Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Clutter Reduction in Multi-Dimensional Visualization by Using Dimension Reduction

Harpreet Kaur
Student of M.tech(CSE)
Swami Vivekanand Institute Of Engineering & Technology,
Banur, Punjab, India

Shelza
Assistant Professor (CSE)
Swami Vivekanand Institute Of Engineering & Technology,
Banur, Punjab, India

Abstract— The volume of Big data is increasing in gigabytes day by day which are hard to make sense and difficult to analyze. The challenges of big data are capturing, storing, searching, sharing, analysis and visualization of these datasets. Big data leads to clutter in their visualization. Clutter is a crowded or disordered collection of graphical entities in information visualization. It can blur the structure of data. In this paper, we present the concept of clutter based dimension reduction. Our purpose is to reduce clutter without reducing information content or disturb data in any way. Dimension reduction is a technique that can significantly reduce the dimensions of the datasets. Dimensionality reduction is useful in visualizing data, discovering a compact representation, decreasing computational processing time and addressing the curse of dimensionality of high-dimensional spaces.

Keywords— multidimensional visualization, dimension reduction, in homogeneity measure, time measure, visual clutter.

I. Introduction

Visualization is the process of transforming data into graphical representation. A good visualization clearly reveals structure of the data. The goal of visualization is to facilitate the user to gain a qualitative understanding of the information. An ideal visualization needs to maximize the visibility of patterns and structure and minimize the clutter present. Earlier visualization was done by constructing a visual image in mind but nowadays visualization is like a graphical representation that supports in decision making which extracts a lot of information in one vision without reading a lot of data files. On the other hand, clutter is a crowded or disordered collection of graphical entities in information visualization. Clutter is undesirable because it makes viewers difficult to understand the displayed content. When the dimensions or number of data items grow high, it is necessary for users to encounter clutter. Clutter reduces information gain from visualization. Clutter [1] is a state of confusion that degrades both the accuracy and ease of interpretation of information displays.

There are many techniques which are used to reduce the clutter and make the visualization better. However, many clutter reduction techniques may results in information loss and accuracy of data. Many clutter reduction techniques deal with data of high volume or high dimensionality, such as hierarchical clustering, sampling, and filtering. But they may result in some information loss. In order to complement these approaches, helping the user to reduce clutter in some traditional visualization techniques while retaining the information in the display, we propose a clutter reduction technique using dimension reduction.

1.1 Why it is important to reduce the clutter

- Increases information gain.
- Increases visibility of hidden datasets.
- Increases insights into datasets.
- Reduces mental overload and stress.
- Saves time and improve effectiveness.
- Improved data accuracy.
- Reorganizing makes it easier to access information and make things more accessible.
- Increases understanding and interpretation analysis of data.

Clutter reduction is a visualization-dependent task because visualization techniques vary largely from one to another. The basic goal of this paper is to present clutter reduction approaches for several visualization techniques. In order to automate the clutter reduction for dimension reduction, we first analyse the dataset and measure the dimensions of the original dataset. By using Dimension Reduction, the clutter in the dataset and dimensions of the dataset are reduced. After that the difference is calculated between before and after cluttered dataset. Our technique targets on small to middle-size dataset in terms of dimensionality. Although we only chose four visualization techniques to experiment with, there are many more traditional visualization techniques can benefit from this concept.

II. PREVIOUS WORK

To overcome the clutter problem, many approaches have been proposed. Multi-resolution approaches are used to group the data into hierarchical clusters and display them at a desired level of detail. These approaches do not retain all the information in the data, since many details will be filtered out at low resolutions.

Shelza[3] used clustering technique to reduce the clutter in CAD images and made visualization Framework that will incorporate clustering based on features of CAD images. The results shows that Visualization has been identified as a critical technique for exploring data sets and for this best abstraction technique is chosen based upon data abstraction quality from the number of available data abstraction techniques. They used data abstraction quality measure 'Histogram Difference Measure (HDM)' to find out how well the abstracted dataset represents the original dataset. Among three clustering algorithms, the result shows that proposed clustering algorithm gives the best results and then visualization is done for clustered dataset created by proposed clustering approach.

Wei Peng[2] proposed a dimension reordering technique for clutter reduction and uses the heuristic algorithms. By using heuristics algorithms, they did work on dimension reordering with much higher dimensions with relatively good results.

III. METHODOLOGY

The methodology to reduce the clutter incorporates the following steps:

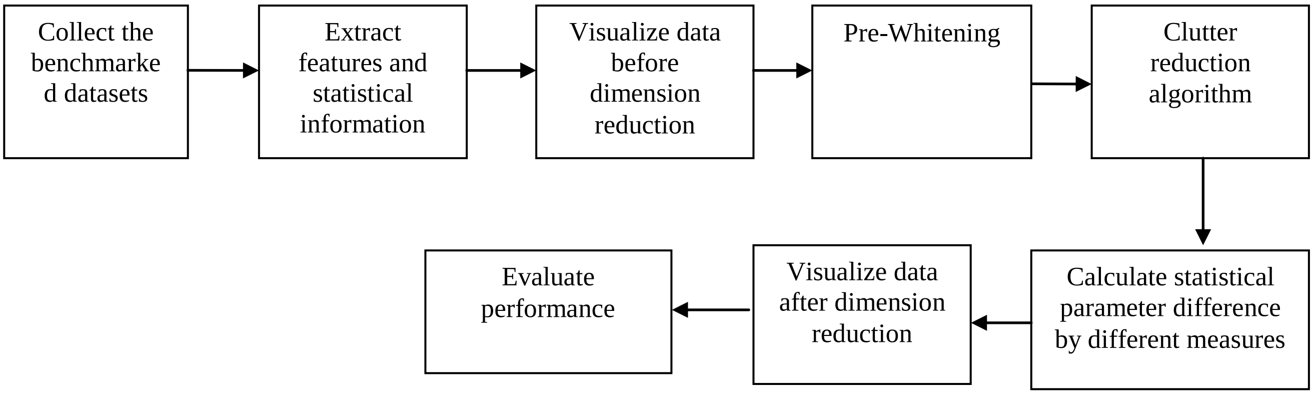
Collect the benchmarke d datasets
Extract features and statistical information
Visualize data before dimension reduction
Pre-Whitening
Clutter reduction algorithm
Calculate statistical parameter difference by different measures
Visualize data after dimension reduction
Evaluate performance

FIG 1.PROCEDURE FOR CLUTTER REDUCTION IN MULTI-DIMENSIONAL VISUALIZATION

A. COLLECT THE DATASETS

The first step for the clutter reduction is to collect the dataset. The datasets 3D Clusters, Helix, Twin Peaks are used.

B. EXTRACT FEATURES AND STATISTICAL INFORMATION

The features and statistical information are extracted according to import features and dataset is created. The features and information are extracted by using mean, standard, Variance, Co-variance methods. The mean is used to refer to central value of a discrete set of numbers. In statistics, standard deviation shows how much variation or dispersion exists from the mean, or expected value. Variance is a measure of how far a set of numbers is spread out. Co-variance is a measure of how much two random variables change together.

C. *PRE-WHITENING*

After extracting the features and information the pre-whitening method is used. Pre-whitening [8] concentrates the main variance in the data in a relatively small number of dimensions. Thereby, it separates noise from the data. Therefore, pre-whitening is recommended before performing any dimensionality reduction. Pre-whitening is helpful for trends in data. It is useful for find the co-relation in data. Correlated data means there is some relationship exists in data. So pre-whitening is useful to keep the data which has some relationship. It will remove the data which has no relationship.

D. VISUALIZE DATA BEFORE CLUTTER REDUCTION

After pre-whitening the data is visualized. The data is visualized before clutter reduction and dimensions are calculated before clutter reduction.

E. IMPLEMENTING CLUTTER REDUCTION ALGORITHM

After visualizing the data, cutter reduction based dimension reduction algorithm [7] is implemented. The dimension reduction technique is used for high-dimensional datasets. When analysing large data of multiple dimensions, it may be necessary to perform dimensionality reduction techniques to transform the data into a smaller, more manageable set. In dimension reduction, the dimensions of the datasets can significantly reduce. Dimensionality reduction is useful in visualizing data, discovering a compact representation, decreasing computational processing time and addressing the curse of dimensionality of high-dimensional spaces. Reducing the number of dimensions can separate the important features or variables from the less important ones, thus providing additional insight into the data. The GPLVM, CFA, LPP techniques are used to reduce the dimensions of the dataset.

The Gaussian process latent variable model [5] is a flexible approach to probabilistic modelling in high dimensional spaces. A major advantage of the approach is its ability to effectively model probabilistically data of high dimensionality.

GPLVM is a probabilistic approach. This approach can be used to handle missing data. CFA [6] stands for Coordinated Factor Analysis. In Statistics Data CFA is used to reduce the co-ordinates to the lower dimensional space. Locality preserving projection (LPP) is a linear projective map. This technique is an alternative to PCA a classical technique that is used to projects the data along the directions of maximal variance.

F. CALCULATE STASTICAL PARAMETER DIFFERNCE

In this step, the difference is calculated between before dimension reduction values and after dimension reduction values. The difference should be minimum between these values. The difference is calculated by using time, testing for equal distribution and inhomogeneity measures. The results will be better pronounced if the difference between before dimension reduction value and after dimension reduction value is minimum.

G. VISUALIZE DATA AFTER CLUTTER REDUCTION

In this step, the data is visualized after clutter reduction by using dimension reduction. The number of dimension has been reduced after dimension reduction and clutter is reduced to the much extent. The visualization of data is much better than before clutter reduction. Visualization helps to graphically depict the underlying knowledge in the data.

H. EVALUATE PERFORMANCE

In this step, the performance is calculated on the basis of statistical parameter difference by using different measures. After the clutter reduction it is clear that CFA technique is producing the good result in every measure.

IV. RESULTS & DISCUSSIONS

In our experiment we have seen that the statistical information before and after dimension reduction for removal of clutter thus, normally minimum and maximum values remain the same and there is no effect on it. If there is minimum difference between the mean, standard, variance values this shows that the technique producing good results greater the difference the technique will be fairly good. Due to application of dimension reduction, the visualization before clutter reduction is now more information gain and structure is more clearly revealed. We have used two different visualization techniques to visualize the data. The visualization techniques plot matrix and parallel coordinates are used to visualize the data. The dimension reduction linear techniques GPLVM, CFA, LPP are used to reduce the dimension or clutter from the data. After that the measures are applied on these techniques. The access time measure, test significance measure, Homogeneity measure, are applied to prove that which technique of dimension reduction producing the good results.

A. *TIME MEASURE*

Time measure is a measure which calculates the time between before and after execution with its visualization. For measure the time tic, toc command is used. Tic starts a stopwatch timer to measure performance. The function records the internal time at execution of the Tic command and display the elapsed time with the Toc function. The time before dimension reduction was high. But after removing clutter by using GPLVM the access time get reduced.

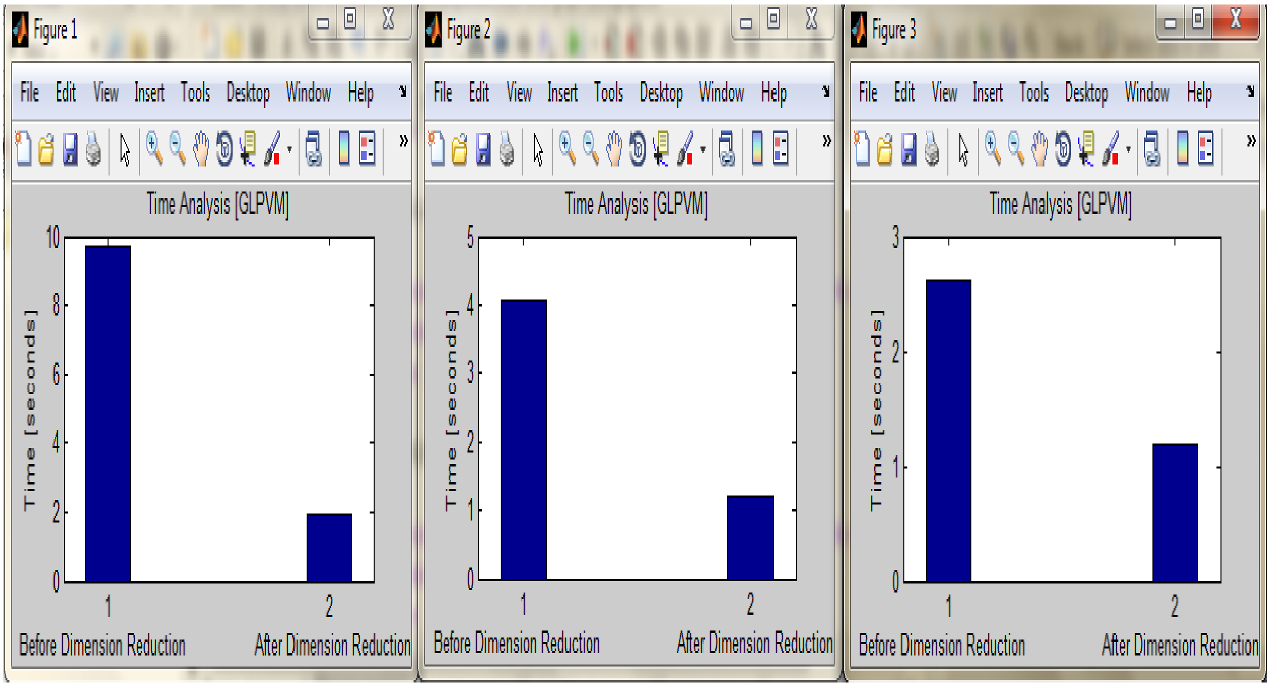

Fig. 2 shows the time measure for GPLVM technique on different datasets

It is clear from the figure 2 that the time is analysed higher than before dimension reduction. The time before dimension reduction is very high. By applying GPLVM method the time get low. The figure shows the time measure for different datasets. The results showing the different access time for different datasets 3D cluster, helix, twin peaks datasets.

The figure 3 shows the results for CFA (Coordinated Factor Analysis) technique of dimension reduction. The CFA reduces the co-ordinates to the lower space. The CFA technique in our case has the least access time.

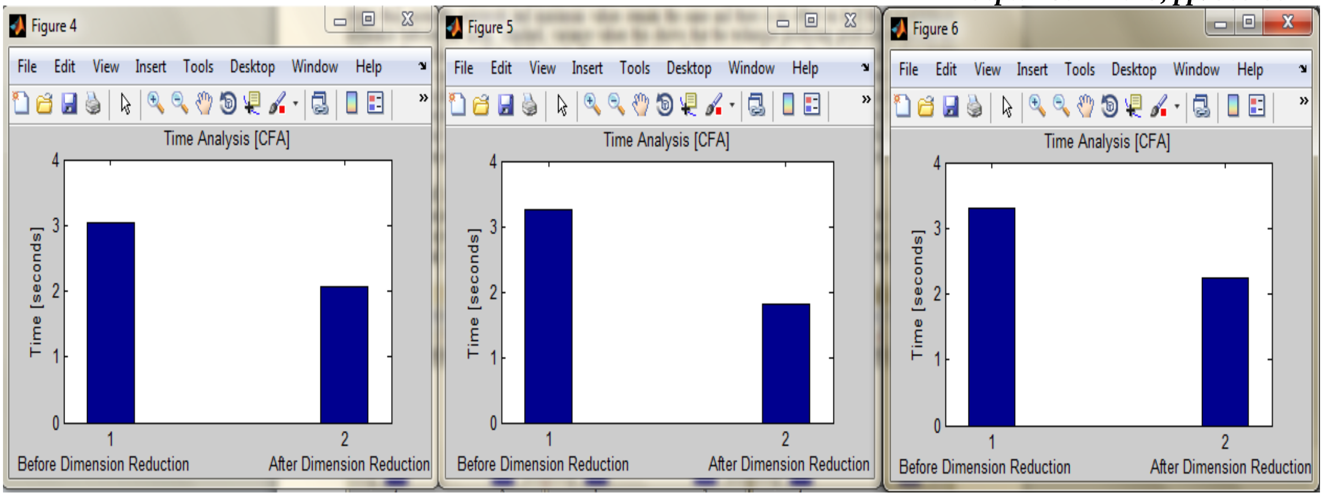

Fig. 3 shows the time measure for CFA technique on different datasets

It is clear from the figure 3 that the access time is low after dimension reduction. In figure 4 the results shown for LPP technique.in this technique the data is projected along with dimensions.

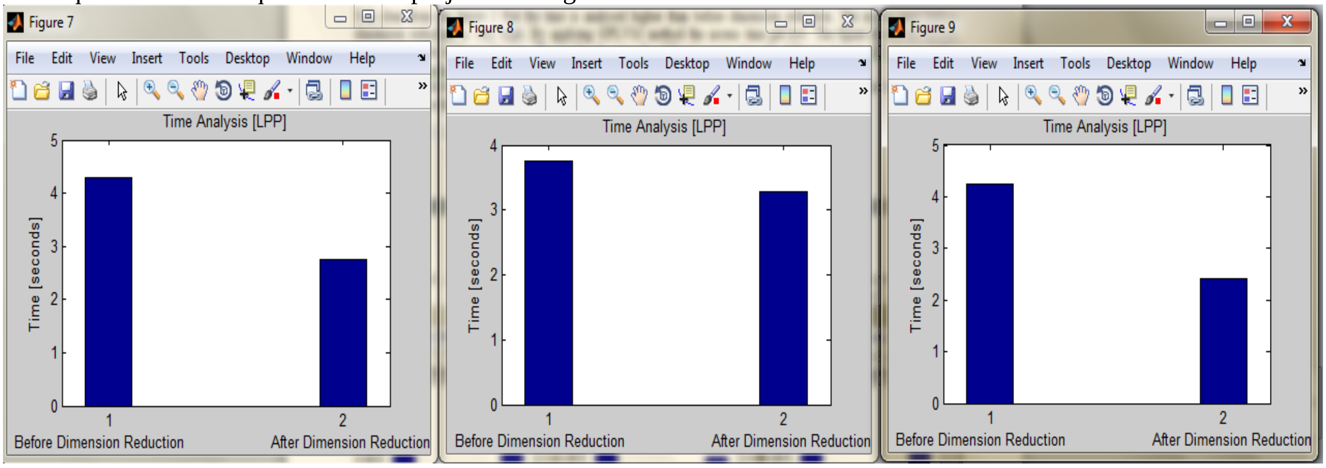

Fig. 4 shows the time measure for CFA technique on different datasets

It is clear from the figure 4 that the times after clutter reduction is get reduced by using LPP technique. The table represents the time for different techniques.

TABLE I

TIME MEASURE FOR DIFFERENT DATASETS BY DIFFERENT TECHNIQUES

S no.	Datasets used	Access time before dimension reduction in seconds	Access time before dimension reduction in seconds	Technique used
1.	3D Cluster	9.735035 sec.	1.950313 sec.	GPLVM
	Helix	4.054816 sec.	1.206065 sec.	GPLVM
	Twin Peaks	2.619645 sec.	1.192811 sec.	GPLVM
2.	3D Cluster	3.043035 sec.	2.062217 sec.	CFA
	Helix	3.265204 sec.	1.814966 sec.	CFA
	Twin Peaks	3.314031 sec.	2.235179 sec.	CFA
3.	3D Cluster	4.299922 sec.	2.757660 sec.	LPP
	Helix	3.759456 sec.	3.285005 sec.	LPP
	Twin Peaks	4.256750 sec.	2.407024 sec.	LPP

It is clear from the measure of access time that the best produced result is by CFA technique because it has less access time.

B. *TESTING FOR EQUAL DISTRIBUTION*

This test basically signifies whether the distributions are equal or not in each distribution. This is based on the null hypothesis that two random variables have X and Y will have same probability distribution μ=v where $X=X_1$-----Xn $Y=Y_1$------Yn. For conducting this test, we conduct arithmetic averages of distances between X and Y samples. To prove:

$$E_{n,m}(X,Y)= 2A-B-C \qquad ….(i)$$

The formula to calculate A,B,C is as follows.

$$A=1/nm\sum |Xi – Yj| \qquad …(ii)$$

ij

$$B=1/n^2\sum|Xi - Xj|$$

ij

$$C=1/m^2\sum |Yi - Yj| \qquad \ldots(iv)$$

ij

therefor, where equation (i) is greater than zero the test hypothesis T will equal to:

m*n/m+n*eqn(i)

The test is implemented at different techniques of dimension reduction. The results are shown below for the GPLVM, CFA, LPP techniques.

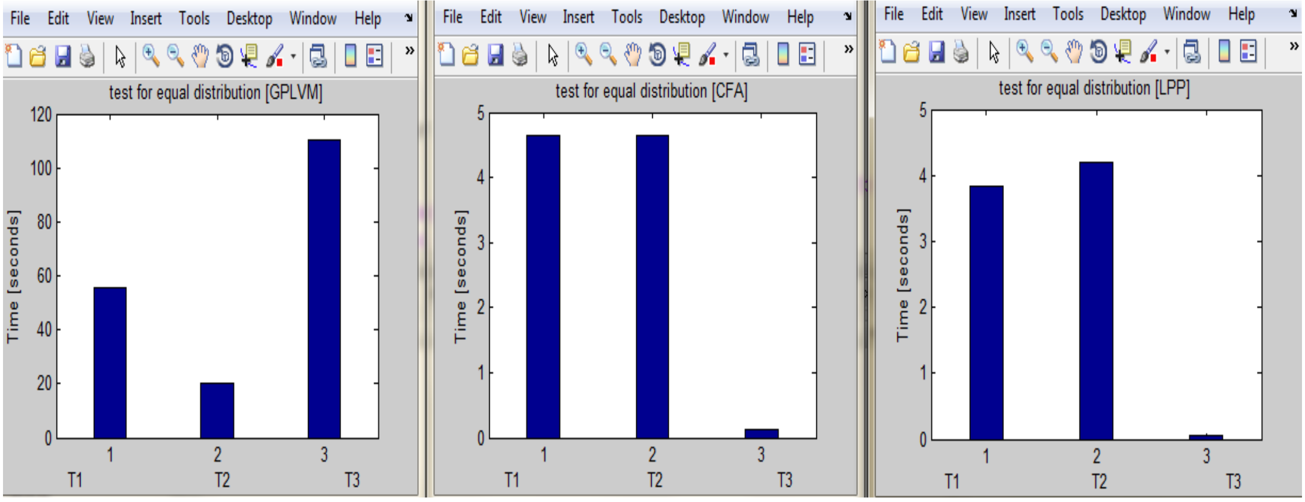

Fig. 5 shows the equal distribution for GPLVM, CFA, LPP techniques

The test has two case:

1) Larger the dimensions less is the value of test significance.
2) Lower the dimensions high is the value of test significance.

In this equal distribution test the CFA technique has fairly good results than other techniques. Higher is the value of test is in GPLVM. From the test it is clear that larger the dimension the value of test will be low. So the value of CFA is less in this test. This test shows that lower the value better it is for the technique.

TABLE IIII

TESTING FOR EQUAL DISTRIBUTION MEASURE FOR DIFFERENT DATASETS BY DIFFERENT TECHNIQUES

S no.	Datasets used	Testing for equal distribution	Technique used
1.	3D Cluster	55.748445485544174	GPLVM
	Helix	19.977251929227460	
	Twin Peaks	110.1515633876394	
2.	3D Cluster	4.640369276191493	CFA
	Helix	4.643927045993644	
	Twin Peaks	0.118749945289754	
3.	3D Cluster	3.829196971452624	LPP
	Helix	4.189795507755489	
	Twin Peaks	0.066791863923225	

C. *E-COEFFICIENT INHOMOGENEITY MEASURE*

In statistics, homogeneity and its opposite, heterogeneity, arise in describing the properties of a dataset, or several datasets. They relate to the validity of the often convenient assumption that the statistical properties of any one part of an overall dataset are the same as any other part. In meta-analysis, which combines the data from several studies, homogeneity measures the differences or similarities between the several studies , which in our case is visualization [plot matrix ,parallel co-ordinates] of dataset after the application of dimension reduction, It also helps in studying the several degrees of complexity of the dataset [3D Cluster, Helix, Twin peaks] values. Homogeneity helps to show the uniformity of data. The homogeneity depends on the test significance. The homogeneity should be lie between 0 and 1. The homogeneity measure for the dimension reduction techniques is as follows:

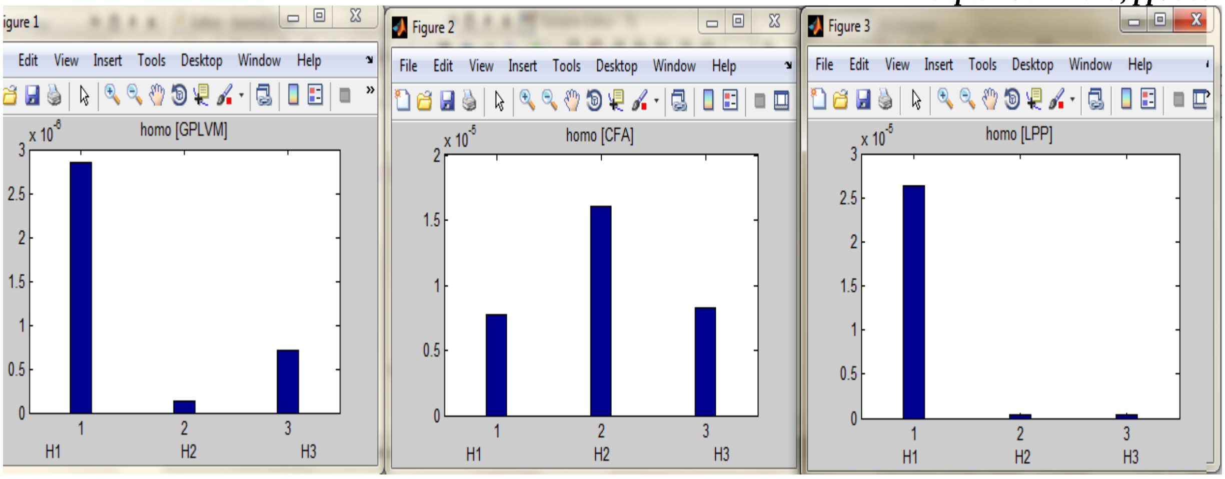

Fig. 6 shows the homogeneity measure for GPLVM, CFA, LPP techniques

TABLE IIIII

HOMOGENEITY MEASURE FOR DIFFERENT DATASETS BY DIFFERENT TECHNIQUES

S no.	Datasets used	Inhomogeneity measure	Technique used
1.	3D Cluster	0.000002848	GPLVM
	Helix	0.000000134	
	Twin Peaks	0.000000711	
2.	3D Cluster	0.000007705	CFA
	Helix	0.000016028	
	Twin Peaks	0.000008300	
3.	3D Cluster	0.000026286	LPP
	Helix	0.000000375	
	Twin Peaks	0.000000428	

From the table it is clear that the H value of each technique is lying between 0 and 1. CFA technique has best result in this homogeneity measure.

V. Conclusion

In this paper, we have used the concept of visual clutter reduction by using dimension reduction in Multi-Dimensional Visualization. By using dimension reduction, our purpose is to reduce clutter by reducing the dimensions of the original dataset. The different measures are applied on different techniques. In our study GPLVM, CFA, LPP techniques are used for clutter reduction. From the above measures it is clear that CFA technique is producing good result in every case. By using clutter based dimension reduction the visualization of the datasets is improved. The visualization has now more information gain and reveals the structure more clearly.

VI. Future Scope

Future work will include the combination of dimension reduction approach with other approaches. In our paper three techniques GPLVM, CFA, LPP are used. In future more techniques can be used to reduce the clutter and make the visualization more better and more other measures than those used in this paper can be used to measure the clutter in data.

References

[1] Clutter Measurement and Reduction for Enhanced Information Visualization by Natasha Lloyd in Computer Science December 2005.

[2] Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering by Wei Peng, Matthew O. Ward and Elke A. Rundensteiner.

[3] A Novel System for Abstraction and Visualization of CAD Images by shelza and Balwinder singh

[4] Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets by Jing Yang, Matthew O. Ward and Elke A. Rundensteiner.

[5] pre-whitening of data by co-variance weighted pre-processing by Harald Martens1*, Martin Høy2, Barry M. Wise3, Rasmus Bro1 and Per B. Brockhoff4.

[6] learning an internet co-ordinate system by dilip antony joseph

[7] Dimensionality Reduction:AComparative Review by L.J.P. van der Maaten _ , E.O. Postma, H.J. van den Herik.

[8] To prewhiten or not to prewhiten in trend analysis? By M Bayazit & B Önöz.

[9] Data Dimensionality Estimation Methods:A survey by Francesco Camastra

[10] Measuring Data Abstraction Quality in Multiresolution Visualization by Qingguang Cui1,†, Matthew O. Ward1, Elke A. Rundensteiner1 and Jing Yang2