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 2nd 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 attempts 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 form 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-dimenional 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

List of Figures

List of Tables

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