High Dimensional Data Visualization: Advances and Challenges

Fisseha Gidey G.
Big Data Visualization Algorithms
National Advanced School of Engineering,
University of Yaoundé I

Charles Awono Onana
Big Data Technology
National Advanced School of Engineering,
University of Yaoundé I

ABSTRACT
Recent technological advances and availability of computing resources resulted in a massive growth of data size, dimensions and complexity. Data visualization is a good approach when dealing with large scale high dimensional datasets as it will provide the opportunity to understand what's in the data and where to focus. However, the ever increasing dimensions of datasets, the physical limitations of the display screen (2D/3D), and the relatively small capacity of our mind to process complex data at a time pose a challenge in the process of visualization. This paper describe the advancements made so far in visualizing high dimensional data and the challenges that should be addressed in future researches.

General Terms
Data Visualization, Dimension Reduction

Keywords
Big Data, Data Visualization, Dimension Reduction, PCA, Sammon's Mapping, and MDS

1. INTRODUCTION
In the era of big data there is a need to process large amounts of data to find some patterns and hidden structures in the data to use it for further analysis. It is very important and yet difficult to find any insight if the data have many dimensions. A data with large number variables of observation relative to the sample size is called high dimensional data and occur in diverse fields such as Science, Astronomy, Chemistry, Engineering, Economics, and Biology to name a few. The problem of driving an insight from a very large, complex, and high dimensional data to make a knowledge based decision is an active area of research that attracted the attention of many researchers from diverse fields of study. Visualization is an important approach when dealing with large masses of high dimensional data.

A great number of dimensionality reduction approaches for high dimensional data visualization have been introduced in the last decades. However, recent technological advances and emergence of large scale datasets have clearly indicated limitations of existing methods in these new settings [7, 15, and 16] and there remains a clear need for the development of novel visualization approaches. Indeed, dimensionality reduction transforms a high dimensional data in to a lower dimensional space but with unavoidable information loss. The focus of researches in visualization is moving in to finding a novel approach that reduces dimensions with the lowest information loss.

The remaining part of this paperis organized in the following manner: Section 2 provides the definition of data visualization and the developments it has made so far, section 3 presents the different types of dimensionality reduction for visualizing high dimensional data and make a systematic comparison between these techniques. The paper ends with a conclusion and future research directions.

2. DATA VISUALIZATION
Data visualization, the presentation of data in a pictorial or graphical form, has been practiced for hundreds of years. It is an ancient practice that dates back to the 2nd century AD and has shown several developments since then [4, 9 and 21] and the challenge arises from the characteristics of large scale high dimensional data. Since then it has shown several developments and recently becomes an integral part of academia and business. In fact, most of the developments occurred in the last two and half centuries, most importantly in the last 30 years. Their wide availability, increasing size, and complexity have led to new challenges and opportunities for their effective visualization as the role of visualization is always to manifest the right decision or action but not replacing critical thinking [15 and 22].

In the past decades, a variety of approaches have been introduced to visually convey high dimensional structural information by utilizing low dimensional projections or abstractions: from dimension reduction to visual encoding, and from quantitative analysis to interactive exploration. Despite its long history, there is a clear need for further development of visualization methods when working with large scale high dimensional data where commonly used visualization tools are either too simplistic to gain a deeper insight in to the data properties, or are too cumbersome or computationally costly [16]. Most importantly, the ever increasing dimensionality of data, the physical limitations of the displaying devices (2D/3D), and the relatively low capacity of our mind to process complex information at a time made the visualization of high dimensional data difficult [22].

A great number of approaches for data visualization have been introduced in the last 30 years [7, 8, 9, 12, 21, and 22] such as histogram, x-y plots, line plots, scatter plots, flow charts, time line, bubble chart, Venn diagram, and pie charts often called conventional methods of visualization which has been used for more than a century in a nearly unchanged form. These techniques are useful for data exploration but are limited to relatively small and low dimensional datasets. However, recent technological advances and emergence of large scale datasets have clearly indicated limitations of the existing methods in these new settings [7, 15, and 16] and there remains a clear need for the development of novel visualization approaches.

Visualization of high dimensional data in low dimensional space is an essential tool for exploratory data analysis, especially to reveal hidden relationships concealed by the inherent complexity of data. That's what dimension reduction

is supposed to do as visualization is moving from the traditional approach, displaying results, in to finding a meaningful information and relationship for further analysis.

Consider the following figure on unemployment rate of an ideal country over ten years to understand how visualization simplifies data. For instance, the gap between men and women unemployment rate was narrow at some point and one might inquire and infer why or how that happened which would be difficult to notice otherwise in tabular form.

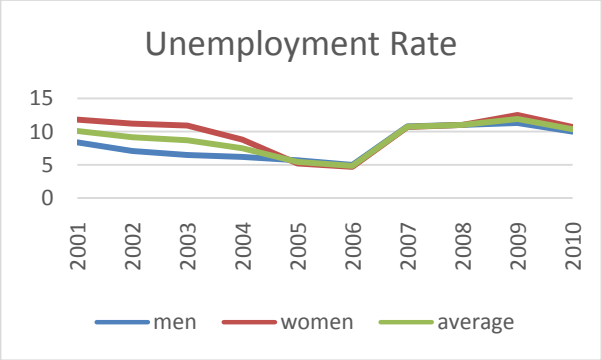

Fig 1: Visualizing unemployment rate of men and women of an ideal country

3. DIMENSIONALITY REDUCTION (DR) TECHNIQUES

Methods of dimensionality reduction provide a way to understand and visualize the structure of complex datasets. It is used to reduce redundancy and variance, improve accuracy, visualize high dimensional data, and recover intrinsic dimensions. In visualization, dimensionality reduction is an approach used to transform high dimensional data (D)in to a lower dimensionality vector space$(d, d \ll D)$while trying to preserve as much information and relationships as possible from the original data. However, it is impossible to avoid information loss but minimize it and hence transforming high-dimensional data into a lower-dimensional version that preserves as much information and relationship as possible from the original data is a research area widely studied [18, 19, 22, 23, and 24], given its ability to reduce the computational cost and/or improve the performance of both pattern recognition and information visualization systems.

Tree Map, Parallel Coordinate, Cone Tree, Heat Maps, Star Coordinates, Self-organizing map, and Chernoff faces has emerged as extensions to the conventional visualization techniques to handle big data yet far from enough [5, 7, 11, 13, 15, 16, and 19] as they suffer from dimensions. The other extensions such as Principal Component Analysis and Multidimensional Scaling suffer from being based on linear models. Only recently, few strategies are able to reduce the data dimensionality in a nonlinear way.

Parallel coordinate is a century old dimension reduction techniques useful in displaying multidimensional data. It is used to plot individual data elements across many dimensions and is very useful when to display multidimensional data for comparison [15]. It maps the k dimensional space on to the two dimensions by using k-equidistant axes which are parallel to one of the display axis [7 and 15]; however, it falls short as the dimension increases. The star coordinate models are probably the most scalable techniques for visualizing large datasets compared with other multidimensional visualizations methods such as parallel coordinates and scatter plot matrix [13].

Generally, techniques for dimensionality reduction are subdivided in to convex and non-convexwhere the convex techniques optimize an objective function that doesn't contain any local optima and the non-convex techniques optimizes the objective function that do contain local optima [14]. Comparatively, the nonlinear techniques are in a better position to deal with complex nonlinear data than the linear dimensionality reduction techniques for visualization.

3.1 Principal Component Analysis (PCA)

Principal component analysis which is also known as Hotteling or Karhunen is the most commonly used classical algorithms for dimensionality reduction. Unlike the metric and non-metric dimensionality reduction techniques PCA tries to preserve the variance of the data than preserving the distance or the global ordering relations of the objects. For a given high dimensional dataset, PCA finds the vectors along which the data has maximum variance [1, 3, 11, and 22]. Generally, PCA transforms the data in to a new coordinate system in such a way that the largest variance by any projection of the data comes to lie in the first coordinate , the second largest variance on the second coordinate and so on. PCA is useful when the data lies on or close to a linear subspace of the dataset. Suppose $x \in \mathbb{R}^{N \times D}$is a matrix whose rows are D-dimensional data points. We are looking for the d orthogonal vectors along which the data has maximum variance. If in fact the data lies perfectly along a subspace of $\mathbb{R}^{D}$, PCA will reveal that subspace; otherwise PCA will introduce some error.

It optimizes the objective function

$$\underbrace{Maximize}_{V} \; Var\,(XV)$$

$$Subject\ to\ V^T V = I$$

Where$x \in \mathbb{R}^{N \times D}$, data matrix $V \in DX$ has its columns as the direction of maximum variance.

As PCA is a linear dimensionality reduction it cannot unfold the low dimensional manifolds embedded in to the high dimensional vector space. The power of PCA algorithm has been extended by applying a kernel trick named as Kernel PCA yet it falls short in handing nonlinear high dimensional data.

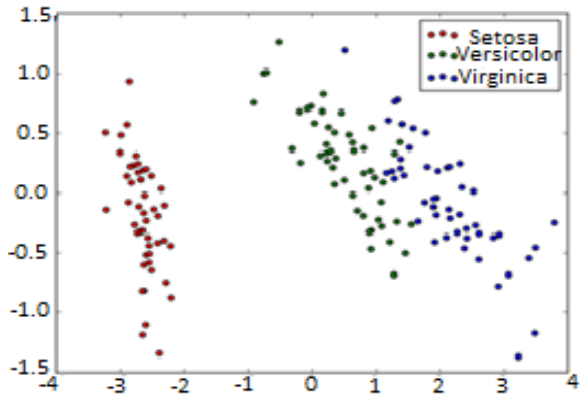

Fig 2: PCA mappings of Iris data set

3.2 Sammon's Mapping

Sammon's mapping is one of the most popular metric, nonlinear dimensionality reductions method [2, 11, and 17] widely applied in visualizing high dimensional data. It is a simple yet very useful nonlinear projection algorithm that maps the high dimensional (n) features in the original data in

to fewer variables (m dimensions,$m < n$) by preserving the inherent structure of the data. Generally speaking, Sammon mapping attempts to preserve the structure of high(n) dimensional data by finding N points in a much lower (m) dimensional data space such that the inter-point distance measured in the m-dimensional space, approximate the corresponding inter-point distance in the n-dimension space.

While PCA tries to preserve the variance of the data, Sammon mapping try to preserve the inter-pattern distances by optimizing a cost function that describes how well the pair wise distances in a dataset are preserved. This can be achieved by minimizing an error criterion, called Sammon's stress.

The Sammon's stress is looking for Y by minimizing

$$E = \frac{1}{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d(i,j)} \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \frac{\left(d(i,j) - d^*(i,j)\right)^2}{d(i,j)}$$

The minimization of the error function E is an optimization problem in the new variables$Y_{ij}\,(i = 1, 2, \ldots N; j = 1, 2, \ldots m)$. Here, we assume that $X = \{x_k : x_k = (x_{k1}, x_{k2}, \ldots, x_{kn})^T, k =$ *1, 2, ..., N* is the set of n input vectors, and *Y=yk: yk=(yk1, yk2, ...,ykm)T, k=1, 2, ..., N* is the unknown projected vectors to be found where $d_{ij} = d(x_i, x_j) and d_{ij}{}^* = d(y_i, y_j)$ are the Euclidean distances between $x_i \& x_j$ and $y_i \& y_j$ respectively. Similar to PCA, the Sammon mapping cannot unfold the nonlinearly embedded two dimensional manifolds. Moreover, it is not applicable to large N as the algorithm involves a large number of computations in every iteration step which require the computation of $\frac{N(N-1)}{2}$ distances.

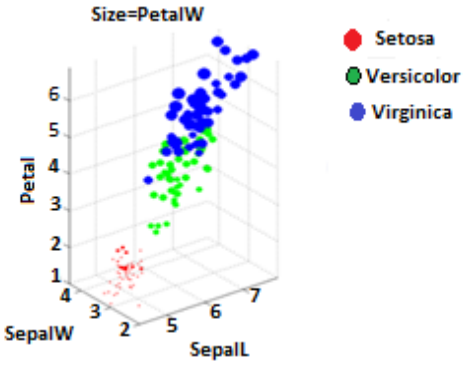

Fig 3: Sammon mapping of Iris data set

3.3 Multidimensional Scaling (MDS)

Multidimensional Scaling refers to a group of unsupervised data visualization techniques. It is used to visualize a given high dimensional data in a low dimensional space by generating a configuration of the given data utilizing the Euclidean low dimensional space. The idea is based on the analysis of similarity or dissimilarity data on a set of objects [10, 11, and 20]. The purpose of MDS is mainly to represent similarity or dissimilarity of data as distance in low dimensional space to make data accessible for visual inspection and exploration. Given a set of data in a high dimensional feature space, MDS maps them in to a low dimensional data space in a way that objects that are very similar to each other in the original space are placed near each other on the map, and objects that are very different from each other are place far away from each other.

There are two kinds of MDS known as metric and non-metric MDS. Metric MDS discovers the underlying structure of dataset by preserving similarity information (pair wise distance) while the non-metric MDS attempts preserve the rank order among the dissimilarities. Moreover, these two approaches are different in that while metric MDS algorithm is an algebraic method, the non-metric MDS is an iterative mapping process. Both approaches utilize optimization problem though the main goal of the optimization differs significantly.

The metric MDS, similar to Sammon's mapping, tries to minimize a stress function given by the following equation

$$E_{metric\ MDS} = \frac{1}{\sum_{i<j}^{N} d_{i,j}{}^{*2}} \sum_{i<j}^{N} \left(d_{i,j}{}^* - d_{i,j}\right)^2,$$

Where $d_{i,j}{}^*$ denote the distance between vectors$x_i \& x_j$ and $d_{i,j}$ between $y_i \& y_j$, respectively.

MDS enables a data analyst to literally look at a data and explore the structure visually by displaying the correlation graphically. It is a good approach for global structure but it is complex, not suitable for large graphs, and not applicable to nonlinear dimensional.

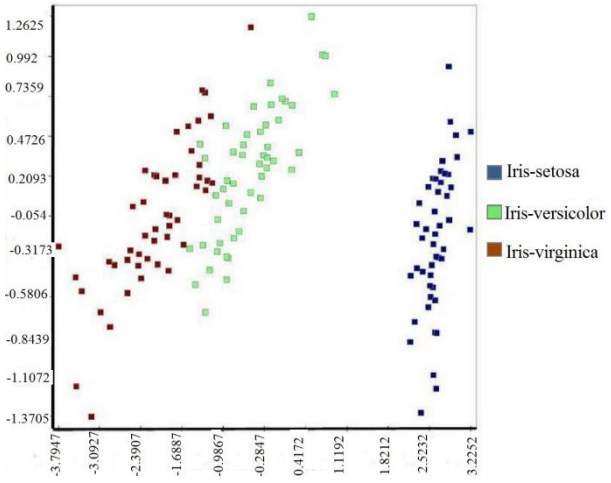

Fig 4: MDS mapping of Iris data set

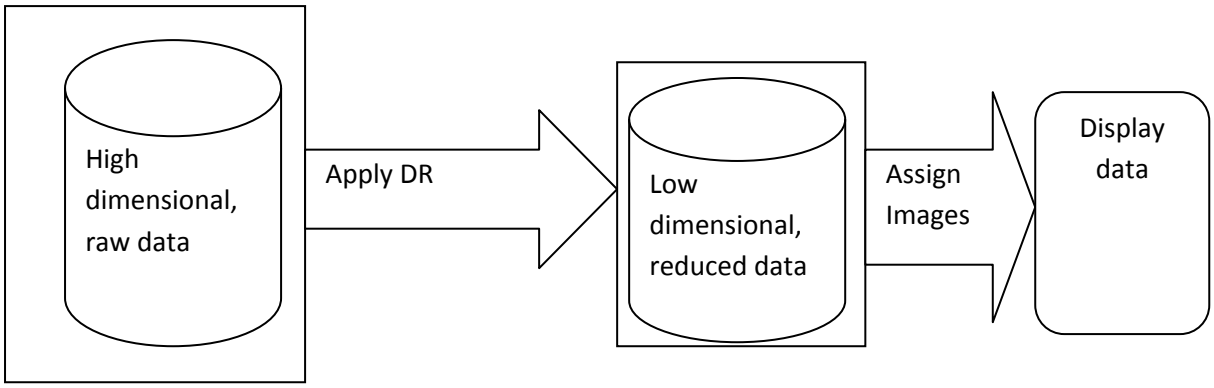

Fig 5: Visualization process for high dimensional data

4. CONCLUSION

Visualization is important when analyzing multidimensional datasets since it helps humans discover complex relationships, yet it is challenging. Most of the developments occurred in the past three decades where visualization transformed itself as a tool for exploratory data analysis than being used for result communication through conventional visualization methods. Visualization techniques can be evaluated and compared with respect to their suitability for certain data characteristics such as data types, number of dimensions, number of data items, and category [3, 6]. From the observation, it can be deduced that there is no best dimension reduction techniques as the result vary from data to data and hence every dimensionality reduction have its own pros and cons.

Despite the existence of varieties of new dimensionality reduction techniques to visualize high dimensional data, there is a clear need for further development. This is because either existing techniques fall short as dimension increases or are designed to handle only linear problems. Even those few strategies available to handle nonlinear problems don't specify to what extent they preserves the data structure and relationship after reduction.

The observationsmade on different studiesregarding dimensionality reduction for visualization indicates that, future researches should focus in developing better algorithms for visualizing high dimensional data that consider large dimensions, physical limitations of the display screen, and small capacity of our mind to process complex information at a time. Moreover, how much of the information has been preserved from the original data after dimension reduction is made and how can we evaluate that is another problem that should be addressed in future researches. The future focus of this paper is however limited to building a novel visualization algorithms for high dimensional datatofill the aforementioned gaps based on existing techniques.

5. REFERENCES

[1] Agnes Vathy-Fogarassy and Janos Abonyi. 2013. Graph Based Clustering and Data Visualization Algorithms, Springer.

[2] Anna Maria K. and Janos Abonyi. Visualization of Fuzzy Clustering Results by Modified Sammon Mapping. University of Veszprem

[3] ArpitJangid and Soren Goyal. 2015. Techniques for visualizations of high dimensional data, convex optimization.

[4] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE

[5] C. Donalek, S G., Djorgovski, et al. 2014. Immersive and collaborative data visualization using virtual reality platforms. IEEE International conference

[6] Daniel A. Kein. 2001. Visual Exploration of Large Datasets. Communication of the ACM

[7] Daniel A. Keim. 2002. Information Visualization and Visual Data Mining. Transaction on visualization and computer graphics, IEEE, vol 7

[8] Datos.gob.es. 2013. Data Processing and Visualization Tools. European Public Sector Information Platform, Topic Report No: 2013/07

[9] Edward R. Tufte (1997): Visual Explanations: Images, Quantities, Evidence, and Narrative. Graphics Press.

[10] Ingwer Borg and Patrick Groenen. 1997. Modern Multidimensional Scaling: Theory and Applications. Springer

[11] Johan A.K Suykens. 2008. Data Visualization and Dimensionality Reduction Using Kernel Maps with a Reference point. IEEE Transaction on Neural Networks

[12] Justin Choy. 2012. Visualization Techniques from Basics to Big Data with SAS Visual Analytics. SAS Global Forum

[13] K. Chen. 2014. Optimizing Star-Coordinate Visualization Method for Effective Interactive Cluster Exploration on Big Data. Intelligent Data Analysis

[14] Laurens Van der Maaten, Eric Postme, Jaap Van den Herik. 2009. Dimensionality reduction: A comparative Review. Tilburg University.

[15] Lidongwang, Guanghuiwang, and Cheryl Ann Alexander. 2015. Big Data Visualization: Methods, Challenges, and Technology Progress

[16] NemanjaDjuric. 2014. Big Data Algorithms for Visualization and Supervised Learning

[17] NicolaeApostolescu and Daniela Baran. 2016. Sammon Mapping for preliminary Analysis in Hyperspectral Imagery. The 36th IEEE "Caius Iacob" conference on Fluid mechanics

[18] Olga Kurasove, VirginijusMarcinkevicius, and Victor Medvedev. 2014. Strategies for Big Data Clustering.

IEEE 26th International Conference on Tools with Artificial Intelligence

[19] PoojaChenna. 2016. Comparative Study of Dimension Reduction Approaches with Respect to Visualization in 3-Dimensional Space. Kennesaw State University, Springer.

[20] SeungHeeBae and Judy Qiu. 2012. High Performance Multidimensional scaling for Large High Dimensional Data Visualization. IEEE Transaction of parallel and Distributed System.

[21] Stephane Few. 2007. Data Visualization Past, Present, and Future. Perceptual edge

[22] S.Lui, D. Maljovich, B. Wang, P.-T. Bremer and V.Pascucci. Visualizing High dimensional data: Advances in the past decade

[23] http://www.gartner.com:Big Data means Big Business. Douglas Laney. Gartnerinc.

[24] http://SSrn.com: Big Data for development:-From Information to Knowledge Societies,2013