculating the factors

STEP 4: Construction of scales or factor scores to use in further analyses

STEP 5: When the map list is opened, the pixel values of the input maps are transformed into the new Raster maps (factors).

Autoencoder: The auto encoder algorithm and its deep version as traditional dimensionality reduction methods have achieved great success via the powerful represent ability of neural networks [11].It is a dimensionality reduction from manifold learning. Auto encoder consists of two parts they are encoder x_i and decoder y_i .Parameters (W, W_o).Weight S_i is calculated from $x_{i,}$, equation is minimized using stochastic gradient feature.

The hidden representation is computed using $\{y_i\}$ n ₁and the notations are updated. The Auto encoder algorithm is given in Table4.

Auto encoder Algorithm

Input: training set $\{x_i\}$ n

Parameters: $\Theta = (W, W_o)$

Notation: Ωi: reconstruction set for x_i

Si: the set of reconstruction weight for x_i $\{y_i\}$

$n_{1:}$ hidden representation

STEP 1: Compute the reconstruction weights S_i from $\{x_i\}$ n ₁ and determines the reconstruction set Ω_i

STEP 2: Minimize E in Eqn.4 using the stochastic gradient descent and update Θ for t steps

STEP 3: Compute the hidden representation $\{y_i\}$ n ₁ and update S_i and Ω_i from $\{y_i\}$ n ₁.

STEP 4: Repeat the step 2 and 3 until convergence

Experimental Result: 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 number of features reduced by dimensionality reduction techniques is illustrated in Table 5. It describes features extracted by MDS, Auto Encoder, ProbPCA, CFA to medical datasets (thyroid, Oesophagal, Heart).The feature extracted is based on number of input features.Figure2 depicts the original features of the dataset, number of features reduced by the dimensionality reduction technique and the time required to perform this reduction process.

Execution Time: Execution time is calculated in milliseconds. Tic Toc time is taken as execution time for performing dimensionality reduction.

Table.5. Number of features extracted

DATASET	ALGORITHM USED	ORIGINAL FEATURES	NUMBER OF FEATURES REDUCED	TIME TAKEN(milliseconds)
THYROID	MDS	22	11	1019
	PPCA	22	4	1003
	CFA	22	5	1017
	Auto encoder	22	5	1016
OESOPHAGAL	MDS	13	5	1011
	PPCA	13	4	1001
	CFA	13	5	994
	Auto encoder	13	4	994
HEART	MDS	12	6	1007
	PPCA	12	3	1008
	CFA	12	4	1007
	Auto encoder	12	3	1005

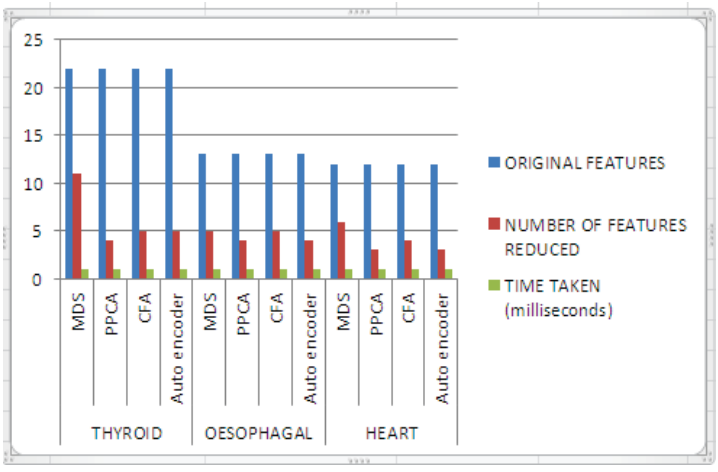

Figure.2. Features extracted

Clustering accuracy before dimensionality reduction: Pre-clustering is done to the dataset to find the accuracy of original features. The performance factors considered are clustering accuracy before dimensionality reduction and time taken.Table6 explains about the pre-clustering accuracy. Figure 3 explains about the dataset with

original features and the PAM clustering is performed to the dataset to find the accuracy and the time taken .

Table.6. Clustering Accuracy for Original Features

DATASET	ORIGINAL FEATURES	PAM CLUSTERING ACCURACY	TIME(milliseconds)
THYROID	22	66.99%	4148
OESOPHAGAL	13	40.58%	1094
HEART	12	66.66%	3618

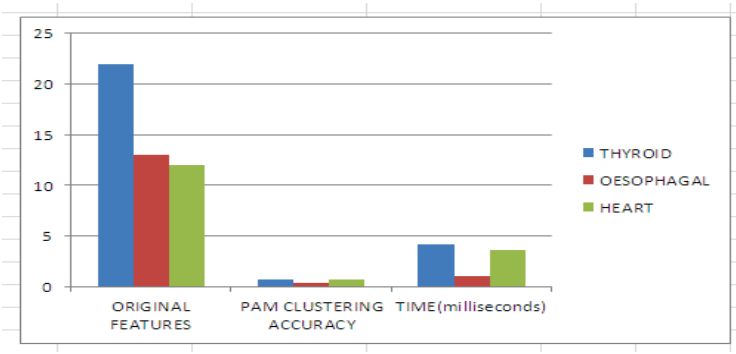

Figure.3. Clustering Accuracy for Original Features

Clustering accuracy after dimensionality reduction: Post-clustering is done to the dataset to find the accuracy of reduced features after performing the dimensionality reduction. The performance factors used for the comparison are clustering accuracy, Time taken is explained Table.7. Based on the reduced attributes Figure.4 describes the performance metrics like accuracy and execution time.

Table.7. Clustering Accuracy for Reduced Features

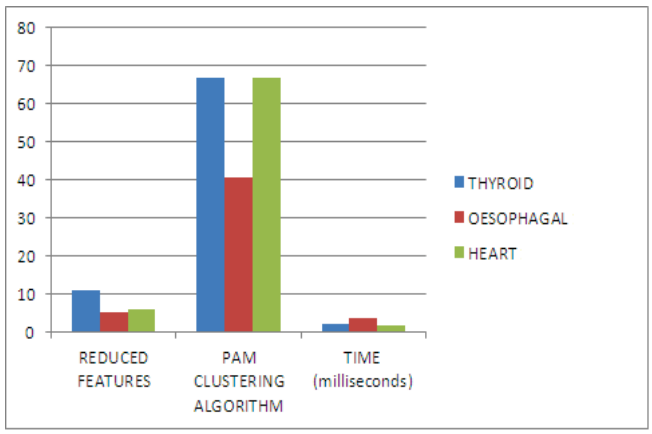

Figure.4. Clustering Accuracy for Reduced Features

Conclusion: Feature extraction is the concept of merging the related features together and eliminating other unwanted features. Reduction of features increases the accuracy of data mining techniques and also decreases the time consumption. In this work the performance of four different feature extraction algorithms MDS, PPCA, CFA and Auto encoder have been analyzed. Time taken and number of features reduced are considered as the performance measures. From the analysis, it is observed that MDS algorithms performance is more efficient than other algorithms. After using the dimensionality reduction techniques, we have analyzed the performance of the PAM clustering algorithm efficiency in the reduced data set. Results proved that the PAM clustering accuracy is same with the original, i.e. before reducing features

and also observed that the time required to perform the clustering is also minimized. In future the existing algorithms are to be enhanced and new algorithms are to be developed to extract more number of features and other clustering algorithms are used for the comparison.

Reference:

1. Dr. S. Vijayarani, S. Maria Sylviaa-Comparative Analysis of Dimensionality Reduction Techniques-International Journal of Innovative Research in Computer and Communication Engineering Vol. 4, Issue 1, January 2016
2. http://www.comp.dit.ie/btierney/Oracle11gDoc/datamine.111/b28129/feature_extr.html
3. Yong Seog Kim, W. Nick Street, and Filippo Menczer, University of Iowa, USA- Feature Selection in Data Mining
4. 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
5. W. Michael Brown,ShawnMartin, Sara N. Pollock,Evangelos A. Coutsias,4 and Jean-Paul Watson1,b- Algorithmic dimensionality reduction for molecular structure analysis- The Journal of Chemical Physics 129, 064118 (2008).
6. Mojie Duan, Li Han, Lee Rudolph and Shuanghong Huo- Geometric Issues in Dimensionality Reduction and Protein ConformationSpacehttps://www.cs.unm.edu/amprg/rss14workshop/PAPERS/Duan.pdf
7. https://en.wikipedia.org/wiki/Multidimensional_scaling.
8. https://en.wikipedia.org/wiki/Principal_component_analysis
9. https://people.cs.pitt.edu/~milos/courses/cs3750-Fall2007/lectures/class17.pdf
10. http://www.unt.edu/rss/class/mike/6810/FA.pdf
11. Wei Wang, Yan Huang, Yizhou Wang , Liang Wang- Generalized Auto encoder: A Neural Network Framework for Dimensionality Reduction [12].
12. http://www.cvfoundation.org/openaccess/content_cvpr_workshops_2014/W15/papers/Wang_Generalized_Autoencoder_A_2014_CVPR_paper.pdf.
13. https://www.knime.org/files/knime_seventechniquesdatadimreduction.pdf
14. Luigi D'Ambra1, Pietro Amenta2, and Michele Gallo3-Dimensionality Reduction Methods – Metodolo skizvezki, Vol. 2, No. 1, 2005, 115-123.
15. G. N. Rama devi, K.Usharani -Study on Dimensionality Reduction Techniques and Applications- Publications of Problems & Application in Engineering Research – Paper-http://Ijpaper.Com/ Csea2012 Issn: 2230-8547; E-Issn: 2230-8555

S. Maria Sylviaa,
Assistant Professor, Department of Computer Science,
Nirmala College for Women, Coimbatore, India
Dr. S. Vijayarani
Assistant Professor, Department of Computer Science,
School of Computer Science and Engineering,
Bharathiar University, Coimbatore, India