



Faculty of Cognitive Sciences and Human Development

**DATA MINING AND DATA VISUALIZATION USING
SELF-ORGANIZING MAP (SOM)**

Ahmad Naufal Khan Bin Jamil Khan

(45433)

**Bachelor of Science with Honours
(Cognitive Science)
2017**

**DATA MINING AND DATA VISUALIZATION USING SELF-ORGANIZING MAP
(SOM)**

AHMAD NAUFAL KHAN BIN JAMIL KHAN

This project is submitted
in partial fulfilment of the requirements for a
Bachelor of Science with Honours
(Cognitive Science)

Faculty of Cognitive Sciences and Human Development

UNIVERSITI MALAYSIA SARAWAK

(2017)

UNIVERSITI MALAYSIA SARAWAK

Grade: A -

Please tick one

Final Year Project Report

Masters

PbD

DECLARATION OF ORIGINAL WORK

This declaration is made on the 8 day of JUNE year 2017.

Student's Declaration:

I, AHMAD NAUFAL KHAN BIN JAMIL KHAN, 45433, FACULTY OF COGNITIVE SCIENCES AND HUMAN DEVELOPMENT, hereby declare that the work entitled, DATA MINING AND DATA VISUALIZATION USING SELF-ORGANIZING MAP(SOM) is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

9 JUNE 2017

AHMAD NAUFAL KHAN (45433)

Supervisor's Declaration:

I, AP DR. TEH CHEE SIONG, hereby certify that the work entitled, _____ was prepared by the aforementioned or above mentioned student, and was submitted to the "FACULTY" as a *partial/full fulfillment for the conferment of BACHELOR OF SCIENCE WITH HONOURS (COGNITIVE SCIENCE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by:


(AP DR. TEH CHEE SIONG)

Date.

9 JUNE 2017

I declare this Project/Thesis is classified as (Please tick (√)):

- CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*
- RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*
- OPEN ACCESS**


I declare this Project/Thesis is to be submitted to the Centre for Academic Information Services (CAIS) and uploaded into UNIMAS Institutional Repository (UNIMAS IR) (Please tick (√)):

- YES**
- NO**

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself / herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature : 
Date: 9 June 2017

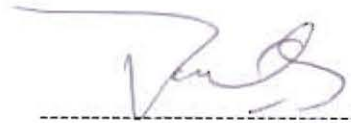
Supervisor's signature : 
Date: 9 June 2017

Current Address: Jalan Datuk Mohammad Musa, 94300 Kota Samarahan, Sarawak, Malaysia

Notes: * If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.

The project entitled Data Mining and Data Visualizations Using Self-Organizing Map (SOM) was prepared by Ahmad Naufal Khan bin Jamil Khan and submitted to the Faculty of Cognitive Sciences and Human Development in partial fulfillment of the requirements for a Bachelor of Science with Honours (Cognitive Science).

Received for examination by:



(AP DR. TEH CHEE SIONG)

Date:

9th June 2017

Grade

A-

ACKNOWLEDGEMENT

I would like to thank God for giving me this opportunity to complete my thesis in a manner I truly am satisfied. I thank god for showing me the way when I felt lost, for giving me hope when giving up seems like the only option. Without Him, my work will never be completed.

To my beloved and respected supervisor, AP Dr Teh Chee Siong, you are my biggest inspiration of all time. I would like to give the highest appreciation to my dear supervisor for showing the path to complete this thesis. Thank you for handing me the faith to work on this wonderful topic. You've showed me what it means to be a responsible person and to work through every hardship.

I also would like to thank my parents. They have always encouraged me to continue work and convince me that I am capable of doing it. They always believe in me and that has been a major motivation for me to go on. Whenever I feel like giving up, they convince me to continue till the end. I thank them for that. To the rest of my family, I thank you all for your undying prayers and help for all my life. To those people who have stick with me through thick and thin, you will always be in my prayers. You are truly a precious gem

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iii
LIST OF FIGURES	vi
ABSTRACT	vii
ABSTRAK	viii
CHAPTER 1	1
1.0 INTRODUCTION.....	1
1.0.1 Preliminaries	1
1.0.2 Artificial Intelligence (AI).....	2
1.0.3 Biological Neural Network (BNN).....	3
1.0.4 Artificial Neural Network (ANN).....	4
1.0.5 Data visualization.....	5
1.1 Problem statement	5
1.2 Research Objectives	7
1.3 Specific Objectives	7
1.4 Research Scopes.....	7
1.5 Research Methodologies.....	7
CHAPTER 2	9
2.0 LITERATURE REVIEW.....	9
2.1 Introduction	9
2.2 Dimension Reduction	9
2.3 Data visualization.....	9
2.4 Visualization of high-dimensional data items.....	10
2.5 Data visualization in the past and current	11
2.6 Principal Component Analysis (PCA).....	12
2.7 Sammon's Mapping (SM).....	13
2.8 Self-Organizing Maps (SOM).....	14
CHAPTER 3	16
3.0 THE PROPOSED DATA VISUALIZATION METHOD	16
3.1 Introduction	16
3.1 Self-Organizing Map (SOM).....	16
3.2 Advantages of Self-Organizing Map (SOM).....	20
3.3 Issues Related to SOM	21

CHAPTER 4	23
4.0 EXPERIMENTS ON BENCHMARK DATASETS	23
4.1 Introduction	23
4.2 Data Visualization.....	23
4.3 Visualizations comparison of various datasets	23
4.4 Iris Datasets.....	24
4.5 Wine Datasets	35
4.6 Glass Identification Dataset	40
CHAPTER 5	43
5.0 DISCUSSION AND CONCLUSION.....	43
5.1 Introduction	43
5.2 Data visualization Analysis	43
5.3 Data construction.....	43
5.4 Data normalization.....	43
5.5 Map Training	44
5.6 Data Visualization.....	44
5.7 SOM_SHOW command in SOM Toolbox	45
5.8 SOM_SHOW_ADD command in SOM Toolbox	45
5.9 SOM_GRID command in SOM Toolbox.....	46
5.10 Future applications of SOM	46
REFERENCES	48

LIST OF FIGURES

Figure 1 Representation of Kohonen SOM algorithm	17
Figure 2 Flow chart of training process in SOM	18
Figure 3 Map grid showing how clustering works	19
Figure 4 Histogram and scatter plots of Iris Datasets	24
Figure 5 U-matrix / Distance matrix of Iris Dataset	27
Figure 6 U-matrix and empty label grid	28
Figure 7 Labelled U-matrix for each class	39
Figure 8. Hit histogram showing the BMU	30
Figure 9. Coloured hit histograms according to their classes	31
Figure 10. Visualizations of Iris datasets in map grid (3-dimensional)	33
Figure 11. Many forms of U-matrix representations	37
Figure 12. U-matrix are labelled by Wine classes	39
Figure 13. Map grid of Wine datasets (3-dimensional)	40
Figure 14. Distance matrix of glass identification data and its variables	41
Figure 15. Labelled distance matrix of glass identification data.	42
Figure 16. Colored distance matrix. 6 classes mean 6 colors are used to label.	42
Figure 17. Distance matrix of glass-identification	46

ABSTRACT

Data mining and data visualizations are becoming essential parts in information technology in recent era. Without the existence of data mining technology, big data can be near impossible to be extracted and have to be done manually. With the aid of data mining technology, now information can be gathered from datasets at much shorter time. The discovery of data visualizations also aids in managing data into presentable form that can be understood by everyone. Big dimensions can now be reduced to help data be more understandable. In this thesis, Kohonen self-organizing map(SOM) technique is discussed and examined for data mining and data visualizations. SOM is a neural network technique that can performs data mining, data classification and data visualizations. SOM Toolbox was used on MATLAB. All steps in SOM are explained in details from weight initialization until training is stopped. Graphical explanations of how SOM works are also used to help visualize SOM algorithm. Many methods of visualizations are displayed in the thesis. All methods are critically analysed. Few benchmark datasets are used as examples of the visualization techniques. Such examples are Iris datasets and Wine datasets. The strength and weaknesses are listed out in discussion section.

Keywords : Data mining, data classification, data visualization and self-organizing map,

ABSTRAK

Pemahaman makna data dan data visualisasi adalah sesuatu yang sangat penting dalam bidang teknologi maklumat pada era baru ini. Tanpa kewujudan teknologi pemahaman makna data ini, data yang besar boleh dikatakan hampir mustahil untuk diolah dan diekstrak. Oleh itu, data besar ini perlu dilakukan secara manual. Dengan adanya teknologi pemahaman makna data ini, kini maklumat dapat dikumpulkan dari set data pada masa yang lebih singkat. Penemuan penggambaran data juga membantu dalam menguruskan data ke dalam bentuk yang rapi dan boleh difahami semua lapisan masyarakat. Dimensi besar juga boleh melalui proses pengurangan dimensi untuk menjadikan data lebih mudah difahami. Dalam tesis ini, Kohonen telah mereka suatu teknik yang dipanggil *Self-Organizing Map* (SOM). Teknik ini dibincang dan diteliti bagi proses pemahaman makna data dan data visualisasi. SOM ialah satu teknik rangkaian neural yang boleh melakukan pengolehan data, klasifikasi data dan penggambaran data dengan tepat. *SOM Toolbox* telah digunakan dalam aplikasi MATLAB. Semua langkah dalam SOM dijelaskan secara terperinci dari permulaan pemberat sehinggalah latihan diberhentikan yang menandakan proses sudah tamat. Penjelasan grafik tentang gerak kerja SOM juga digunakan untuk membantu menggambarkan algoritma SOM. Banyak kaedah penggambaran yang dipaparkan dalam tesis ini. Kesemua kaedah penggambaran ini juga dianalisa secara kritikal. Set data penanda aras digunakan sebagai contoh untuk memaparkan teknik-teknik visualisasi ini. Antara contoh set yang digunakan ialah set Iris Flower dan data Wine. Kekuatan dan kelemahan bagi teknik SOM dirungkaikan dalam seksyen 5 tesis ini.

Kata kunci : Perlombongan data, klasifikasi data, visualisasi data dan *Self-Organizing Map*.

CHAPTER 1

1.0 INTRODUCTION

1.0.1 Preliminaries

The pattern recognition related studies are concerned with teaching a machine to identify pattern of interest from its background. Little did we know, humans are born with such gifted abilities that we tend to take for granted. Humans are able to identify patterns even as babies. This given ability may be better than what machines are able to do even at current level of technology.

However, this is limited to a certain extend at which humans are better than machines. Given a big dataset, humans' minds may be able to perform data recognition but at slow speed compared to machines. These massive datasets are also demanding a thrust in data visualization technology, producing massive big leaps in related researches. This field of study aims to increase our understanding of datasets, what useful information is latent in it and how to detect the portion of it that are of strong interest (Fayyad, Grinstein & Wierse, 2002). A good portion of data visualization related researches focuses on making it easier to visualize more dimensions effectively using 2D-3D display technologies available (Fayyad et al., 2002).

Data visualization is the presentation of data in a pictorial or graphical format. It enables us to see analytic data in visual form instead of plain numbers. This will allow humans to identify patterns or understand hard concepts. With the help of such interactive visualization, humans can use this technology to present information in charts or graphs for more

detailed-oriented presentation, enabling humans mind to encapsulate new information easier (Tufte, 1983).

This study is associated with Artificial Intelligence (AI), which focuses on creating computer systems that can engage on behaviors that humans consider intelligent. In current time, computational computations use algorithmic approach where the computers must know the specific problem solving steps. However, Artificial Intelligence proves to be different by depicting paradigm for computing rather than the conventional computing (Teh, 2006).

Artificial Neural Networks (ANN), one of the highlighted study area under Artificial Intelligence is created through inspiration of how biological nervous system processes information especially in the brain. Most of the work in ANN is related to neurons and how they train data. In data visualization, it is also important to classify data first so that we can visualize data better and clearer. ANN is useful for this purpose because it can be programmed to perform specific tasks (Teh, 2006).

1.0.2 Artificial Intelligence (AI)

AI has been defined in many ways to represent different point of views in literature (Kumar, 2004). In general, AI aims to develop paradigms or algorithms which in attempt to causes machines to perform tasks that requires cognition or perception while performed by humans (Sage, 1990). Simon (1991) defined AI as :

“We call a program for a computer artificially intelligent if it does something which, when done by a human being, will be thought to require human intelligence.”

Wilson (1992) also provided a definition for AI as following:

“Artificial Intelligence is the study of computations that make it possible to perceive reason and act.” (Wilson,1992,as cited in Kumar 2004,p.12).

According to Kumar (2004), any artificial should consists three elements in essence. First of all, AI system has to be able to handle knowledge that is both general and domain specific, implicit and explicit and at different levels of abstractions. Secondly, it should have suitable mechanism to constrain the search through the knowledge base, and able to arrive at conclusion from premises and available evidence. Lastly, it should possess mechanism for learning new information with minimal noise to the existing knowledge structure, as provided by the environment in which the system operates at.

The AI study area is often related to biological neural system as it is hugely motivated by the biological intelligence system provided by nature.

1.0.3 Biological Neural Network (BNN)

The human brain is one of the most complex system ever discovered. It is more complex than any machine ever created. This is because our brains contains billion of cells that regulates the body, stores information and retrieving them as well as making decisions in a way that humans are still not able to explain (Swerdlow, 1995). This complex organ inside humans are made of hundred billion information-processing units called neurons. These neurons are probably the most vital units that enables humans brain to perform all the actions that they are capable of. Adding to the massive number of neurons, each of them consists of an average of 10,000 synapses that pass along information between them in incredibly fast speed. This results in total number of connections in the network to reach up to a huge figure of approximately 10^{15} .

In depth, each neurons contain dendrites spreading from each neurons. These tree-like structures transports signals from one neuron to other neuron. At the end of the neuron, there exists soma, a portion of the neuron that contains the nucleus and other vital components that decide what kind of output signals the neuron will give off in accordance to input signals received by other neurons. Afterwards, there's another important structure called axon, which carries the output signal produced in soma towards other neurons. At the end of the axon, there is located a synaptic terminal which is the transfer point of information. The synaptic terminals are not connected with each other. There are gaps between them called the synapses. This completes the so-called sophisticated neural networks in our brains. This is a single process that occurs while in fact in our brain, there are billions of same processes occurring within milliseconds.

1.0.4 Artificial Neural Network (ANN)

ANNs are humans attempt on making biological neural network into computational network. Just like neurons in our brains, it is able to solve complex computations through self-organizing abilities by learning information (Graupe,1977).

However, there are perspectives that ANN models are not as accurate and effective until they are able to replicate the work of neurons as precise as possible. Kohonen (2001) argued that even though nature has its own way that no machines can replicate, the compliment is also true. Machines are able to do what humans brain cannot do as well. Human brains have limitation that are varied and prove to be a weakness at times. So it is not logical to copy in details unless it can guarantee full benefits.

1.0.5 Data visualization

As mentioned before, data visualization is the field of study that enables user to explore the properties of a dataset based on the person's own intuition and understand of domain knowledge, since it is meant to convey hidden information about the data to the user (Teh,2006). Data mining, data warehousing, retrieval systems and knowledge applications have widely increase the demand of the expansion of data visualization field. A huge advantage of data visualization over non-data visualization technique is it provides direct feedback to enhance the quality of data mining (Ankerst, 2000). To further explain the necessity and capability of data visualization, König (1998) explains more :

“To cope with today's flood of data from rapidly growing databases and related computational resources, especially to discover salient structures and interesting correlations in data, requires advanced methods of machine learning, pattern recognition, data analysis and visualizations. The remarkable abilities of the human observers to perceive clusters and correlations, and thus structure in data, is of great interest and can be well exploited by effective systems for data projection and interactive visualizations. “(König, 1998, p.55).

1.1 Problem statement

Data visualization is a field that has been supported by many respected approaches such as Principal Component Analysis (PCA) (Johnson & Wichern, 1992), Multidimensional Scaling (MDS) (Shepard & Carroll, 1965), Sammon's Mapping (Sammon, 1969), Principal Curves (Hastie & Stuetzle, 1989) and Principal Surfaces (LeBlanc & Tibshirani, 1994). PCA, MDS and Sammon's Mapping are considered to be not an effective way for data visualization since they demonstrate major disadvantages. For example, PCA loses certain useful information upon dimension reduction while MDS and Sammon's Mapping require heavy

computation which essentially leads to impractical methods. In addition, MDS and Sammon's Mapping cannot adapt to new data sample without first re-computing the existing data (Wu & Chow, 2005). This also loses its practicality.

Self-Organizing Maps (SOM) (Kohonen, 1984) is a data visualization focused unsupervised ANN method. SOM's visualization is able to preserve data topology from N-dimensional space to low-dimensional display space (Kohonen, 2001). However, it is not able to reveal the underlying structure of data as well as the inter-neuron distances from N-dimensional input space to low-dimensional output space (Yin, 2002).

In order to overcome these weaknesses, SOM proposed to new variants called ViSOM (Yin, 2002) and PRSOM (Wu & Chow, 2005) to preserve the inter-neuron distances and data structure in addition to data topology preservation. It is proven that ViSOM and PRSOM are better than PCA, MDS and Sammon's Mapping or even SOM itself in terms of inter-neuron distances, data topology and data structure preservation by Yin (2002) and Wu and Chow (2005) respectively. Also, they are discovered to be able to overcome the aforementioned limitations of PCA, MDS and Sammon's Mapping.

However, SOM faces problems in optimizing the accuracy of classification as compared against supervised classification focused ANN methods such as LVQ (Kohonen, 1988). Even its variants ViSOM and PRSOM are not able to perform data classification although they contributed a significant boost to data visualization.

Although the primary focus of this research is data visualization, but it can never stand on its own. Data visualization usually comes hand in hand with data classification. In order to correctly visualize data more accurately, the data has to be classified first so it can present the data in more presentable manner. If big datasets are presented without being classified first, it will show confusing data.

1.2 Research Objectives

The aim of this research is to propose SOM as a model of data visualization. The visualization ability is expected to be able to preserve data structure, inter-neuron distances and data topology from N-dimensional input space to low-dimensional output space.

1.3 Specific Objectives

- To study how SOM toolbox performs various methods of data visualization.
- To investigate the feasibility of the proposed method for data visualization.
- To test a developed SOM toolbox to visualize numerous benchmark datasets.
- To evaluate the efficiency of data visualization methods.

1.4 Research Scopes

The scope of this research is confined in developing an ANN method to support data visualization with computational efficiency. In depth, data structure, data topology and inter-neuron distances will be preserved as much as possible.

Empirical studies on the benchmark datasets will be conducted to demonstrate data visualization abilities using proposed method. Both qualitative and quantitative comparisons will be conducted.

1.5 Research Methodologies

The research kicks off with thorough literature reviews of the recent data visualization methods such as PCA, Sammon's Mapping(SM) and SOM in an effort to clearly state and compare the strengths and limitations of each method. A literature review is also conducted in order to define a way of enhancing a selected supervised ANN with adequate data visualization abilities.

The proposed methods are implemented using MATLAB 7.0 software packages (Education version) and SOM toolbox. The empirical studies on the benchmark datasets are conducted to evaluate the performance of selected method. Iris flower dataset will first be tested and followed by other multi-dimensional datasets such as wine and car characteristics.

CHAPTER 2

2.0 LITERATURE REVIEW

2.1 Introduction

Data visualization is a field being explored rapidly by researchers in this modern society. Various approaches have been introduced along the line. In this chapter, methods for data visualization will be discussed and the algorithms involved, highlighting their strengths and weaknesses.

This chapter discusses the techniques involved, specifically dimension reduction, Sammon's Mapping and Self-Organizing Maps (SOM) These are all related techniques for data visualization which will also be discussed in this chapter. Their performances will be highlighted for each of the techniques.

2.2 Dimension Reduction

The objective of dimension reduction is to represent data of higher dimensions usually more than three, in a much lower dimensional space with minimum loss of information. Humans are unable to view data if the dimension is higher than three. So data visualization aims to aid humans in better understanding the data and to make use of them for good use. The methods that will be used and discussed in later sections are Sammon's Mapping (Sammon,1969) and SOM (Kohonen,1984).

2.3 Data visualization

Visualization is the transformation of the symbolic into the geometric. There are many reasons why visualizations are created such as to answer questions, make decisions, seeing data in a clearer context, expanding memory, finding patterns and to name a few. To serve

the purpose, visualization has to be constructed obeying the design principles as proposed by Mackinlay (1986). Those principles are ;

- 1) Expressiveness - a set of data is considered expressible in visual if it is also expressible in the sentence of the language including all the facts in the data.
- 2) Effectiveness - a visualization is considered effective if the information portrayed by the visualization is more readily perceived than the information provided by any other form of visualization.

Data visualization methods represent the underlying structures of data and the multivariate relationship among the data samples, which are in researcher's interest for decades. In order to achieve this, several methods have been developed to perform quick visualization of simple data summaries for low-dimensional datasets. As an example, the so-called five number summaries (Tukey, 1977) which consists of highest and lowest value, median and the two quartiles of data, which can be drawn to visualize the summary of low-dimensional datasets. However, as the number of dimensions increase, their ability of visualizing data summary decreases (Kaski, 1997). From here on afterwards, visualization methods will be discussed briefly for high-dimensional data.

2.4 Visualization of high-dimensional data items

To help visualize data better, numerical approaches related to graphical representation have been used for this purpose. Tan, Steinbach and Kumar (2006) have discussed the similarities between them for high-dimensional data. Parallel Coordinate approach is one them related to data visualization. It is a very simple form of data visualization which provides one coordinate axis for each attributes spaced parallel in the x-axis and the corresponding value is plotted by drawing lines connecting the parallel attributes in the y-axis

for each data sample. If the connections are correctly arranged, the patterns will form and produce good pattern. However, relationships between data samples are not visible. Vice versa, if the data are not properly arranged in order, it may cause confusing data visualization (Tan et al.,2006).

Andrews (1972) proposed a visualization technique where one curve for each data item is obtained by using the components of the data vectors as coefficients of orthogonal sinusoids (Kaski,1997). Similar to Parallel Coordinate, it is also greatly affected by the ordering the data attributes.

2.5 Data visualization in the past and current

Data visualization dates back to the 2nd century AD where rapid development has occurred during that period. The earliest table ever recorded was created in Egypt where it was used to organize astronomical information as a tool for navigation. In general, a table is a textual representation of data but contains visual attributes that help humans recognize better. According to Few (2007), graphs didn't exist until 17th century where it was invented by the infamous mathematician Rene Descartes who is also known for his philosophical thoughts.

In contemporary use, data visualization has shown its importance in business intelligence but still being ignored by a fraction of business people. It is still a misunderstood concept where people take it for granted and to be taken lightly. However, there are still good news related to the advancement of data visualization especially in academic fields. In university as an example, all courses require data visualization as part of explanation in any matter at all. Due to this rapid advancements, many universities nowadays have given their attentions on visualization and a few have excellent programs that serve the needs of many graduate students whom may require it for their researches and prototype application. There also exist bad trends in data visualization in current days especially when people rush to embrace the

technology and fail to fully understand this matter without first studying it. Data visualization is captivating to the eyes and intrigue people to try and use them. This is especially bad in field of advertisements when they try to implement visualizations technique without considering the design principle. Many people are yet to grasp the idea of data visualizations fully. While many may be under the impressions that data visualization is about making things pretty and cute, truth is it is about dressing up presentations to tackle the audience. It doesn't necessarily have to be colorful just as long as it catches the attention of the users it is already considered a success (Few, 2007).

2.6 Principal Component Analysis (PCA)

PCA (Hotelling, 1933) is a linear data analysis method to find the orthogonal principal directions in a dataset, along which the dataset shows the largest variances (Yin, 2002). It displays data in a linear projection on such a subspace of the original data space that best preserves the data variances (Kaski, 1997). By selecting the major components, the PCA is able to effectively reduce the data dimensions and display the data in terms of selected and reduced dimensions in low-dimensional space. As described by Yin (2002), PCA is an optimal linear projection of the mean square error between original data points and the projected one, as shown in Eq(1).

$$\min_{\mathbf{x}} \left[\mathbf{x} - \sum_{j=1}^m (q_j^T \mathbf{x}) q_j \right]^2 \quad \text{Eq(1)}$$

Here, $\mathbf{x}=[x_1, x_2, x_3 \dots x_8]$ is the N-dimensional input vector and $\{q_j, j=1,2,3,\dots,m, m \leq N\}$ are the first m principal eigenvectors of the data and the term q_j^T represents the projection of \mathbf{x} onto the j^{th} principal dimension.

However, linear PCA loses certain useful information about the data while dealing with data having much higher dimension than two (Yin,2002). Linear PCA is not able to display the non-linearity of data, since it displays data in a linear subspace. For high-dimensional dataset, linear visualization methods may have difficulty to represent the data well in low-dimensional spaces (Kaski, 1997). Thus several approaches have been proposed to produce non-linear low-dimensional display of high-dimensional data.

2.7 Sammon's Mapping (SM)

Sammon's Mapping (Sammon, 1969) is a non-linear data projection method usually associated with multidimensional scaling (MDS). It preserves the pair wise distances between two data points in N -dimensional space to low-dimensional display space. The advantage of Sammon's Mapping is that the errors are divided by the distances in the N -dimensional space, which emphasizes the small distances (Kohonen, 2001). With δ_{ij} representing the distance between two points i and j in N -dimensional data space and d_{ij} representing their corresponding distances in low-dimensional display space, the cost function of Sammon's Mapping can be read according to Kohonen (2011) as :

$$Es(t) = \sum_{i \neq j} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}} \quad \text{Eq(2)}$$

As in MDS, Sammon's Mapping performs dimensional reduction by displaying multidimensional data in a low-dimensional nonlinear display space. However, both MDS and Sammon's Mapping display weakness in providing explicit mapping function to accommodate new data samples in the visualization without re-computing the existing data (Mao & Jain, 1995). Sammon's Mapping requires high computational complexity which causes impracticality when using Sammon's Mapping in multiple occasions. Although