



*FACULTY OF COMPUTER
& INFORMATION SCIENCES
AIN SHAMS UNIVERSITY
Abbassia, Cairo, Egypt*

KNOWLEDGE DISCOVERY WITH ARTIFICIAL NEURAL NETWORKS

A Thesis submitted to the Department of Computer
Science, Faculty of Computer and Information Sciences,
Ain Shams University

In partial fulfillment of the requirements for the degree
of Master in Computer and Information Sciences

BY

Ayad Fekry Ayad

B.Sc. in Computer Science
Demonstrator, Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University

Under the Supervision of

Prof. Dr. Abdel-Badeeh Mohamed Salem

Prof. Of Computer Science, Computer Science Department,
Faculty of Computer and Information Sciences
Ain Shams University

Prof. Dr. Mostafa Mahmoud Syiam

Professor and Head of Computer Science Department,
Vice Dean for Student affairs
Faculty of Computer and Information Sciences
Ain Shams University

2004

Acknowledgments

My utmost thanks to ALLAH for all his gifts in all my life.

My gratitude should go to Prof. Mohammed Said Abdel-Wahab, Dean of the faculty of Computer and Information Sciences, for his valuable support.

My greatest thanks go to Prof. Abdel-Badeeh Salem and Prof. Mostafa Syiam for their teaching, supporting, guiding and encouragement they provide me throughout my work. Their comments and suggestions helped me not only in this thesis, but also in my career. They have suffered a lot with me since it is my first step in research studies. They taught me how to read and write a paper and how to conduct experiments and analyze the results.

I would like to thank my fathers Kerolous and Marqus for helping and supporting me to complete this thesis.

Special Thanks to Mr. Tamer Mustafa who helped me during writing this thesis.

Thanks are not enough to be given to my parents. Rather, there are no words that are valuable to equalize their love and support for me. They are behind any success in my life.

Table of Contents

	Page
1 Introduction	1
1.1 Motivations & Objectives.....	5
1.2 Contribution	6
1.3 Thesis Organization	7
2 Knowledge Discovery And Data Mining	9
2.1 Data Mining and Information Age: Emerging Quests	9
2.2 Defining Knowledge Discovery	10
2.3 Architectures Of Knowledge Discovery..	13
2.4 Business Understanding	17
2.5 Data Understanding	20
2.6 Data Preparation	22
2.6.1 Data Cleaning	23
2.6.2 New Features	25
2.6.3 Transformations	26
2.6.4 Prepared Information Environment	27
2.7 Modeling	28
.....	28
2.7.1 Modeling Tools	34
.....	35
2.7.2 Model Assessment	36
.....	
2.8 Evaluation	37
.....	41

2.9	Deployment	43
	47
2.10	Fundamental Issues In Knowledge	49
	Discovery.....	49
	51
2.11	Tasks Of Data Mining	54
	58
2.12	Examples Of Knowledge Discovery Systems	64
		66
2.13	Summary	68
	68
		70
3	Self-Organizing Map (SOM)	71
3.1	SOM Structure	72
	78
3.2	Basic SOM Training Algorithm	
	80
3.3	Mathematical Treatment and Properties	
3.4	SOM Visualization.....	81
3.5	Variants of SOM	81
3.6	Related Algorithms to SOM.....	
3.7	Data Analysis Using SOM	83
	3.7.1 Quantization	85
	3.7.2 Projection	
	3.7.3 Benefits and Pitfalls	89
3.8	Using SOM in Data Mining	
3.9	Summary	94
		95

4	The Proposed Modification To Traditional SOM Training Algorithm	100
4.1	Feature Maps	102
	
4.2	Kohonen Self-Organizing Feature Maps	102
4.3	Disadvantages Of The Basic SOM	105
	Training Algorithm	108
4.4	The K-Means Algorithm	110
	115
4.5	The Proposed K-Means Initialization For SOM Training Algorithm	117
		118
4.6	Complexity Analysis Of The Proposed Approach.....	119
	120
	120
4.7	Experimental Results & Discussion.....	122
4.8	Summary	122
	124
5	The Proposed Growing Hierarchical SOM (GHSOM) For Document Clustering	125
		126
5.1	Introduction	
	129
5.2	Overview On Dynamic Neural Network	134

	Models.....	
	148
5.3	The Growing Hierarchical SOM (GHSOM)	148
5.3.1	The principles	159
	167
5.3.2	GHSOM Training Algorithm	
5.3.3	Analysis Of GHSOM Characteristics	
5.4	Data Set	
	
5.4.1	Document Preprocessing	
	
5.4.2	Generating Characteristic Document Vectors.....	
	
5.5	Experimental Results & Discussion	
5.5.1	A SOM Of ICICIS Abstracts Collection.....	
	
5.5.2	A GHSOM Of ICICIS Abstracts	

Collection.....

.....

5.5.2.1 D
e
e
p

H
i
e
r
a
r
c
h
y

...

...

.

.

5.5.2.2 S
h
a
l
l
o
w

H
i
e
r
a
r
c
h
y

...

.

5.5.3 Comparison Of
SOM and GHSOM
Representation

5.6 Summary
.....

6 Summary, Conclusions And Future Work

References

Appendix A : Implementation Code In C++

A.1 SOM Training Algorithms

A.2 Proposed Modification to SOM Training
Algorithms

A.3 GHSOM Training Algorithms

List Of Figures

Figure		Page
2.1	A general scheme of knowledge discovery	13
2.2	Main functional phases of the knowledge discovery process	14
2.3	(a) CRISP-DM: process model for data mining (b) Brachman's KDD process and (c) Pyle's model-building outline. Both of the latter process models have been simplified	17
2.4	Relation of domain knowledge, data and the problem.....	18
2.5	PIE in information flow	27
3.1	Neighborhoods (size 1, 2 and 3) of the unit marked with black	50
3.2	Updating the best matching unit (BMU) and its neighbors.....	52
3.3	The two basic neighborhood functions	53
3.4	(a) A topologically good, (b) A folded 1-dimensional SOM	57
3.5	Component planes representation of a SOM in 2D (a) and 3D (b)	59
3.6	U-matrix presentations of a SOM: shades of gray (a) and 3D mesh (b)	60
3.7	Sammon's projection of a SOM	61
3.8	Measurements from a computer system from two	

	days depicted as a data set histogram	62
3.9	Trajectory plotted on top of the u-matrix of a SOM. The arrows show the consecutive BMUs	63
3.10	Visualization of quantization errors of two input vectors.....	63
3.11	Data analysis using SOM as an intermediate step	68
3.12	Using SOM in data mining	73
4.1	Pattern classes that are evident by proximity ...	85
4.2	An example of the procedure involved in the first step.....	91
4.3	An example of the procedure involved in the second step.....	92
4.4	The assignment of layers In an N x N network	93
4.5	The whole procedure to arrange N ² cluster centers	93
4.6	810 2-D data samples generated using DDA line algorithm	95
4.7	The resultant feature maps constructed by the two methods for 810 data set: left column is method 1 (conventional), and right column is method 2 (proposed).....	97
4.8	The resultant calibrated maps constructed by the two methods for the iris data set: left column is method 1 (conventional), and right column is method 2 (proposed).....	98
4.9	Avg.quantization error by the two methods for 810 data samples	100
4.10	Avg.quantization error by the two methods for iris data samples	100
5.1	Architecture of a trained GHSOM	110

.....

5.2	Insertion of units: A row (a) or a column	
	(b) of units (shaded gray)	114
5.3	Document preprocessing and encoding	119
5.4	5 x 6 SOM of the ICICIS conference	121
5.5	Top and second level maps	
	(a) Layer 1 map: 4x3 units; Main topics	
	(b) Layer 2 map: 2x2 units; Knowledge discovery	123
5.6	Layer 1 map: 5x4 units (shallow hierarchy)	124

List Of Tables

Table	Page
4.1 The looking up table. Only upper triangle need to be computed	93
4.2 Learning parameters of the first data set	96
4.3 Learning parameters of the second data set	99

List of Publications

1. Abdel-Badeeh M. Salem, Mostafa M. Syiam, and Ayad F. Ayad “***Improving Self-Organizing Feature Map (SOFM) Training Algorithm Using K-means initialization***” In Proc. 5th International Conference on Enterprise Information Systems ICEIS, vol.1, 2003, pp.399-405, France.

Published also in Proc. 7th IEEE International Conference On Intelligent Engineering Systems (INES), vol.40, 2003, pp.41-46. Egypt.

2. Abdel-Badeeh M. Salem, Mostafa M. Syiam, and Ayad F. Ayad “***A Hybrid Dynamic Self-Organizing Map For Clustering Of Document Collections***” In Proc. 2nd WSEAS Intern. Conf. on ARTIFICIAL INTELLIGENCE, KNOWLEDGE ENGINEERING, and DATA BASES (AIKED 2003), vol.2, 2003, pp.201-206, Greece.
3. Abdel-Badeeh M. Salem, Mostafa M. Syiam, and Ayad F. Ayad “***Unsupervised Artificial Neural Networks For Clustering Of Document Collections***” The 17th International FLAIRS Conference, 2004, USA.

Accepted also for publication in 6th International Conference on Enterprise Information Systems ICEIS, 2004, France.

Abstract

Data mining is a part of a large area of recent research in artificial intelligence and information management. The purpose is to find new knowledge from databases where the dimensionality, complexity, or amount of data is prohibitively large for manual analysis. A large data set may be of very high dimensionality and consists of complex structure that even the most well planned data mining techniques might have difficulty extracting meaningful patterns from it. An unsupervised technique such as cluster detection becomes useful in such situations.

Clustering is a useful tool when it is used to deal with a large complex data set with many variables and unknown internal structure. In such situations, clustering would be the best tool to obtain an initial understanding of the structure inherent in the data. Once automatic cluster detection has discovered regions of the data space that contains similar records, other data mining tools and techniques could be used to discover rules and patterns within the clusters.

The Self-Organizing Map (SOM) has been used as a tool for mapping high-dimensional data into a one- (or two-) dimensional feature map. It is then possible to visually identify the clusters from the map. The main advantage of such a mapping is that it would be possible to gain some idea of the structure of the data by observing the map, due to topology preserving nature of the SOM.

Usually, SOM can be initialized using random values for the weight vectors. In this thesis we present a different approach for initializing SOM. This approach depends on the K-means algorithm as an initialization step for SOM.