A New Dimensionality Reduction Technique for K-Means Clustering Performance Improvement

[1]Poonam Rani, [2]Neha Madan

[1]Faculty, [2]Student
NSIT - Division of Computer Engineering, Delhi University, Delhi - 110078

***Abstract*: This paper proves the hypothesis that Auto Encoder Neural Network is better dimensionality Reduction technique as compared to the Principal Component Analysis and K-means itself improves performance of K-means technique. This technique is used in various fields like medical data image analysis, weather maps or satellite images .This paper uses Tensor flow and machine learning libraries like sklearn, programming language python, graph plot libraries like matplotlib and numpy for implementation.In terms of cluster purity, the Auto Encoder gives the best K-means performance improvement than Principal Component Analysis and K-means itself.**

***Keywords*: Auto Encoder Neural Network Model, PCA, Tensorflow, sklearn, matplotlib, numpy**

I. INTRODUCTION

Clustering is a task to combine objects into groups (called clusters) based on similarity measure. For instance, in document clustering the documents are assigned real number between 0 and 1.The number 0 indicates dissimilar document and number 1 indicates similar documents. The similarity measures refer as cosine similarity functions or Euclidean distance. Clustering technique is used in a wide area of applications like data analysis, image analysis and so on. Clustering can be of 1.Hierarchical Clustering: Clusters are arranged in hierarchy. 2. Partitional Clustering divides into distinct clusters such that each data object has one cluster. 1. K-medoids Methods: Data points are chosen to be medoids.

2. K-means Methods:
It chooses mean of data to be centroids clusters such that each data object has one cluster. 3. Density-Based Clustering are categorized in high density regions separated by regions of lower density.4. Grid-Based Clustering Data objects are divided into cells that are known as grids. Dimensionality Reduction means to covert high dimensional data to low dimensional data. This technique gives better performance in terms of K-means clustering. It consumes lesser space and is easier to compute. In machine learning, classification is done on huge number of factors. The larger factors makes difficult to work and visualize the training set data. It is used in image compression, data analysis, computer vision, data mining and several other domains. PCA stands for Principal Component Analysis. The high dimensional data is reduced to low dimensions sorted by retaining maximum data variance. The PCA dimensionality reduction process involves Normalization of image, Creating Covariance Matrix of image data, Calculating Eigen values and Eigenvectors for Covariance Matrix, Transforming the original image dataset into output with reduced dimensions. Auto Encoder neural network is automated functional model which undergoes training to regenerate it's input. The output layer generated is exactly the same as the input with reconstruction loss error function The output values depend upon the behaviour of input values and weighted values between input layer and hidden layer.As per the result,AutoEncoder is a better model than PCA and K-Means itself.

II. RELATED WORK

Clustering technique [1]Kanungo(2002) *et al*, proposed K-means Clustering algorithm for better efficiency and easy implementation. [2]Hans-Hermann, B (2008) *et al*, proposed the origin of k-means algorithm and modifications and enhancements in k-Means for better performance in cluster analysis. Dimensionality Reduction technique. [3] Lee, D (2001) *et al*, proposed the non-matrix factorization method to minimize the least mean square error. [4] Kaski (2011) *et al*, used dimensionality reduction technique to reveal hidden and useful information related to financial data. This also helped in visualization of data. The data is first reduced in smaller dimensions using Random Projection Dimensionality Reduction Technique. Then Multidimensional Scaling method is applied to visualize the financial data. Random projection DR method represents two matrices and initializes the matrix with random values and assign probabilities to it. Multidimensional Scaling produces k-dimensional features. [5] Talebibeyrami, Nayer (2013) used the feature selection methods for email classification. Feature selection techniques like document frequency, IG and X2 statistic methods proved to be best and MI to be weakest. Document Frequency calculates number of documents. Information Gain calculates the information gained by training set as a whole.X2 method distributes the information in weighting features. MI Technique calculates the mutual information. [6] Krzysztof siwek 1(2013) *et all*, uses dimensionality reduction technique in multidimensional medical data related to blood cells. It proves that TSNE is a better dimensionality reduction technique for it than any other linear and nonlinear methods. TSNE stands for t-distributed Stochastic Neighbor Embedding. [7] Mrs Juliet Rozario(2014) *et all* proved that PCA Dimensionality Reduction Technique applied high dimensional dataset improves quality and performance of high dimensional dataset. P CA technique stands for principal component analysis which extracts principal components with maximum variance. [8] Neha (2015) *et all*, introduced clustering technique in the field of data mining to analyse large data easily. A method of K-Means clustering technique was introduced to produce high quality clusters of data based on the similarity measure which discovers hidden data and makes the data analysis easier. [9] Sebring, Rahmat Widia Zane (2015) *et all*, proves that dimensionality reduction technique with clustering gives better performance. dbscan with svd has the lowest processing time and lowest attribute. dbscan groups points which are close to each other on basis of distance measure and marks a outlier of points in the low density region.SVD stands for single value decomposition which factorizes the matrix whether it is real or complex. [10] balodhi (2015) *et all*, proved that PCA Dimensionality Reduction technique applied to data in cloud environments makes representation of data correctly. In the paper[11] El Moudden, Ismail Ouzir(2016) *et all* proves that PCA Dimensionality Reduction technique improves the performance of K-nearest neighbour in recognition of daily human activities like playing, walking ,sitting etc. In the paper [12] UsmanAli (2017) *et all*, used PCA Dimensionality Reduction Technique and Factor Analysis to Leukaemia Bioinformatics data set which makes data analysis easier. PCA stands for principal component analysis which extracts principal components with maximum variance. Factor Analysis extracts meaning attributes which are required for large data analysis. [13] Ravet *et all* proves that dimensional reduction technique PCA applied on sounds and music collections makes its pattern analysis easier. Dimensionality Reduction and Clustering technique: [12] Sudha Ramkumar (2016) *et all* proposed that text document clustering using K-Means with feature selection Dimensionality Reduction technique improves performance. It is better than applying K-Means which may not cluster the document properly. Feature selection Dimensionality Reduction Method used here is Information Gain technique. It calculates the information gained by training set as whole.

III. PROPOSED WORK

Dimensionality reduction technique of neural network Autoencoder model on the MNIST image dataset has been applied. Since it is an automated non-linear functional model it gives the best performance as compared to other human designed dimensionality reduction techniques like linear PCA, Information Gain and so on.

MNSIT DATASET:

It is handwritten digits database containing training set of 60,000 examples, and a test set of 10,000 examples. It contains MNSIT data with digits 0-9 as sample with 28 * 28 dimensions i.e. 784 dimensions. Auto Encoder Model:The input neurons{x1, x2, x3.., xn} are taken for dimensionality reduction and output is reduced .The reduced output dimensions are used for better analysis of data. The image is reconstructed with error loss function and output is generated with neurons {y1, y2, y3...,yn}.The reconstruction error loss function gives values of hidden layer of neuron {z1,z2…,zn}.This output can be used for better performance in data analysis and clustering of data. The clustering is done by K-clustering technique. Auto Encoder is a neural network model.

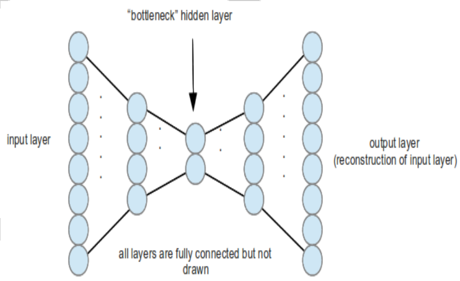


FIGURE:1.1

Training input image data with AutoEncoder Neural Network Model for Image Compression
Input: Image x with features (x1, x2…... xn).
1. Train the dataset X with Autoencoder Neural Network Model.
2. A reconstructed image Y with features (y1, y2… yn) similar to Input image X is generated.
3. Minimize the reconstruction error loss function (y-x) ^2 to generate Z.
4.Z is output in hidden layer of Auto Encoder with features(z1, z2…, zn).
5. Apply K-Means Clustering technique to output Z.

Output: The feature features (z1, z2…, zn) are grouped according to similarity or dissimilarity into k-clusters which yields better performance.

K-Means Clustering
1.The image data is taken as input and apply watershed algorithm from primary segmentation to final segmentation
2. The data observations are partitioned into K-clusters and they are randomly assigned to clusters
3.Calculate the shortest distance between data observation and centroid of cluster.
4. We generate K clusters by assigning each data observation point closest to cluster.
5. We calculate new centroid for each cluster.

ISSUES FACEDIN K-MEANS CLUSTERING:

1. Clustering quality and performance degrades with high dimensional data. Irrelevant features make the clustering difficult. High dimensional data causes overfitting. The dimensionality problem is reduced by the dimensionality reduction techniques.

2. Clustering depends on distance measure i.e. Euclidean Distance. In high dimensions space distance measure does not work. We have to apply dimensionality reduction technique and reduce dimensions to experiment with distance metric.

APPLICATIONS OF K-MEANS CLUSTERING

1. DATA MINING:

It is used to observe huge patterns of data. It is used in the field of machine learning, AI

Database technology

2. IMAGE ANALYSIS:

It means we are extracting quantitative information from images. It takes data as input and returns numerical output.

3. SEARCH ENGINE:

It is a software application system that search information on the net and return output as search results.

4. BIOINFORMATICS:

It is field for analysing the biological data with tools and techniques.

5. MACHINE LEARNING

It is a field in which statistical methods are developed to make ability of computer to learn.

6. DATA ANALYSIS

It is a field to analyse large number of datasets, extract and apply suitable models to it.It is useful to extract suitable information.

7. DOCUMENT BASED CLUSTERING

It is used to cluster the documents for information retrieval and extraction.

ADVANTAGE OF K-MEANS CLUSTERING:

It is fast and computationally efficient algorithm as it gives good clustering results and produces high quality clusters practical applications. It is used for high dimensional data. It works well when the clusters are hyper-spherical.

Some Techniques used to improve the performance of K-Means clustering:

1. It can select effective centroid to improve performance of k-means clustering.

2. It can use the Dimensionality Reduction technique as it is suitable for high-dimensional data.

DIMENSIONALITY REDUCTION:

Dimensionality Reduction means to convert high dimensional data to low dimensional data. This reduced feature output performs better performance in terms of K-means clustering. It consumes lesser space and is easier to compute. In machine learning, classification is done on huge no of factors. The larger factors makes difficult to visualize the training set data and work on it. It is used in image compression. data analysis, computer vision, data mining and several other domains.

PCA stands for Principal Component Analysis. The high dimensional data is reduced to low dimensions sorted by retaining maximum data variance. The PCA dimensionality reduction process follows 4 steps:

1.Normalization of image
2.Creating Covariance Matrix of image data
3.Calculating Eigenvalues and Eigenvectors for Covariance Matrix.
4.Transforming the original image dataset into output with reduced dimensions retaining maximum variance.

AutoEncoder neural network is an automated functional model which undergoes training to regenerate it's input.The output layer generated is exactly the same as the input with reconstruction loss error function i.e.The hidden layer compresses the input from high dimensions to lower dimensions. Let Z be the output of hidden layer. Z will give better performance in K-means Clustering. It makes the data visualization easier.

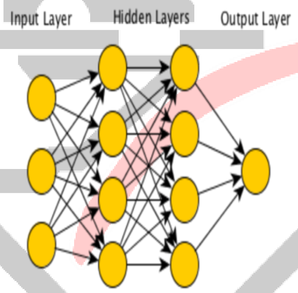


FIGURE:1.2

ADVANTAGE AND APPLICATIONS OF PROPOSED WORK:
The Auto encoder model neural network model is better model in performance than the PCA and other dimensionality reduction techniques introduced so far. It is an automated functional model whereas other are human designed models.It is used in data analysis and image analysis.

IV: IMPLEMENTATION

LIBRARY:

We use the tensorflow neural network library for training the data with Autoencoder model for dimensionality reduction. It gave better performance with K-means clustering.

TECHNOLOGY USED:

1.Machine learning libraries used like sklearn.

2.Programming language used is python.

3.Neural Network library used for implementation is tensorflow.

4.Graph plot libraries used are matplotlib

5.Other python libraries used is numpy.

V: EXPERIMENT AND RESULTS:

CLUSTER PURITY:

Purity of cluster refers to the largest class of objects assigned to the cluster.For e.g if 5s are in majority in a cluster then label name 5 is assigned to the cluster.5 is the cluster purity

EXPERIMENT RESULTS:

We load the Mnist dataset into representations(X)and their true labels(Y). We then randomly sample 10000 images and keep it separate, as test data. Using sklearns K-means clustering implementation, we measure the average performance of running k means directly on these 10000 images, where each image is 784 dimension representation vectors. We get a cluster purity of 0.5947 on 10 runs.

Next, we use sklearn's PCA implementation to reduce this 784 dimension vector into a n dimensional vector. We try various values for n, but find that with n=30, we get the best cluster purity on average. This value comes out to be 0.6002.

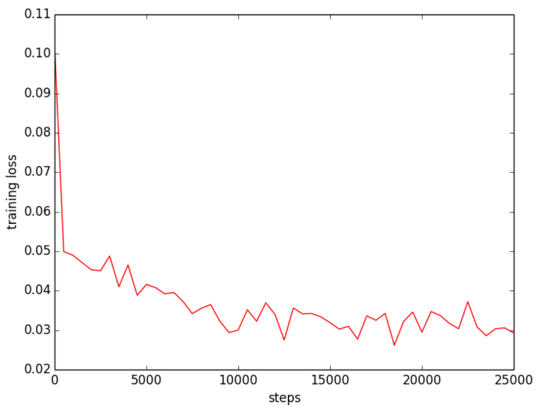
Finally, we apply our main approach of first training an auto encoder neural network model to reduce the dimensions. We use the 60000 images training dataset and pass each image through the input layer. Our AE model is structured as

F.C. (784,50) -> (50,50) -> (50,5) -> (5,50)-> (50,50)->(50,784)

We use all tanh activations except at the final layer, where we use ReLU.

The objective function is to minimise the reconstruion loss of the image, which is just the MSE. We use mini-batch gradient descent with a batch_size=50 and learning rate of 3e-4. The loss curve is shown below.

As it is evident we force the image through a 5 dimensional bottleneck layer, which we extract out for the unseen test data containing 10,000 images and then apply K-Means on top, getting a cluster purity of 0.62. The highest amongst the tested methods. Also, noteworthy here is that while we get a slight improvement for keeping 30 eigenvectors in PCA, we get a major bump with just 5 dimensions of the autoencoder.

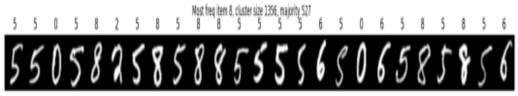


We show below the clusters generated by our three experiments.

K-Means Clusters with dimensionality reduction by Auto Encoder:

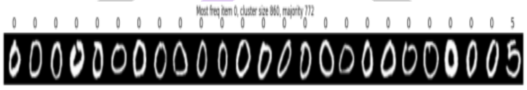
After training of data with Auto Encoder, the image is reconstructed with output Y .We minimize the reconstruction error (y-x)^2 to generate values of Z as hidden layer with (z1,z2,….zn). We apply k-means clustering for better performance. K-clusters are generated with n-observations.

The MNSIT dataset contains handwritten digits from 0-9.The data is trained with Auto Encoder model for dimensionality reduction. This improves performance with K-Means Clustering. The reduced output undergoes K-means Clustering. The digits are assigned to different cluster according to the cluster purity.

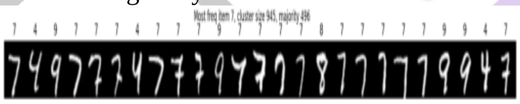


Cluster 1: As shown in the top of the cluster, it clusters images closest to "4". The number of elements in this are 1275, and majority element (which is 4) occurs 546 times. Cluster Homogeneity for the above cluster is defined as 527/1356=0.38.

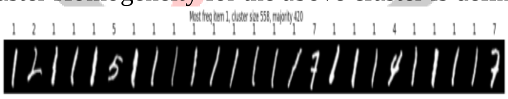
Cluster 2:10000 images are being put into 10 clusters. This image shows 25 randomly sampled images from a cluster.As shown in the top of the cluster, it clusters images closest to "4". The number of elements in this are 1468, and majority element(which is 4) occurs 546 times.Cluster homogeneity can be defined as majority / size of cluster. In this case it is 545/1275 = 0.31.



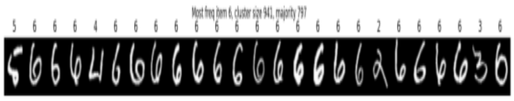
Cluster 3: As shown in the top of the cluster, it clusters images closest to "0". The number of elements in this are 860, and majority element (which is 0) occurs 772 times. Cluster Homogeneity for the above cluster is defined as 772/860=0.89.



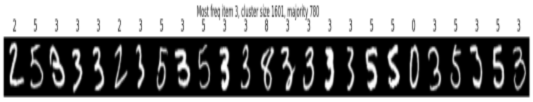
Cluster 4: As shown in the top of the cluster, it clusters images closest to "7". The number of elements in this are 945, and majority element (which is 7) occurs 496 times. Cluster Homogeneity for the above cluster is defined as 496/945=0.52



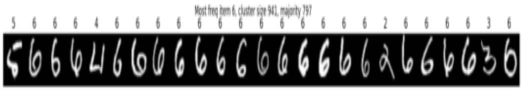
Cluster 5: As shown in the top of the cluster, it clusters images closest to "1". The number of elements in this are 558, and majority element (which is 1) occurs 420 times. Cluster Homogeneity for the above cluster is defined as 420/558=0.75



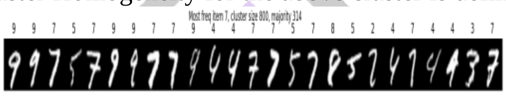
Cluster 6: As shown in the top of the cluster, it clusters images closest to "6". The number of elements in this are 940, and majority element (which is 6) occurs 797 times. Cluster Homogeneity for the above cluster is defined as 797/940=0.84



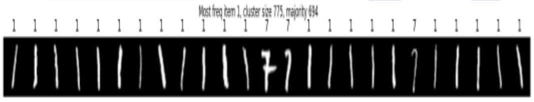
Cluster 7: As shown in the top of the cluster, it clusters images closest to "3". The number of elements in this are 1600, and majority element (which is 3) occurs 780 times. Cluster Homogeneity for the above cluster is defined as 780/1600=0.48



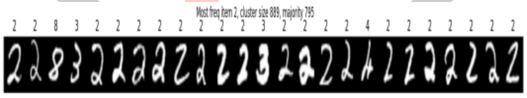
Cluster 8: As shown in the top of the cluster, it clusters images closest to "6". The number of elements in this are 941, and majority element (which is 6) occurs 797 times. Cluster Homogeneity for the above cluster is defined as 797/941=0.84



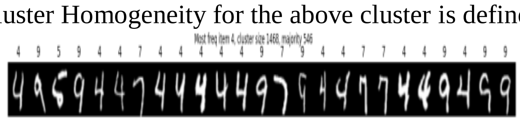
Cluster 9: As shown in the top of the cluster, it clusters images closest to "7". The number of elements in this are 800, and majority element (which is 7) occurs 314 times. Cluster Homogeneity for the above cluster is defined as 314/800=0.39



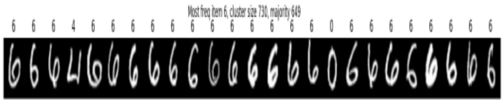
Cluster 10: As shown in the top of the cluster, it clusters images closest to "1". The number of elements in this are 775, and majority element (which is 1) occurs 694 times. Cluster Homogeneity for the above cluster is defined as 694/775=0.89.



Cluster 11: As shown in the top of the cluster, it clusters images closest to "2". The number of elements in this are 809, and majority element (which is 2) occurs 795 times. Cluster Homogeneity for the above cluster is defined as 795/809=0.98.



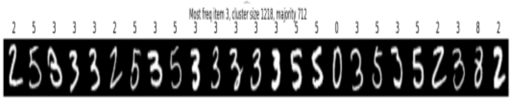
Cluster 12: As shown in the top of the cluster, it clusters images closest to "2". The number of elements in this are 809, and majority element (which is 2) occurs 795 times. Cluster Homogeneity for the above cluster is defined as 795/809=0.98.



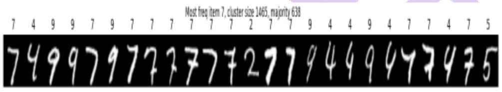
Cluster 13: As shown in the top of the cluster, it clusters images closest to "2". The number of elements in this are 809, and majority element (which is 2) occurs 795 times. Cluster Homogeneity for the above cluster is defined as 795/809=0.98.

K-Means Clusters generated by Dimensionality Reduction with PCA:

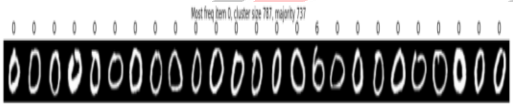
The MNSIT dataset contains handwritten digits from 0-9.PCA Dimensionality technique is applied to the dataset.The reduced output undergoes K-means Clustering. This improves performance with K-Means Clustering. The digits are assigned to different cluster according to the cluster purity.



Cluster 1: As shown in the top of the cluster, it clusters images closest to "3". The number of elements in this are 1218, and majority element (which is 3) occurs 732 times. Cluster Homogeneity for the above cluster is defined as 732/1218=0.600.

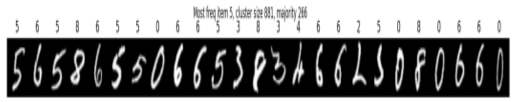


Cluster 2: As shown in the top of the cluster, it clusters images closest to "3". The number of elements in this are 1218, and majority element (which is 3) occurs 732 times. Cluster Homogeneity for the above cluster is defined as 732/1218=0.600.

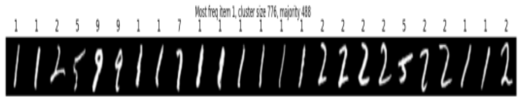


Cluster 3: As shown in the top of the cluster, it clusters images closest to "0". The number of elements in this are 787, and majority element (which is 0) occurs 737 times. Cluster Homogeneity for the above cluster is defined as 737/787=0.93

Cluster 4: As shown in the top of the cluster, it clusters images closest to "2". The number of elements in this are 700, and majority element (which is 2) occurs 772 times. Cluster Homogeneity for the above cluster is defined as 772/700=1.10



Cluster 5: As shown in the top of the cluster, it clusters images closest to "5". The number of elements in this are 880, and majority element (which is 5) occurs 226 times. Cluster Homogeneity for the above cluster is defined as 226/880=0.256.

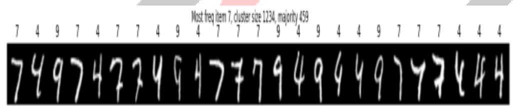


Cluster 6: As shown in the top of the cluster, it clusters images closest to "1". The number of elements in this are , and majority element (which is 1) occurs 488 times. Cluster Homogeneity for the above cluster is defined as 488/776=0.62

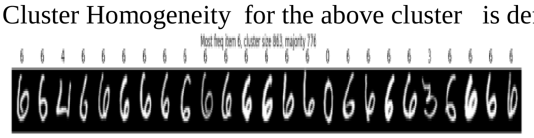
K-Means Clusters generated:

The MNSIT dataset contains handwritten digits from 0-9.The dataset undergoes K-means clustering.The digits are assigned to different cluster according to the cluster purity.

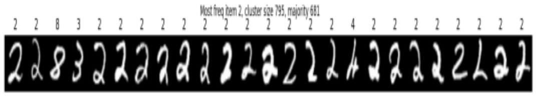
Cluster 1: As shown in the top of the cluster, it clusters images closest to "7". The number of elements in this are 373, and majority element (which is 7) occurs1089 times. Cluster Homogeneity for the above cluster is defined as 1089/373=2.9.



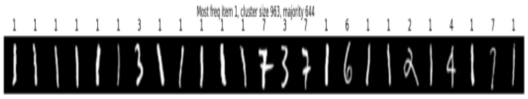
Cluster 2: As shown in the top of the cluster, it clusters images closest to "7". The number of elements in this are 373, and majority element (which is 7) occurs 1089 times. Cluster Homogeneity for the above cluster is defined as 1089/373=2.9.



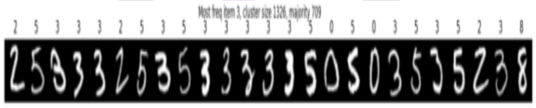
Cluster 3: As shown in the top of the cluster, it clusters images closest to "6". The number of elements in this are 863, and majority element (which is 6) occurs 776 times. Cluster Homogeneity for the above cluster is defined as 776/863=0.89



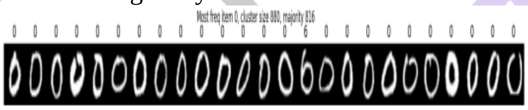
Cluster 4: As shown in the top of the cluster, it clusters images closest to "2". The number of elements in this are 795, and majority element (which is 2) occurs 681 times. Cluster Homogeneity for the above cluster is defined as 681/795=0.85



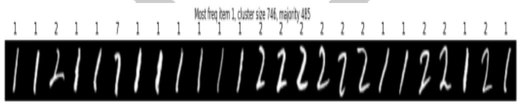
Cluster 5: As shown in the top of the cluster, it clusters images closest to "1". The number of elements in this are 963, and majority element (which is 1) occurs 644 times. Cluster Homogeneity for the above cluster is defined as 644/963=2.9.



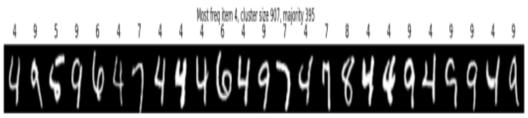
Cluster 6: As shown in the top of the cluster, it clusters images closest to "3". The number of elements in this are 1326, and majority element (which is 3) occurs 709 times. Cluster Homogeneity for the above cluster is defined as 709/1326=0.53.



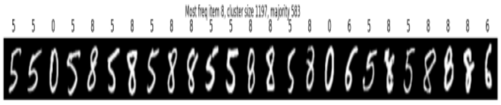
Cluster 7: As shown in the top of the cluster, it clusters images closest to "0". The number of elements in this are 880, and majority element (which is 0) occurs 816 times. Cluster Homogeneity for the above cluste is defined as 816/880=0.92



Cluster 8: As shown in the top of the cluster, it clusters images closest to "1". The number of elements in this are 746, and majority element (which is 1) occurs 485 times. Cluster Homogeneity for the above cluste is defined as 485/746=0.65



Cluster 9: As shown in the top of the cluster, it clusters images closest to "4". The number of elements in this are 907, and majority element (which is 4) occurs 395 times. Cluster Homogeneity for the above cluster is defined as 395/907=0.43



Cluster 10: As shown in the top of the cluster, it clusters images closest to "8". The number of elements in this are 583, and majority element (which is 8) occurs 1197 times. Cluster Homogeneity for the above cluster is defined as 583/1197=0.48

VI: CONCLUSION AND FUTURE SCOPE

First, we establish that using dimensionality reduction methods give a boost in k-means performance. The possible explanation of this lies in the inaccuracy of Euclidean similarity as we increase the dimensions. We further show here that nonlinear neural network based dimensionality reduction method is superior than using a linear method like PCA, not only in terms of cluster purity performance but also in terms of representation efficiency. It would be interesting to further experiment with the neural network structure and its activation functions to understand its representational capacity.

REFERENCES

[1].Lee, D., &Seung, H. (2001).Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, (1), 556–562.
[2].Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D.Silverman,R.,& Wu, A. Y. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 881–892.
[3].Hans-Hermann, B. (2008). Origins and extensions of the k-means algorithmin cluster analysis. *Journal Electroniqued'Histoire Des Probabilities et deLa Statistique*, *4*(2), 1–18
[4]. Sembiring, Zain, &Embong, 2011) Sembiring, R. W., Zain, J. M., &Embong, A. (2011). Dimension Reduction of Health Data Clustering, 1(4), 1018–1026.
[5].(Kaski, S., Peltonen, 2011)Kaski, S., Peltonen, J. (2011).Dimensionality reduction for data visualization.IEEE Signal Processing Magazine, 28(2),100-104.https://doi.org/10.1109/MSP.2010.940003
[6].(Paper, 2013) Paper, C. (2013). Analysis of Medical Data Using Dimensionality Reduction Techniques Analysis of Medical Data Using DimensionalityReductionTechniques,(April),279–281.
[6].(Talebibeyrami&Feiziderakhshi,2013) Talebibeyrami, N., &Feiziderakhshi, M. (2013). The Feature Selection and Dimensionality Reduction Methods for Email Classification, 3(Mi), 633–636.
[7].(Of, 2014) Of, I. J. (2014). Research in computer applications and robotics dimensionality reduction of high dimensional dataset using, 2(9), 133–139.
[8].(Balodhi, 2015) Balodhi, S. (2015). Dimensionality Reduction Problems Targeting Private Cloud Environments, 8491, 32–37.
[9].Neha, D., &Vidyavathi, B. (2015). A Survey on Applications of Data Mining using Clustering Techniques.*International Journal of Computer*, *126*(2),7–12.
[10].(Ravet, n.d.)Ravet, T. (n.d.). Nonlinear dimensionality reduction approaches applied to music and texual sounds St ´ ecile Picard-Limpens, Christian Frisson NUMEDIART Institute for New Media Art Technology University of Mons (Belgium) 31 Boulevard Dolez, B-7000 Mons, Belgium.
[11].(El Moudden, Ouzir, Benyacoub, & El Bernoussi, 2016) El Moudden, I., Ouzir, M., Benyacoub, B., & El Bernoussi, S. (2016). Mining human activity using dimensionality reduction and pattern recognition. Contemporary Engineering Sciences, 9(September),1031–1041.
[12]. (Sudha Ram Kumar &Poorna, 2016) SudhaRamkumar, A., &Poorna, B. (2016). Text Document Clustering Using Dimension Reduction Technique.International Journal of Applied Engineering Research ISSN, 11(7), 973–4562.
[13].(Usman Ali, Ahmed, Ferzund, Mehmood, &Rehman, 2017) Usman Ali, M., Ahmed, S., Ferzund, J., Mehmood, A., &Rehman, A. (2017). Using PCA and Factor Analysis for Dimensionality Reduction of Bio-informatics Data. IJACSA) International Journal of Advanced Computer Science and Applications, 8(5), 415–426.