

AIN SHAMS UNIVERSITY  
Faculty of Computer  
& Information Sciences  
Computer Science Department



# **A STUDY ON VISUALIZATION ALGORITHMS FOR DATA MINING**

A Thesis Submitted to Computer Science  
Department, Faculty of Computer &  
Information Sciences, Ain Shams University

In partial fulfillment of the requirements for  
Master of Science Degree

By

**Emad Monier Mosaad Ibrahim**

B.Sc. in Computer Science, 2000.  
Demonstrator, Computer Science Department,  
Faculty of Computer & Information Sciences,  
Ain Shams University, Cairo, Egypt.

Under Supervision of

**Prof. Dr. Abdel-Badeeh M. Salem**

Professor of Computer Science, Computer Science Department,  
Faculty of Computer & Information Sciences,  
Ain shams University, Cairo, Egypt.

**Dr. Khaled Ahmed Nagaty**

Lecturer, Computer Science Department, Faculty of Computer &  
Information Sciences,  
Ain shams University, Cairo, Egypt.

2004

# **Acknowledgement**

First and foremost, I humbly give my deep thanks to God for giving me the opportunity and the strength to accomplish this work.

I would like to express my deep appreciation and thanks to Prof. Dr. Mohamed Said Abdel-Wahab, our dean, for his great support and encouragement during the execution of this work.

I would like to express my greatest gratitude to Prof. Dr. M. Siam the head of Computer Science department and Prof. Dr. Ali Elnaiem the head of the basic sciences department for their support and indispensable advices during the progress of the work.

I would like to express my deep appreciation to Prof. Dr. Abdel-badeeh Mohammed Salem for his full supervision of this work and to his planning, on going advises, guidance and for his comprehensive help in the interpretation of the outcome results by valuable suggestions in writing this thesis.

I am indebted to Dr. Khaled Ahmed Nagaty for his valuable and helpful guidance in reviewing and directing the work presented in this thesis.

I would like to express my deep thanks to my family and my colleague friends, Safaa El-Said, Abeer Mahmoud and Mohamed Abdel-Megeed, those who were encouraging me strongly during the execution of this work.

# Publications

1. Abdel-Badeeh M.Salem, Khaled A. Nagaty and Emad Monier, “***Enhancement of Cluster Visualization in Self-Organizing Map***”, Proceedings of International Conference on Soft Computing, Mendel, Brno, Czech Republic, pp. 183-187, June 4-6, 2003.
2. Abdel-Badeeh M.Salem, Khaled A. Nagaty and Emad Monier, “***Free Projection SOM: A New Method for SOM-Based Cluster Visualization***”, Proceedings of WSEAS 2<sup>nd</sup> international Conference on Artificial Intelligence, Knowledge Engineering, and Data Bases (AIKED 2003), Crete Island, Greece, pp. 128-132, August 11-13, 2003.

Also accepted for publication in The Third IASTED International Conference on Artificial Intelligence and Applications AIA 2003, September 8-10, 2003, Benalmadena, Spain.

3. Abdel-Badeeh M.Salem, Khaled A. Nagaty and Emad Monier, “***New Techniques for Enhancement of SOM-Based Cluster Visualization***”, The Official Journal of the Egyptian Computer Society, Vol. 25. No. 1, pp. 26-38, January 2003.

Also accepted for publication in Journal of Institute of Mathematics & Computer Sciences (Comp. Sc. Series), Vol. 14. No.2, December 2003, Calcutta-700 007, India.

## **Abstract**

In recent years, the need to extract knowledge automatically from large databases has grown increasingly acute. In response, the closely related fields of knowledge discovery in database (KDD) and data mining have developed processes and algorithms that attempt to intelligently extract interesting and useful information (i.e., knowledge) from vast amounts of raw data. Such techniques are used in various application domains, ranging from department stores to catalogs of stellar objects.

One of the techniques used in data mining is visualization, which may be used at the beginning of the data mining step to obtain a rough feeling of the clusters and structures in the data. The visualization process consists of two main phases: vector quantization; which tries to find a smaller but still representative set for the original data set, and vector projection; which deals with the problem of finding a representation of the data in a human-perceptible dimension, i.e. 2- or 3-dimensional display for the data. So the visualization can be obtained by applying one of the vector quantization algorithms and then a vector projection algorithm.

Self-Organizing Map (SOM) is a neural network algorithm based on unsupervised learning. It is usually used for data visualization because it performs both the vector quantization and projection together. The projection implemented by the SOM is done on a regular 2- or 3- dimensional grid. Many techniques have been developed to visualize the shape of the data on the map grid.

In this study, two new techniques for the enhancement of the visualization of the SOM have been developed. The first technique is referred to as Segmented Distance Matrix (SDM)

depends on the enhancement of the distance matrix visualization by performing a segmentation process on it to reduce the number of gray levels used for visualization. This causes a clear cluster assignment for the SOM vectors. The SDM was tested on artificial and real world data sets and it has been shown that it gives qualitative information about the clusters present in the data and intuitive assignment of clusters within the input data.

The second technique is referred to as Free Projection Self-Organizing Map (FP-SOM). It is an extension to the standard SOM algorithm in order to enhance the projection implemented by the SOM so that the clusters and possibly the structures present in the data could be visualized efficiently and without need to run other computationally expensive algorithms such as Sammon's projection algorithm. FP-SOM algorithm has been tested on artificial and real world data sets of varying sizes and complexity. It has been shown that the projection error of the standard SOM can be reduced by values that range between 10% and 36% depending on the size and complexity of the data set and the time complexity of the standard SOM is not affected by the modifications of the FP-SOM. Both techniques do not affect the basic training process or the parameters of the standard SOM, because they work as extensions to the basic SOM algorithm. They can be used together as a powerful tool for high dimensional and large data visualization.

# Table of Contents

Acknowledgment .....	II
Publications.....	III
Abstract.....	IV
Table of Contents.....	VI
List of Figures.....	VIII
List of Tables.....	XII
<b>CHAPTER</b>	<b>Page</b>
<b>1 Introduction.....</b>	<b>2</b>
1.1 Problem Overview.....	2
1.2 Thesis Motivation.....	5
1.3 Thesis Objective.....	5
1.4 Thesis Organization.....	6
<b>2 Data Mining and Knowledge Discovery in Databases.....</b>	<b>9</b>
2.1 Introduction. ....	9
2.2 The KDD Process.....	9
2.3 Data Mining Step.....	12
2.4 Data Mining Versus Query Tools.....	15
2.5 Data Mining Tasks.....	16
2.6 Data Mining Techniques.....	19
2.7 Complex Data for Mining. ....	22
2.8 Data Mining Applications.....	23
2.9 Summary.....	25
<b>3 Visualization Algorithms and Techniques..</b>	<b>27</b>
3.1 Introduction.....	27
3.2 Visualization Process.....	27
3.3 The Self-Organizing Map (SOM).....	30
3.3.1 Basic Algorithm.....	32

3.3.2	Variants of SOM.....	35
3.3.3	Related Algorithms.....	37
3.3.4	Data Analysis Using SOM .....	40
3.3.5	Benefits and Pitfalls.....	42
3.3.6	Scalability.....	44
3.3.7	Using SOM in Data Mining .....	46
3.4	SOM-Based Data Visualization .....	46
3.4.1	Cluster Structure and Shape .....	46
3.4.2	Components of the Map .....	48
3.4.3	Data on the Map .....	51
3.5	Vector Quantization and Projection (VQ-P) Algorithms.....	53
3.5.1	Vector Quanization.....	53
3.5.2	Vector Projection.....	60
3.6	Differences Between SOM and VQ-P.....	69
3.7	Evaluation of Data Visualization Techniques..	70
3.7.1	Proposed Characteristics of an Evaluation Environment.....	70
3.7.2	Basic Testing Criteria and Measures...	72
3.8	Summary.....	74
<b>4</b>	<b>Enhancement of SOM Distance Matrix Visualization .....</b>	<b>76</b>
4.1	Introduction.....	76
4.2	Distance Matrices (DM) .....	77
4.3	Segmentation of the Distance Matrix (SDM)...	78
4.4	Experimental Results.....	79
4.4.1	The Animal Data Set.....	79
4.4.2	Auto Mpg Data Set .....	82
4.4.3	Breast Cancer Data Set.....	85
4.5	Discussion.....	88
4.6	Summary.....	90
<b>5</b>	<b>Free Projection Self-Organizing Map.....</b>	<b>92</b>

5.1	Introduction.....	92
5.2	The Projection of the Standard SOM.....	92
5.3	The Free Projection Self-Organizing Map (FP-SOM).....	94
5.3.1	Motivation of FP-SOM.....	95
5.3.2	The FP-SOM Algorithm.....	95
5.4	Experimental Results.....	97
5.4.1	Animal Data Set.....	97
5.4.2	Iris Flower Data Set .....	103
5.4.3	Auto Mpg Data Set.....	108
5.4.4	Breast Cancer Data Set.....	113
5.4.5	Thyroid Disease Data Set.....	120
5.5	Discussion.....	126
5.6	Summary.....	131
<b>6</b>	<b>Conclusions and Future Work.....</b>	<b>134</b>
	<b>References.....</b>	<b>140</b>
	<b>Appendix.....</b>	<b>A-1</b>

## List of Figures

<b>Figure 2.1</b>	KDD is a multi-disciplinary field.....	10
<b>Figure 2.2</b>	Knowledge discovery process.....	11
<b>Figure 2.3</b>	Architecture of a typical data mining system.....	14
<b>Figure 3.1</b>	Visualization process.....	29
<b>Figure 3.2</b>	Linking small multiples by similar position or color....	31
<b>Figure 3.3</b>	Neighborhood sets (at radius 0,1, and 2) of the center most units: (a) hexagonal, (b) rectangular lattice.....	32
<b>Figure 3.4</b>	Updating the best matching unit (BMU) and its neighbors towards the input sample marked with x.....	34
<b>Figure 3.5</b>	Summary of SOM algorithm.....	34
<b>Figure 3.6</b>	Data analysis using SOM as intermediate step.....	41

<b>Figure 3.7</b>	One step of SOM algorithm as C-code.....	44
<b>Figure 3.8</b>	Clustering visualizations: (a) U-matrix with values indicated using grayscale, (b) distance matrix using hexagon size to show the values, (c) similarity coloring, (d) the map network in 3-dimensional space. The colors in (c) and (d) match each other.....	50
<b>Figure 3.9</b>	U-matrix (top left) and three component planes. The different plots are linked together using similar position. It can be seen, for example, that high "X-coord" values are typical for the cluster on the bottom of the map.....	51
<b>Figure 3.10</b>	Visualization of the response of the map to two data samples.....	53
<b>Figure 3.11</b>	The $k$ -means quantization algorithm.....	55
<b>Figure 3.12</b>	Movement of estimated cluster centers $m_1$ and $m_2$ to the actual cluster centers.....	56
<b>Figure 3.13</b>	The $k$ -means applied on the same data in Fig 3.12 but with $k=3$ .....	56
<b>Figure 3.14</b>	The Fuzzy- $k$ -means algorithm.....	57
<b>Figure 3.15</b>	The sequential $k$ -means algorithm.....	58
<b>Figure 3.16</b>	The modified sequential $k$ -means algorithm.....	58
<b>Figure 3.17</b>	The adaptive $k$ -means algorithm.....	60
<b>Figure 3.18</b>	PCA projection of the statistical indicators of 77 countries.....	62
<b>Figure 3.19</b>	Sammon's mapping of the same data set that was projected using the PCA in Fig. 3.18.....	67
<b>Figure 3.20</b>	A nonlinear projection constructed using nonmetric MDS. The data set is the same as in Fig. 3.19 and Fig. 3.18.....	68
<b>Figure 3.21</b>	Evaluation environment for visualization.....	71
<b>Figure 4.1</b>	Summary of the SDM visualization method.....	79
<b>Figure 4.2</b>	12×12 SOM trained on animal data set (a) Standard representation (b) standard average distance matrix representation (c) Segmented distance matrix with 3 levels (d) Segmented distance matrix with 4 levels.....	82
<b>Figure 4.3</b>	Standard average DM for 15×15 SOM trained on auto Mpg data.....	84

<b>Figure 4.4</b>	SDM with 3 levels for 15×15 SOM trained on auto Mpg data.....	84
<b>Figure 4.5</b>	SDM with 4 levels for 15×15 SOM trained on auto Mpg data.....	85
<b>Figure 4.6</b>	Standard average DM representation for 20×20 SOM trained on breast cancer data.....	86
<b>Figure 4.7</b>	3 levels SDM representation for 20×20 SOM trained on breast cancer data.....	87
<b>Figure 4.8</b>	4 levels SDM representation for 20×20 SOM trained on breast cancer data.....	88
<b>Figure 5.1</b>	The inaccuracy of the projection of the standard SOM representation.....	94
<b>Figure 5.2</b>	Summary of FP-SOM algorithm.....	97
<b>Figure 5.3</b>	Standard representation of 8×8 SOM trained on Animal data set.....	99
<b>Figure 5.4</b>	Sammon's projection of SOM prototype vectors in shown in Fig.5.3.....	99
<b>Figure 5.5</b>	FP-SOM Representation of 8×8 SOM trained on Animal data set.....	100
<b>Figure 5.6</b>	Sammon's mapping of the original Animal data.....	100
<b>Figure 5.7</b>	Projection errors of SOM, FP-SOM and SAM-SOM visualizations of Animal data for different map sizes...	101
<b>Figure 5.8</b>	Time in seconds for obtaining standard SOM, FP-SOM and SAM-SOM visualizations of Animal data for different map sizes.....	102
<b>Figure 5.9</b>	Standard representation of 1212 SOM trained on Iris data set.....	104
<b>Figure 5.10</b>	Sammon's projection of SOM prototype vectors in shown in Fig.5.9.....	104
<b>Figure 5.11</b>	FP-SOM Representation of 12×12 SOM trained on Iris data set.....	105
<b>Figure 5.12</b>	Sammon's mapping of the original Iris data.....	105
<b>Figure 5.13</b>	VQ-P of the Iris data.....	106
<b>Figure 5.14</b>	Projection errors of SOM, FP-SOM and SAM-SOM visualizations of Iris data, for different map sizes.....	107
<b>Figure 5.15</b>	Time in seconds for obtaining standard SOM, FP-SOM and SAM-SOM visualizations of Iris data for different map sizes.....	108
<b>Figure 5.16</b>	Standard representation of 18×18 SOM trained on Auto Mpg data set.....	109

<b>Figure 5.17</b>	Sammon's projection of SOM prototype vectors in shown in Fig.5.16.....	110
<b>Figure 5.18</b>	FP-SOM Representation of 18×18 SOM trained on Auto Mpg data set.....	110
<b>Figure 5.19</b>	Sammon's mapping of the original Auto Mpg	111
<b>Figure 5.20</b>	VQ-P of the Auto Mpg data.....	111
<b>Figure 5.21</b>	Projection errors of SOM, FP-SOM and SAM-SOM visualizations of Auto Mpg data for different map sizes.....	112
<b>Figure 5.22</b>	Time in seconds for obtaining standard SOM, FP-SOM and SAM-SOM visualizations of Auto Mpg data for different map sizes.....	113
<b>Figure 5.23</b>	Standard representation of 20×20 SOM trained on Cancer data set.....	115
<b>Figure 5.24</b>	Sammon's projection of SOM prototype vectors in shown in Fig.5.23.....	116
<b>Figure 5.25</b>	FP-SOM Representation of 20×20 SOM trained on Cancer data set.....	117
<b>Figure 5.26</b>	Sammon's mapping of the original Cancer data.....	118
<b>Figure 5.27</b>	VQ-P of the Cancer data.....	118
<b>Figure 5.28</b>	Projection errors of SOM, FP-SOM and SAM-SOM visualizations of Cancer data for different map sizes...	119
<b>Figure 5.29</b>	Time in seconds for obtaining standard SOM, FP-SOM and SAM-SOM visualizations of Cancer data for different map sizes.....	120
<b>Figure 5.30</b>	Standard representation of 60×60 SOM trained on Thyroid data set.....	122
<b>Figure 5.31</b>	Sammon's projection of SOM prototype vectors in shown in Fig.5.30.....	123
<b>Figure 5.32</b>	FP-SOM Representation of 60×60 SOM trained on Thyroid data set. ....	124
<b>Figure 5.33</b>	Sammon's mapping of the original Thyroid data.	125
<b>Figure 5.34</b>	VQ-P of the Thyroid data.....	126
<b>Figure 5.35</b>	Comparison between SOM, FP-SOM and SAM-SOM applied on the 5 data sets with respect to the projection error.....	129
<b>Figure 5.36</b>	Comparison between FP-SOM and SAM-SOM applied on the 5 data sets with respect to the percentage of the increase in time of obtaining both visualization with respect to time of training of SOM...	130

<b>Figure A.1</b>	Block diagram of the visualization package.....	A-2
<b>Figure A.2</b>	Entering parameters of the SOM.....	A-3
<b>Figure A.3</b>	The SOM after training on IRIS data.....	A-4
<b>Figure A.4</b>	Entering the type of Distance Matrix.....	A-4
<b>Figure A.5</b>	After computing the Distance Matrix.....	A-5
<b>Figure A.6</b>	Entering the number of segmentation levels of Distance Matrix.....	A-5
<b>Figure A.7</b>	After segmentation of the Distance Matrix.....	A-6
<b>Figure A.8</b>	Zooming and other view control tools in the visualization package.....	A-6
<b>Figure A.9</b>	Previous visualizations of the data in addition to the FP-SOM visualization.....	A-7
<b>Figure A.10</b>	Previous visualizations in addition to SAM-SOM visualization.....	A-7
<b>Figure A.11</b>	Previous visualizations in addition to Sammon's Projection of the data.....	A-8
<b>Figure A.12</b>	Three-dimensional viewing tools.....	A-8

## List of Tables

<b>Table 3.1</b>	Some vector quantization algorithms.....	54
<b>Table 3.2</b>	Some vector projection algorithms.....	62
<b>Table 4.1</b>	The Animal data set.....	80
<b>Table 4.2</b>	Attribute information of the Auto Mpg data set.....	83
<b>Table 5.1</b>	Summary of simulation results of different visualizations for Animal data.....	101
<b>Table 5.2</b>	Summary of simulation results of different visualizations of Iris data.....	108
<b>Table 5.3</b>	Summary of simulation results of different visualizations for Auto Mpg data.....	113
<b>Table 5.4</b>	Summary of simulation results of different visualizations for Breast Cancer data.....	120
<b>Table 5.5</b>	Summary of simulation results of different visualizations for Thyroid disease data.....	126
<b>Table 5.6</b>	Summary of the results obtained from the visualization of the 5 data sets.....	128

# **Chapter 1**

## **Introduction**

## *Chapter 1*

# **Introduction**

### **1.1 Problem Overview**

Data mining is an emerging area of new research efforts, responding to the presence of large databases in commerce, industry and research. It is also a title for a large number of widely divergent methods ranging from belief networks and relational learning to statistics and neural networks. Data mining is part of a larger framework, Knowledge Discovery in Databases (KDD) [1], whose purpose is to find new knowledge from databases where dimension, complexity or amount of data is prohibitively large for human observation alone. Data mining is an iterative process requiring that the intuition and background knowledge of humans be coupled with the computational efficiency of modern computer technology. For this reason, visualization is a very important part of data mining.

Visualization techniques are very useful methods of discovering patterns in data sets, and may be used at the beginning of data mining processes to get a rough feeling of the quality of the data set and where patterns are to be found. Data mining can also be thought of as a preprocessor for visualization. For example in prediction, one of the data mining tasks, the goal of data mining is to select the dimensions (variables or attributes) relevant to accurately predicting one variable based on others [2]. It is hence possible for an algorithm to look for a few dimensions out of the possibly hundreds or thousands necessary to determine a predicted variable. As such, a data mining algorithm can be viewed as a preprocessor for the visualization engine, discarding the bulk of

irrelevant dimensions and avoiding combinatorial number of subsets of dimensions that could be considered, to focus the analyst's attention on the few relevant variables.

The visualization process consists of two phases: vector quantization and vector projection. Vector quantization reduces the original data set to a smaller, but still representative set to work with. At the same time it suppresses noise. Some vector quantization algorithms are:  $k$ -means [3],  $c$ -means [4,5], maximum entropy [6,7], neural gas [6] and SOM [7]. In order to visualize multidimensional vectors, a projection from the original high-dimensional input space to at most 3-dimensional output space has to be found. Some algorithms of vector projection: multi-dimensional scaling (MDS) [8,9], Sammon's nonlinear mapping [10], Curvilinear component analysis (CCA) [11] and SOM.

Self-Organized Map (SOM) is a neural network algorithm based on unsupervised learning. The SOM has proven to be a valuable tool in data mining and KDD with applications in full-text and financial data analysis. It has also been successfully applied in various engineering applications in pattern recognition, image analysis, process monitoring, and fault diagnosis [12,13,14]. SOM has several beneficial features, which make it a useful method in data mining. It implements an ordered dimensionality-reducing to the mapping of the training data. The map follows the probability density function of the data and is robust to missing data. It is readily explainable, simple and – perhaps most importantly – easy to visualize. Visualization of complex multidimensional data is indeed one of the main application areas of the SOM. It is clear that SOM is an example of a method that accomplishes both vector quantization and projection. Other methods can be constructed by first making vector quantization and then vector projection using the previously mentioned algorithms. The SOM differs from a serial combination of a vector quantization and vector