Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.8 No.S3, 2019, pp.66-71
© The Research Publication, www.trp.org.in
DOI: https://doi.org/10.51983/ajcst-2019.8.S3.2084

Performance Analysis of Dimensionality Reduction Techniques in the Context of Clustering

T. Sudha[1] and P. Nagendra Kumar[2]
[1]Professor, [2]Assistant Professor, [1&2]Department of Computer Science,
[1]Sri Padmavathi Mahila Visva Vidyalayam, Tirupati, Andhra Pradesh, India
[2]Geethanjali Institute of Science and Technology, Sri Potti Sreeramulu, Andhra Pradesh, India
E-Mail: thatimakula_sudha@yahoo.com, nagendra.gudur@gmail.com

(Received 12 April 2019; Revised 24 April 2019; Accepted 2 May 2019; Available online 10 May 2019)

***Abstract* - Data mining is one of the major areas of research. Clustering is one of the main functionalities of datamining. High dimensionality is one of the main issues of clustering and Dimensionality reduction can be used as a solution to this problem. The present work makes a comparative study of dimensionality reduction techniques such as t-distributed stochastic neighbour embedding and probabilistic principal component analysis in the context of clustering. High dimensional data have been reduced to low dimensional data using dimensionality reduction techniques such as t-distributed stochastic neighbour embedding and probabilistic principal component analysis. Cluster analysis has been performed on the high dimensional data as well as the low dimensional data sets obtained through t-distributed stochastic neighbour embedding and Probabilistic principal component analysis with varying number of clusters. Mean squared error; time and space have been considered as parameters for comparison. The results obtained show that time taken to convert the high dimensional data into low dimensional data using probabilistic principal component analysis is higher than the time taken to convert the high dimensional data into low dimensional data using t-distributed stochastic neighbour embedding.The space required by the data set reduced through Probabilistic principal component analysis is less than the storage space required by the data set reduced through t-distributed stochastic neighbour embedding.**

***Keywords:* Clustering, Dimensionality Reduction, t-distributed Stochastic Neighbour Embedding, Probabilistic Principal Component Analysis**

I. INTRODUCTION

Data mining [1] refers to the process of extracting non-trivial, implicit, previously unknown and potentially useful information from vast amounts of data. The different functionalities of datamining are class/concept description, classification, clustering, outlier analysis and evolution analysis. Clustering is the process of grouping objects based on similarity. The similarity between objects can be measured using different distance measures such as Euclidean distance, Manhattan distance, Minkowski distance etc. The principle of clustering is to maximize the intra cluster similarity and minimize the inter cluster similarity. Clustering often suffers from curse of dimensionality. Dimensionality reduction can be used to deal with curse of dimensionality. The curse of dimensionality [2] was coined by Richard E bellman and it refers to various phenomena that arise when analysing and organizing data in high dimensional space that do not occur in low dimensional space. Dimensionality reduction [3] refers to the process of reducing the number of random variables under consideration by obtaining a set of principal variables.

Dimensionality reduction techniques are of two types. They are Linear Dimensionality reduction techniques and nonlinear dimensionality reduction techniques. A linear dimensionality reduction technique converts the high dimensional data into low dimensional data using linear transformation. Examples of linear dimensionality reduction techniques [4] include Principal Component analysis, Canonical correlation analysis, linear discriminant analysis, Maximum autocorrelation factors etc.

A nonlinear dimensionality reduction technique converts the high dimensional data into low dimensional data using nonlinear transformation. Examples of Nonlinear dimensionality reduction techniques [5] include Sammon’s mapping, auto encoders, Kernel PCA, Locally linear embedding, Laplacian Eigen maps etc. Different types of dimensionality reduction techniques have been developed. The present work makes a comparative study of two dimensionality reduction techniques such as t-distributed stochastic neighbour embedding and probabilistic principal component analysis in the context of clustering.

A. T-Distributed Stochastic Neighbour Embedding (t-SNE)

T-distributed Stochastic Neighbour Embedding [6] is an algorithm for dimensionality reduction that is well suited to visualizing high-dimensional data. The idea is to embed high-dimensional points in low dimensions in a way that respects similarities between points. Nearby points in the high-dimensional space correspond to nearby embedded low-dimensional points and distant points in high-dimensional space correspond to distant embedded low-dimensional points (Generally, it is impossible to match distances exactly between high-dimensional and low-dimensional spaces). t- SNE has been used for visualization in a wide range of applications including computer security research, music analysis, cancer research, bioinformatics and biomedical signal processing.

The algorithm takes the following general steps to embed the data in low dimensions.

1. Calculate the pair wise distances between the high-dimensional points.
2. Create a standard deviation $\sigma_i$for each high-dimensional point i so that the perplexity of each point is at a predetermined level.
3. Calculate the similarity matrix. This is the joint probability distribution of X defined by $P_{ij} = \frac{P_{j/i}+P_{i/j}}{2N}$ ------------------------------ (1.1), Where N is the number of rows of X.
4. Create an initial set of low-dimensional points.
5. Iteratively update the low-dimensional points to minimize the kullback-leibler divergence between a Gaussian distribution and a t-distribution in the low- dimensional space. This optimization procedure is the most time-consuming part of the algorithm.

B. Probabilistic Principal Component Analysis (PPCA)

Probabilistic Principal Component Analysis [7] is a method to estimate the principal axes when any data vector has one or more missing values. It is based on an isotropic error model. It seeks to relate a p-dimensional observational vector y to a corresponding k-dimensional vector of unobserved variable x, which is normal with mean zero and covariance l(k).The relationship is $y^T = W * x^T + \mu + \varepsilon$ -------------------(1.2) where y is the row vector of observed variable. x is the row vector of latent variables. μ is mean is the isotropic error term. ε is Gaussian with mean zero and covariance of v*l(k), where v is the residual variance. Here, k needs to be smaller than the rank for the residual variance to be greater than zero (v>0).

Standard Principal Component Analysis, where the residual variance is zero, is the limiting case of PPCA. The observed variables, y, are conditionally independent given the values of the latent variables, x. So the latent variables explain the correlations between the observation variables and the error explains the variability unique to a particular y_i

The p-by-k matrix W relates the latent and observation variables and the vector μ permits the model to have a nonzero mean. PPCA assumes that the values are missing at random through the dataset. This means that whether a data value is missing or not does not depend on the latent variable given the observed data values. Under this model, $y \sim N\left(\mu, W * W^T + v * I(k)\right)$---- (1.3). There is no closed-form analytical solution for W and v, so their estimates are determined by iterative maximization of the corresponding log likelihood using an expectation-maximization algorithm. This EM algorithm handles missing values by treating them as additional latent variables. At convergence, the columns of W span the subspace, but they are not orthonormal. PPCA obtains the orthonormal coefficients for the components by orthogonalization of W.

II. RELATED WORK

A comparative analysis of different dimensionality reduction techniques has been presented [8]. The effect of applying dimensionality reduction techniques such as Singular value decomposition, Random projection and Principal component analysis on the performance of trace clustering has been presented in [9]. A comparative study of dimensionality reduction techniques such as Random Mapping, Principal component analysis and Independent component analysis has been studied in the context of text retrieval [10]. A comparative study of dimensionality reduction techniques such as Singular value decomposition, Principal component analysis and Multidimensional Scaling with respect to time in the context of clustering has been studied in [11]. A comparative study of dimensionality reduction techniques such as Singular value decomposition, Principal component analysis and Multidimensional Scaling with respect to space in the context of clustering has been studied in [12].

A comparative study of dimensionality reduction techniques such as Singular value decomposition, Principal component analysis, Self organizing map and Fast ICA has been performed in the context of clustering health data [13]. A brief review and the mathematical insight of different dimensionality reduction techniques have been presented in [14]. An analysis of different weight initialization strategies and various dimensionality reduction measures with Self Organizing Maps has been performed in [15]. The performance of Self Organizing Map and hierarchical clustering methods has been tested with various levels of imperfections [16]. An evaluation of different linear and non-linear dimensionality reduction techniques has been performed and a novel evaluation metric known as Renyi entropy has been introduced [17]. A study on linear dimensionality reduction techniques and nonlinear dimensionality reduction techniques was presented in[18].

The performance of Principal component analysis and nonlinear dimension reduction techniques such as Kernel PCA, Locally Linear Embedding, Isomap, Diffusion maps, Laplacian Eigen maps and Maximum variance Unfolding has been assessed in terms of visualization of microarray data[19].A comparative study of dimensionality reduction techniques such as Principal component analysis, Isometric feature mapping, t-distributed stochastic neighbour embedding and diffusion maps has been performed on cytometry dataset[20]. A Comparative study of dimension reduction techniques in the context of visual exploration of text document archives has been performed in [21]. A comparative study of different dimensionality reduction techniques in combination with Fuzzy C-means clustering techniques has been performed [22].

III. EXPERIMENTAL RESULTS

A 7X7 matrix has been considered as a high dimensional data and it has been reduced to low dimensional data through techniques such as t-SNE and PPCA. Cluster analysis has been performed on the original data set and the two reduced data sets obtained through t-SNE and PPCA. The results have been noted and analysed. A 10 X 20 matrix has been considered as a high dimensional data and it has been reduced to low dimensional data through techniques such as t-SNE and PPCA. Cluster analysis has been performed on the original data set and the two data sets reduced through t-SNE and PPCA. The results have been noted and analysed.

An image of size 1219 X1600 pixels has been considered as a high dimensional data and it has been reduced to low dimensional data through techniques such as t-SNE and PPCA. Cluster analysis has been performed on the original data set and the two data sets obtained after applying t-SNE and PPCA on the original data set. The results have been noted and analysed. The time required to perform cluster analysis on the original data as well as on the two data sets reduced through t-SNE and PPCA has been computed. An analysis of time has been done in a table.

An image of size 177 ×284 pixels has been treated as a high dimensional data. Then this high dimensional data has been reduced to low dimensional data using two dimensionality reduction techniques such as t-SNE and PPCA. Clustering is performed on the original data and as well as the lower dimensional data obtained through t-SNE and PPCA. The storage space required by the original image as well as the images obtained after applying t-SNE and PPCA on the original image has been analysed.

An image of size 1600 ×1024 pixels has been treated as a high dimensional data. Then this high dimensional data has been reduced to lower dimensional data using two dimensionality reduction techniques such as t-SNE and PPCA. Clustering is performed on the original data and as well as the lower dimensional data obtained through t-SNE and PPCA. The storage space required by the original image as well as the images obtained after applying t-SNE and PPCA on the original image has been analysed. From the Table I, it is observed that the mean squared error obtained from the original data is not same as the mean squared error obtained from the data reduced through t-distributed stochastic neighbour embedding and the mean squared error obtained from the data reduced through probabilistic principal component analysis.

From the Table II, it is observed that the mean squared error obtained from the original data is not same as the mean squared error obtained from the data reduced through t-distributed stochastic neighbour embedding and the mean squared error obtained from the data reduced through probabilistic principal component analysis.

TABLE I MEAN SQUARED ERROR OBTAINED BY PERFORMING CLUSTERING ON THE ORIGINAL DATA SET (7X7 MATRIX) AND THE TWO DATA SETS OBTAINED BY APPLYING T-SNE AND PPCA RESPECTIVELY ON THE ORIGINAL DATA SET

Number of Clusters	**Mean Squared Error obtained From original data**	**Mean Squared Error obtained after applying t-SNE on the original data**	**Mean Squared Error obtained After applying PPCA on the original data**
K=2	21.0000 39.2000	1.0e+07*7.6957 1.0e+07*6.2615	19.7785 1.8015
K=3	16.5000 15.0000 0	1.0e+07*7.6957 1.0e+07*0 1.0e+07*2.7005	2.7342 4.3906 0.1374
K=4	0 10.0000 0 5.3333	1.0e+07*3.1262 1.0e+07*0 1.0e+07*0 1.0e+07*1.9996	1.8015 0 0 0
K=5	5.3333 0 0 0 0	1.0e+07*0 1.0e+07*3.1262 1.0e+07*0 1.0e+07*0 1.0e+07*0.8251	0 0.1374 0 0 0
K=6	0 0 0 1 0 0	1.0e+06*0 1.0e+06*0 1.0e+06*0 1.0e+06*0 1.0e+06*0 1.0e+06*8.2511	0 0 0 0 0.0024 0
K=7	0 0 0 0 0 0 0	0 0 0 0 0 0 0	0 0 0 0 0 0 0

TABLE II MEAN SQUARED ERROR OBTAINED BY PERFORMING CLUSTERING ON THE ORIGINAL DATA SET (10X20 MATRIX) AND THE TWO DATA SETS OBTAINED BY APPLYING T-SNE AND PPCA ON THE ORIGINAL DATA SET

Number Of Clusters	Mean Squared Error obtained from original data	Mean Squared Error obtained after applying t-SNE on the original data	Mean Squared Error obtained after applying PPCA on the original data
K=2	8000 8000	1.0e+06*1.0234 1.0e+05*1.2027	1.0e+05*0.4000 1.0e+05*1.4000
K=3	16000 40000 16000	1.0e+06*0.1085 1.0e+06*1.0234 1.0e+06*0.3859	1.0e+04*1.6000 1.0e+04*4.0000 1.0e+04*1.6000
K=4	16000 4000 16000 4000	1.0e+05*4.1296 1.0e+05*2.6099 1.0e+05*0 1.0e+05*2.8675	1.0e+04*0.4000 1.0e+04*4.0000 1.0e+04*1.6000 1.0e+04*0
K=5	4000 16000 0 4000 4000	1.0e+05*2.6099 1.0e+05*1.1077 1.0e+05*0.9628 1.0e+05*0 1.0e+05*1.2065	1.0e+04*1.6000 1.0e+04*0.4000 1.0e+04*0.4000 1.0e+04*0 1.0e+04*0.4000
K=6	160000 0 4000 0 4000 0	1.0e+05*0 1.0e+05*2.6099 1.0e+05*0 1.0e+05*0.9628 1.0e+05*1.2065 1.0e+05*0	1.0e+03*4.0000 1.0e+03*4.0000 1.0e+03*0 1.0e+03*4.0000 1.0e+03*0 1.0e+03*4.0000
K=7	4000 0 0 0 0 0 16000	1.0e+05*1.0768 1.0e+05*0.9628 1.0e+05*1.2249 1.0e+05*0 1.0e+05*0 1.0e+05*0 1.0e+05*0	1.0e+03*4.0000 1.0e+03*4.0000 1.0e+03*0 1.0e+03*0 1.0e+03*4.0000 1.0e+03*0 1.0e+03*0
K=8	4000 0 0 0 4000 0 0 0	1.0e+05*1.0845 1.0e+05*0 1.0e+05*1.2065 1.0e+05*0 1.0e+05*0 1.0e+05*0 1.0e+05*0 1.0e+05*0	1.0e+03*0 1.0e+03*4.0000 1.0e+03*0 1.0e+03*0 1.0e+03*4.0000 1.0e+03*0 1.0e+03*0 1.0e+03*0
K=9	0 0 0 4000 0 0 0 0 0	1.0e+04*0 1.0e+04*0 1.0e+04*0 1.0e+04*9.6282 1.0e+04*0 1.0e+04*0 1.0e+04*0 1.0e+04*0 1.0e+04*0	1.0e+03*0 1.0e+03*4.0000 1.0e+03*0 1.0e+03*0 1.0e+03*0 1.0e+03*0 1.0e+03*0 1.0e+03*0 1.0e+03*0
K=10	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 0 0 0

TABLE III MEAN SQUARED ERROR OBTAINED BY PERFORMING CLUSTERING ON THE ORIGINAL DATA SET (1219 X 1600 MATRIX) AND THE TWO DATA SETS OBTAINED BY APPLYING T-SNE AND PPCA ON THE ORIGINAL DATA SET

Number Of Clusters	Mean Squared Error obtained from original data	Mean Squared Error obtained after applying t-SNE on the original data	Mean Squared Error obtained after applying PPCA on the original data
K=2	1.0e+004*2.9334 1.0e+004*4.2951	1.0e+06*0.7730 1.0e+06*1.0210	1.0e+03*3.7919 1.0e+03*8.0526
K=3	1.0e+004*2.0894 1.0e+004*2.0815 1.0e+004*2.6050	1.0e+05*4.1385 1.0e+05*3.2673 1.0e+05*3.9360	1.0e+03*0.9835 1.0e+03*3.3992 1.0e+03*2.9929
K=4	1.0e+004*0.7671 1.0e+004*2.5085 1.0e+004*2.1082 1.0e+004*0.9453	1.0e+05*2.1799 1.0e+05*1.6252 1.0e+05*1.9568 1.0e+05*2.3386	1.0e+03*0.8911 1.0e+03*0.5157 1.0e+03*1.4011 1.0e+03*2.5075
K=5	1.0e+004*0.8280 1.0e+004*2.0766 1.0e+004*0.7025 1.0e+004*0.6823 1.0e+004*1.8124	1.0e+05*0.7912 1.0e+05*1.2523 1.0e+05*1.2417 1.0e+05*1.5020 1.0e+05*1.4170	1.0e+03*1.0596 1.0e+03*0.8664 1.0e+03*0.2892 1.0e+03*0.6033 1.0e+03*0.7790
K=6	1.0e+004*0.5546 1.0e+004*0.2453 1.0e+004*3.0308 1.0e+004*0.1309 1.0e+004*0.7409 1.0e+004*1.2877	1.0e+05*0.7150 1.0e+05*0.8519 1.0e+05*0.7409 1.0e+05*1.3581 1.0e+05*0.4405 1.0e+05*1.1294	1.0e+03*0.2892 1.0e+03*0.3838 1.0e+03*0.7431 1.0e+03*1.0957 1.0e+03*0.1420 1.0e+03*0.4114
K=7	1.0e+004*1.7249 1.0e+004*0.3662 1.0e+004*2.1066 1.0e+004*0.5507 1.0e+004*0.4683 1.0e+004*0.4930 1.0e+004*0.1134	1.0e+04*9.0203 1.0e+04*4.7865 1.0e+04*9.3964 1.0e+04*3.8777 1.0e+04*7.9737 1.0e+04*0.4317 1.0e+04*6.3488	245.0082 251.9932 240.2228 411.8773 407.4174 401.7320 488.2713

From the above Table III, it is observed that the mean squared error obtained from the original data is not same as the mean squared error obtained from the data reduced through t-distributed stochastic neighbour embedding and the mean squared error obtained from the data reduced through probabilistic principal component analysis.

TABLE IV TIME TAKEN FOR CONVERTING THE HIGH DIMENSIONAL DATA (IMAGE OF SIZE 1219×1600 PIXELS) TO LOW DIMENSIONAL DATA USING TWO DIMENSIONALITY REDUCTION TECHNIQUES SUCH T-SNE AND PPCA

Time taken to convert the high dimensional data in to low dimensional data using t-SNE in seconds	Time taken to convert the highdimensional data in to low dimensional data using PPCA in seconds
16.148554 seconds	1579.750614 seconds

From the above Table IV it is observed that the time taken to convert the high dimensional data into low dimensional data using probabilistic principal component analysis is higher than the time taken to convert the high dimensional data into low dimensional data using t-distributed stochastic neighbour embedding.

Table V Time Taken To Perform Cluster Analysis On The Original Data (Image Of Size 1219×1600 Pixels) As Well As The Low Dimensional Data Obtained Through T-Sne And Ppca

Number of Clusters (k)	Time taken to perform cluster analysis on the original data in seconds	Time Taken to perform cluster analysis on the lower dimensional data obtained through t-SNE in seconds	Time Taken to perform cluster analysis on the lower dimensional data obtained through PPCA in seconds
K=2	1.484000	0.005927	0.021327
K=3	4.079000	0.006344	0.010839
K=4	3.735000	0.011934	0.010455
K=5	4.437000	0.012545	0.013821
K=6	4.797000	0.012029	0.013945
K=7	5.953000	0.013739	0.012538

Table Vi Time Taken To Perform Data Reduction And As Well As Cluster Analysis On An Image Of Size 1219×1600 Pixels

Number of Clusters (k)	Total time taken to perform cluster analysis on the original data in seconds	Total time taken to perform data reduction through t-SNE and to perform cluster analysis on the lower dimensional data obtained through t-SNE in seconds	Total time taken to perform data reduction through PPCA and to perform cluster analysis on the lower dimensional data obtained through PPCA in seconds
K=2	1.484000	16.148554+ 0.005927 =16.1545	1579.750614+ 0.021327 =1.5798e+03
K=3	4.079000	16.148554+ 0.006344 =16.1549	1579.750614+ 0.010839 =1.5798e+03
K=4	3.735000	16.148554+ 0.011934 =16.1605	1579.750614+ 0.010455 =1.5798e+03
K=5	4.437000	16.148554+ 0.012545 =16.1611	1579.750614+ 0.013821 =1.5798e+03
K=6	4.797000	16.148554+ 0.012029 =16.1606	1579.750614+ 0.013945 =1.5798e+03
K=7	5.953000	16.148554+ 0.013739 =16.1623	1579.750614+ 0.012538 =1.5798e+03

From the above Table VI, it is clearly evident that the time taken to perform data reduction and clustering on the dataset reduced through probabilistic principal component analysis is higher than the time taken to perform data reduction and clustering on the dataset reduced through t-distributed stochastic neighbour embedding.

From the Table VII, it is observed that the mean squared error obtained from the original data is not same as the mean squared error obtained from the data reduced through t-distributed stochastic neighbour embedding and the mean squared error obtained from the data reduced through probabilistic principal component analysis.

Table Vii Mean Squared Error Obtained By Performing Clustering On The Original Data Set (177 × 284 Matrix) And The Two Data Sets Obtained By Applying T-Sne And Ppca On The Original Data Set

Number Of Clusters	Mean Squared Error obtained from original data	Mean Squared Error obtained After Applying t-SNE on the original data	Mean Squared Error obtained After Applying PPCA on the original data
K=2	1.0e+003*0.6235 1.0e+003*1.8843	1.0e+04*0.9410 1.0e+04*2.3408	55.0429 746.9416
K=3	603.5758 428.8857 898.0405	1.0e+03*7.7468 1.0e+03*6.5924 1.0e+03*2.2730	80.4787 128.1371 138.5557
K=4	328.2225 821.8787 234.2463 305.8172	1.0e+03*1.2936 1.0e+03*2.1620 1.0e+03*6.2774 1.0e+03*2.0978	84.4867 22.5390 41.3201 48.6506

Table Viii Space Required By Original Dataset (177 × 284 Matrix) And The Five Data Sets Obtained By Applying T-Sne And Ppca On Original Dataset

Space required by original matrix	Space required by reduced matrix obtained through t-SNE	Space required by reduced matrix obtained through PPCA
1206432bytes	4544 bytes	2272 Bytes

Table Ix Mean Squared Error Obtained By Performing Clustering On The Original Data Set (1600 × 1024 Matrix) And The Two Data Sets Obtained By Applying T-Sne And Ppca On The Original Data Set

Number Of Clusters	Mean Squared Error obtained from original data	Mean Squared Error obtained After Applying t-SNE on the original data	Mean Squared Error obtained After Applying PPCA on the original data
K=2	1.0e+004*1.2620 1.0e+004*2.3668	1.0e+05*3.6042 1.0e+05*4.6080	1.0e+04*1.2614 1.0e+04*0.7881
K=3	1.0e+004*1.2691 1.0e+004*1.0068 1.0e+004*0.8239	1.0e+05*0.6504 1.0e+05*1.6408 1.0e+05*2.5294	1.0e+03*4.0139 1.0e+03*3.4347 1.0e+03*1.1232
K=4	1.0e+003*6.5107 1.0e+003*6.8864 1.0e+003*6.6082 1.0e+003*7.0297	1.0e+05*0.6487 1.0e+05*0.6579 1.0e+05*1.1725 1.0e+05*0.6957	1.0e+03*3.1090 1.0e+03*0.2927 1.0e+03*1.1488 1.0e+03*1.1996
K=5	1.0e+003*5.0692 1.0e+003*6.6082 1.0e+003*2.3322 1.0e+003*5.8904 1.0e+003*5.9460	1.0e+04*4.0638 1.0e+04*6.8049 1.0e+04*6.0103 1.0e+04*1.5093 1.0e+04*5.6699	1.0e+03*0.8696 1.0e+03*0.6404 1.0e+03*1.0901 1.0e+03*0.9416 1.0e+03*0.2650
K=6	1.0e+003*4.1468 1.0e+003*2.3723 1.0e+003*1.0172 1.0e+003*4.3285 1.0e+003*9.3882 1.0e+003*4.9953	1.0e+04*1.5093 1.0e+04*2.0644 1.0e+04*2.7117 1.0e+04*2.6796 1.0e+04*3.8693 1.0e+04*3.5150	962.4843 126.3490 264.9569 561.3135 450.1083 925.9719
K=7	1.0e+003*0.1499 1.0e+003*0.6878 1.0e+003*4.5617 1.0e+003*1.5334 1.0e+003*5.3123 1.0e+003*6.6559 1.0e+003*5.4569	1.0e+04*3.0333 1.0e+04*1.5093 1.0e+04*1.8247 1.0e+04*2.4965 1.0e+04*0.8218 1.0e+04*0.5512 1.0e+04*2.7117	22.7646 854.5571 272.4079 300.2551 209.2375 787.1999 438.5545

From the above Table VIII, it is clearly evident that the space required by the data set reduced through Probabilistic principal component analysis is less than the storage space required by the data set reduced through t-distributed stochastic neighbour embedding.

From the above Table IX, it is observed that the mean squared error obtained from the original data is not same as the mean squared error obtained from the data reduced through t-distributed stochastic neighbour embedding and the mean squared error obtained from the data reduced through probabilistic principal component analysis.

TABLE X SPACE REQUIRED BY ORIGINAL DATASET (1600 × 1024 MATRIX) AND THE TWO DATA SETS OBTAINED BY APPLYING T-SNE AND PPCA ON ORIGINAL DATASET

Space required by original Image	Space required by reduced Image obtained through t-SNE	Space required by reduced Image obtained through PPCA
39321600 bytes	16384 bytes	8192 bytes

From the above Table X, it is clearly evident that the space required by the data set reduced through Probabilistic principal component analysis is less than the storage space required by the data set reduced through t-distributed stochastic neighbour embedding.

IV. CONCLUSION

Clustering often suffers from curse of dimensionality and dimensionality reduction can be used as a solution to this problem. A comparative study of two dimensionality reduction techniques such as t-distributed stochastic neighbour embedding and Probabilistic principal component analysis has been performed in the context of clustering. Different types of data have been considered and reduced to lower dimensional data through dimensionality reduction techniques such as t-distributed stochastic neighbour embedding and probabilistic principal component analysis. Cluster analysis has been done on the original data and the lower dimensional data obtained through dimensionality reduction techniques. The results obtained show that the mean squared error obtained from the original data is not same as the mean squared error obtained from the data reduced through t-distributed stochastic neighbour embedding and the mean squared error obtained from the data reduced through probabilistic principal component analysis.The time taken to perform data reduction and clustering on the dataset reduced through probabilistic principal component analysis is higher than the time taken to perform data reduction and clustering on the dataset reduced through t-distributed stochastic neighbour embedding.The storage space required by the data reduced through probabilistic principal component analysis is less than the storage space required by the data reduced through t-distributed stochastic neighbour embedding.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", *Morgan Kaufmann Publishers, Elsevier, Second Edition*

[2] The Wikipedia website [Online] Available at: https://en.wikipedia. org/wiki/Curse_of_dimensionality

[3] The Wikipedia website [Online] Available at: https://en.wikipedia. org/wiki/Dimensionality_reduction

[4] John P. Cunningham, Zoubin Ghahramani "Linear dimensionality Reduction: Survey, Insights and Generalizations", *Journal of Machine Learning Research*, PP.2859-2900, 2015

[5] The Wikipedia website [Online] Available at: https://en.wikipedia.org /wiki/Nonlinear-dimensionality-reduction.html

[6] The Math works website [Online] Available at: www.mathworks .com/help/stats/t-sne.html

[7] The Math works website [Online] Available at: www.mathworks. com/help/stats/ppca.html

[8] Omprakash Saini and Sumit Sharma " A Review on Dimensionality Reduction techniques in Data Mining", *Computer Engineering and Intelligent Systems*, Vol. 9, No.1, pp.7-14, 2018.

[9] Minseok Song, H.Yang, S.H.Siadat and Mykola Pechenizkiy "A comparative study of dimensionality reduction techniques to enhance trace clustering performances", *Expert Systems with applications,* Vol. 40, No. 9, pp. 3722-3734, July 2013.

[10] Vishwa vinay, Ingemar J.cox, Kenwood and Natasa Milic , "A comparison of Dimensionality Reduction Techniques for Text Retrieval", *Proceedings of the Fourth International Conference on Machine Learning and Applications, IEEE*, December 2005.

[11] T. Sudha and P. Nagendra Kumar, "Comparative study of dimensionality reduction techniques in the context of clustering", *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)*, Vol. 6, No.1, pp.19-28, February 2016.

[12] T. Sudha and P. Nagendra Kumar, "Achieving Privacy Preserving Clustering in Images using Multidimensional Scaling", *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR),* Vol. 6, No. 2, pp.9-18, May 2016.

[13] Rahmat widia sembiring, Jasni Mohamad Zain and Abdullah Embong, "Dimension Reduction of Health Data Clustering", *International Journal on New Computer Architectures and their applications*, Vol. 1, No. 3, pp.1041-1050, 2011.

[14] C.O.S. Sorzano, J. vargas and A. Pascual-Montano, "A survey of dimensionality reduction techniques", *arXiv.org, March 2014*.

[15] H. Haripriya, R. Devisree, Dinesh Pooja and Prema Nedungadi, "A Comparative analysis of Self organizing maps on weight initializations using different strategies." *Fifth International conference on Advances in Computing and Communications*, pp.434-438, March 2016.

[16] Paul Mangiameli, Shaw chen and David west, "A comparison of SOM Neural network and hierarchical clustering methods", *European Journal of Operational Research*", Vol. 93, No. 2, pp. 402-417, September 1996.

[17] Ashish Gupta and Richard Bowden, "Evaluating Dimensionality Reduction Techniques for Visual Category Recognition using Renyi entropy ", *19th European Signal Processing Conference*, pp. 913-917, September 2011.

[18] F.S.Tsai, "Comparative study of Dimensionality Reduction Techniques for Data Visualization", *Journal of Artificial Intelligence*, Vol. 3, No.3, pp.119-134, 2010.

[19] Christoph Bartenhagen, Hans-Ulrich Klein, Christian Ruckert, Xiaoyi Jiang and Martin Dugas, "Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data." *BMC Bioinformatics*, November 2010.

[20] Anna konstorum, Nathan Jekel, Emily vidal and Reinhard Laubenbacher, "Comparative analysis of linear and nonlinear dimension reduction techniques on Mass Cytometry Data", *bioRxiv.March 2018*.

[21] Shiping Huang, Matthew O. Ward and Elke A. Rundensteiner, "Exploration of Dimensionality Reduction for Text Visualization", *NSF grant IIS-0119276.*

[22] Kazim yildiz, Yilmaz Camurcu and Buket Dogan, "Comparison of Dimension Reduction Techniques on High Dimensional Datasets.", *The International Arab Journal of Information Technology*, Vol. 15, No. 2, March 2018.