Clutter Reduction in Multi-Dimensional Visualization by Using Dimension Reduction

Harpreet Kaur[1], Shelza[2]

[1]Swami Vivekanand Institute of Engineering and Technology, Banur, Punjab, India

[2]Swami Vivekanand Institute of Engineering and 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: multi-dimensional visualization, dimension reduction, visual clutter, visual structure

1. 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 analyze 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.

2. 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. 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.

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.

Yang et al. [4] proposed a visual hierarchical dimension reduction technique that creates meaningful lower

dimensional spaces with representative dimensions from original data instead of from generated new dimensions. These techniques generate a lower dimensional subspace to reduce clutter but some information in the original data space is also lost.

3. Methodology

The methodology to reduce the clutter incorporates the following steps:

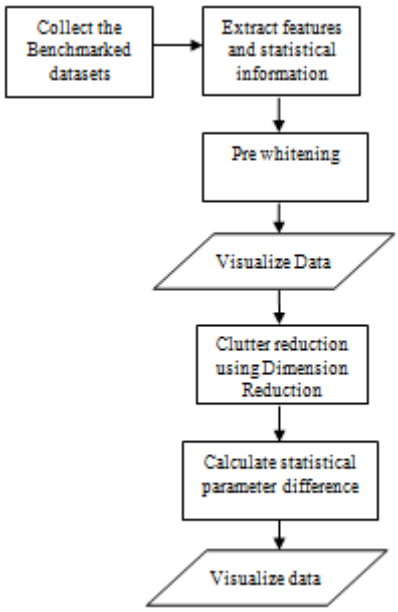


Figure 1: Procedure for clutter reduction in multi-dimensional visualization

3.1 Collect the Datasets

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

3.2 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 a 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.

3.3 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.

3.4 Visualize Data

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

3.5 Implementing Clutter Reduction Using Dimension Reduction

After visualizing the data before clutter reduction, the cutter reduction based dimension reduction algorithm [7] is implemented. The dimension reduction technique is used for high-dimensional datasets. When analyzing 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 and CFA techniques are used to reduce the dimensions of the dataset.

The Gaussian process latent variable model [5] is a flexible approach to probabilistic modeling 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

3.6 Calculate Statistical Parameter Difference

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 results will be better pronounced if the difference between before dimension reduction value and after dimension reduction value is minimum.

3.7 Visualize data

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.

Minimum and maximum values remain the same and there is no effect on it. However, there is large in variance consequently on the standard, which is not good side effect of the dimension reduction. This is widest from the shift in values of mean. Due to application of dimension reduction,

4. Results

It can be seen that the statistical information before and after dimension reduction for removal of clutter thus, normally

the visualization before clutter reduction is now more information gain and structure is more clearly revealed. Following table shows the results for dimension reduction by using GPLVM technique:

Table1: Shows the results for 3D Clusters Dataset by using GPLVM

S. No	Dataset	Min	Max	Mean	Std	Var	Co-var	No. of Dimensions
1.	3D_ Clusters (Before)	0	0	0.5095	0.0582	0	0.0442,0.0121,0.0118, 0.0121,0.0472,-0.0119, 0.0118,-0.0119, 0.0989	500,3
2.	3D_ Clusters (After)	-7.1728	-7.1728	-1.1156e-15	0.5651	9.9653	10.1922, 0.0000, 0.0000, 5.7278	500,2

Table2: Shows the results for Helix Dataset By using GPLVM Technique

S. No	Dataset	Min	Max	Mean	Std	Var	Co-var	No. of Dimensions
1.	Helix (Before)	0	0	0.4984	0.0748	0	0.0602,0.0004,-0.0000, 0.0004,0.0602, 0.0002, -0.0000,0.0002,0.0134	500,3
2.	Helix (After)	-3.0553	-3.0553	-3.1086e-18	0.0070	4.3940e-04	2.2628, 0.0000, 0.0000 ,2.2332	500,2

Table 3: Shows the results for Twin Peaks Dataset by using GPLVM technique

S. No	Dataset	Min	Max	Mean	Std	Var	Co-var	No. of Dimensions
1.	Twin peaks (Before)	0	0	0.4997	0.1435	0	0.0007,-0.0000,0.0002, -0.0000,0.0007,0.0003, -0.0002,0.0003,0.0759	500,3
2.	Twin peaks (After)	-10.6956	-10.6956	2.0206e-17	3.7523	589.5556	34.6771, 0.0000, 0.0000, 0.3389	500,2

The second technique that is used for dimension reduction is CFA. CFA stands for Coordinated Factor Analysis. In Statistics Data CFA is used to reduce the co-ordinates to the lower dimensional space.

Table 4: Shows the results for 3D Cluster dataset by using CFA technique

S. No	Dataset	Min	Max	Mean	Std	Var	Co-var	No. of Dimensions
1.	3D_Clusters (Before)	0	0	0.5095	0.0582	0	0.0442,0.0121,0.0118, 0.0121,0.0472, -0.0119, 0.0118,-0.0119, 0.0989	500,3
2.	3D_Clusters (After)	0.0128	0.0128	0.0539	4.2431e-16	7.3839e-61	1.0e-29 *, 0.1723 , -0.0933 -0.0933 , 0.0508	500,2

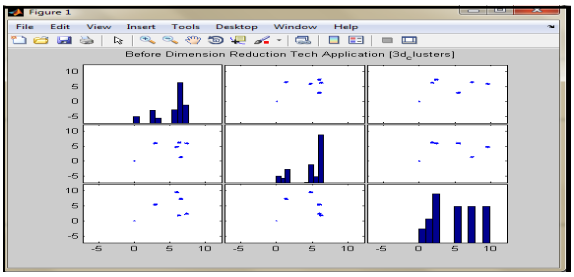
Table 5: shows the results for Helix dataset by using CFA Technique

S. No	Dataset	Min	Max	Mean	Std	Var	Co-var	No. of Dimensions
1.	Helix (Before)	0	0	0.4984	0.0748	0	0.0602,0.0004,0.0000, 0.0004,0.0602, 0.0002, -0.0000,0.0002, 0.0134	500,3
2.	Helix (After)	0.0352	0.0352	0.0462	3.4380e-17	1.6018e-64	1.0e-30 *, 0.4342 , -0.3329 -0.3329, 0.2552	500,2

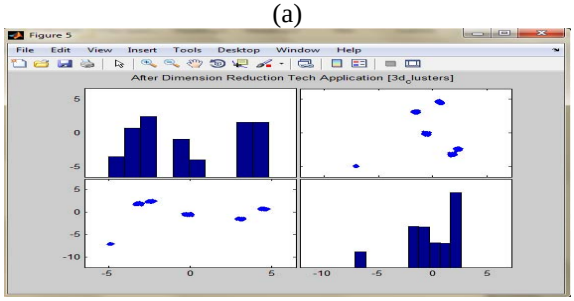
Table 6: Shows the results for twin peaks dataset by using CFA technique

S. No	Dataset	Min	Max	Mean	Std	Var	Co-var	No. of Dimensions
1.	Twin peaks (Before)	0	0	0.4997	0.1435	0	0.0007,-0.0000,-0.0002, -0.0000, 0.0007, 0.0003, -0.0002, 0.0003, 0.0759	500,3
2.	Twin peaks (After)	0.0206	0.0206	0.0303	1.9573e-16	2.3122e-62	1.0e-30 * 0.0464 , -0.0733 -0.0733 , 0.1158	500,2

Visualization Results: Visualization helps to graphical depict the underlying knowledge in data. The dimension reduction GPLVM technique is used. The results are shown in the form of figures before Dimension reduction and after Dimension reduction.

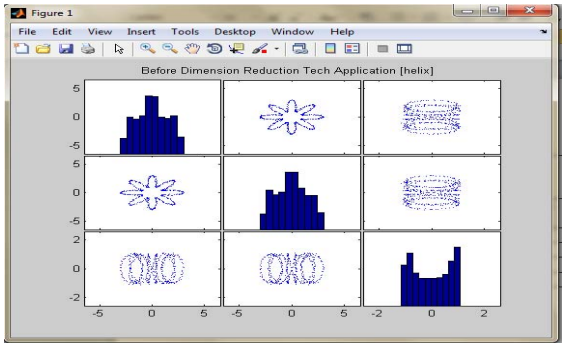


(a)

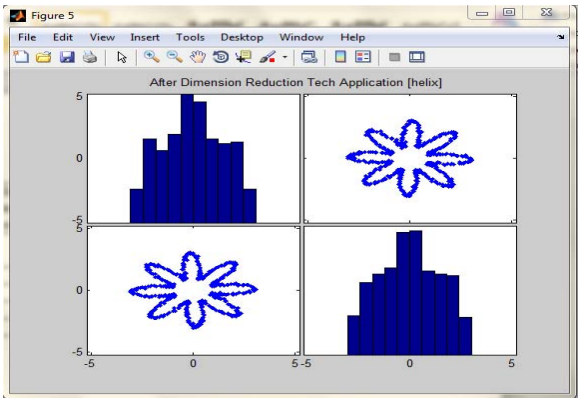


(b)

Figure 2: Plot Matrix visualization for 3D Cluster dataset. (a) Represents the data with original dataset (b) shows the data with clutter reduced.

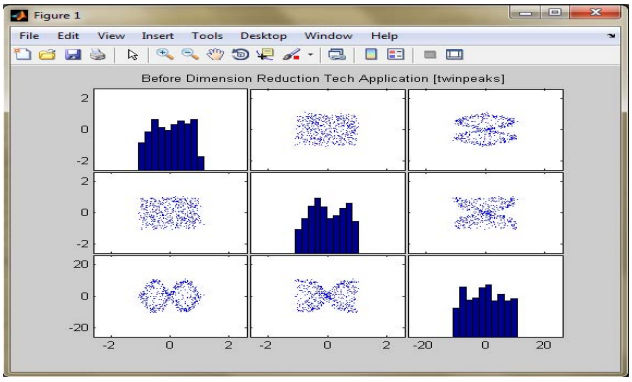


(a)

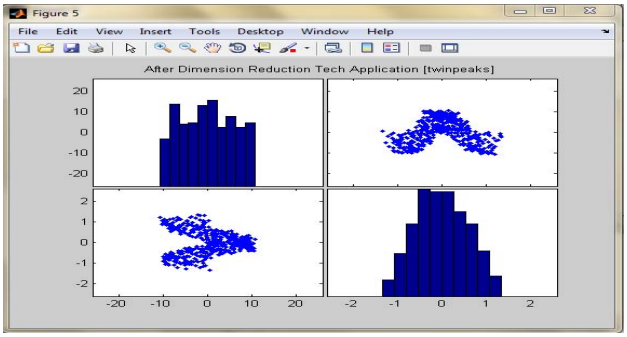


(b)

Figure 3: Plot Matrix visualization for Helix dataset. (a) Represents the data with original dataset (b) shows the data with clutter reduced.



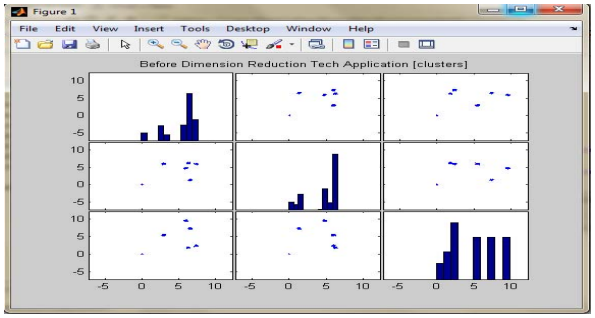
(a)



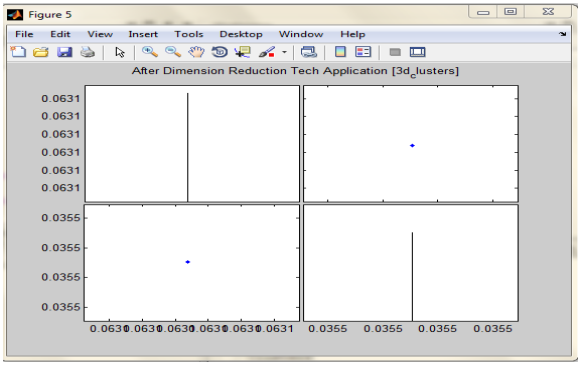
(b)

Figure 4: Plot Matrix visualization for Twin Peaks dataset. (a) Represents the data with original dataset, (b) shows the data with clutter reduced.

After GPLVM Technique, the second dimension reduction CFA technique is used. CFA is used to reduce the co-ordinates to the lower dimensional space. The visualization results for plot matrix visualization by using CFA technique are shown below:

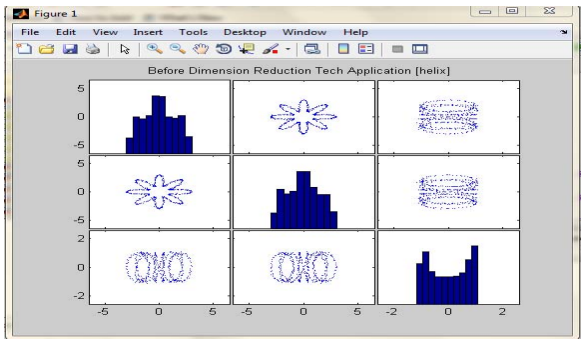


(a)

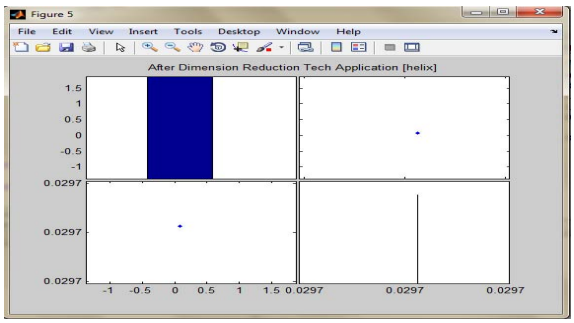


(b)

Figure 5: Plot Matrix visualization for 3D Cluster dataset. (a) Represents the data with original dataset (b) shows the data with clutter reduced.

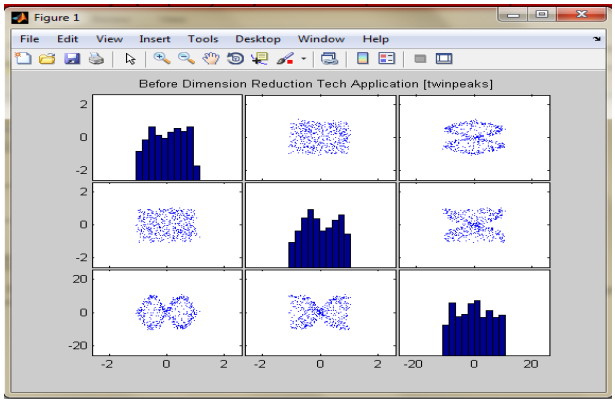


(a)

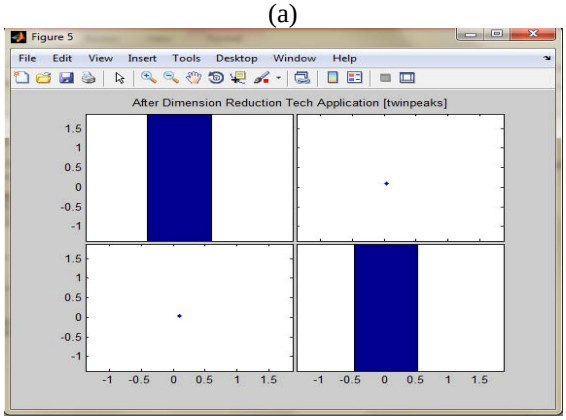


(b)

Figure 6: Plot Matrix visualization for Helix dataset. (a) Represents the data with original dataset (b) shows the data with clutter reduced.



(a)



(b)

Figure 7: Plot Matrix visualization for Twin Peaks dataset. (a) Represents the data with original dataset (b) shows the data with clutter reduced.

5. 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 results show that the minimum and maximum values are same which shows that the clutter is removed from the dataset. By using clutter based dimension reduction the visualization is improved of the datasets. The visualization has now more information gain and reveals the structure more clearly. Coordinates factor analysis (CFA) is used to reduce the N-dimensional coordinate system to a much smaller K-dimensional space. The results are shown before and after dimension reduction CFA technique.

6. Future Scope

Future work will include the combination of dimension reduction approach with other approaches. In our paper two techniques GPLVM and CFA techniques are used. In future more techniques can be used to reduce the clutter and make the visualization better.

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. learning an internet co-ordinate system by dilip antony joseph

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

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

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

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

[10] A Novel Approach for Comparison of Clustering Algorithms on CAD Images by shelza, balwinder singh

Author Profile

Harpreet Kaur is currently persuing the M. Tech in computer science and engineering from Swami Vivekanand Institute of Engineering & Technology, Banur, Punjab. She holds the degree of B. Tech in Computer Science and Technology from Shaheed Udham Singh College of Engineering and Technology, Tangori, Punjab.

Er. Shelza is currently working as Assistant Professor in Computer Science and Engineering Department at Swami Vivekanand Institute of Engineering and Technology, Banur. She has completed her M. Tech in Computer Engineering from Yadwindra College of Engineering and Technology, Talwandi Sabo affiliated to Punjabi University Patiala .she holds the degree of B. Tech in Computer Science and Technology from Sant Longowal Institute of Engineering & Technology, Longowal Distt Sangrur,Punjab. She has 7 years six months experience in teaching