ENHANCED DATA CLUSTERING AND CLASSIFICATION USING AUTO-ASSOCIATIVE NEURAL NETWORKS AND SELF ORGANIZING MAPS

ZALHAN BIN MOHD ZIN

UNIVERSITI TEKNOLOGI MALAYSIA

ENHANCED DATA CLUSTERING AND CLASSIFICATION USING AUTO-ASSOCIATIVE NEURAL NETWORKS AND SELF ORGANIZING MAPS

ZALHAN MOHD ZIN

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

Malaysia-Japan International Institute of Technology
Universiti Teknologi Malaysia

MARCH 2016

To my beloved mother and father

ACKNOWLEDGEMENT

The opportunity for this Newcastle University International Singapore (NUIS) and Universiti Teknologi Malaysia (UTM) Joint-PhD research would have been impossible without the initial effort, support and encouragement of my beloved supervisor from Universiti Teknologi Malaysia (UTM), the late Prof. Dato' Dr. Marzuki Khalid, who did not get the opportunity to see the fruit of his initiative, to him I am most indebted and grateful. In preparing this thesis, I was involved with many academicians and researchers who have contributed a lot to my understanding and thoughts. I wish to express my profound and sincere appreciation to the most important person in this research, my NUIS supervisor, Prof. Dr. Ehsan Mesbahi, for his endless unforgettable support, encouragement, guidance and critics. Without him, this thesis would not have been completed and presented. I am also very thankful to my supervisor from UTM, Prof. Datin Dr. Rubiyah binti Yusof, for her invaluable advice, guidance and motivations in this research.

I would also like to extend my sincere appreciations to my fellow postgraduate friends, researchers and staff of NUIS and Center for Artificial Intelligence and Robotics (CAIRO), UTM Kuala Lumpur for their support and advice, especially to Mrs. Azlin binti Ahmad. Her views and tips are useful indeed. Many thanks to Universiti Kuala Lumpur Malaysia France Institute (UniKL MFI) for their support and financial assistance. My special appreciation also extends to all my colleagues and others who have provided various forms of assistance. Last but not least, I would also like to extend my greatest appreciation to my beloved parents for their spiritual support, of which without I would be lost. My lovely wife, Mardewee Endut, who has been my pillar of strength, her love and understanding encourages me to strive for the success of this PhD. I am also very grateful to all my children and family members for their understanding of my commitment in this research and thesis preparation. I hope they will be inspired to succeed in their life as well.

ABSTRACT

This thesis presents a number of investigations leading to introduction of novel applications of intelligent algorithms in the fields of informatics and analytics. This research aims to develop novel methodologies to reduce dimensions and clustering of highly non-linear multidimensional data. Improving the performance of existing methodologies has been based on two fundamental approaches. The first is to look into making novel structural re-arrangements by hybridisation of conventional intelligent algorithms which are Auto-Associative Neural Networks (AANN) and Self Organizing Maps (SOM) for data clustering improvement. The second is to enhance data clustering and classification performance by introducing novel fundamental algorithmic changes known as M3-SOM in the data processing and training procedure of conventional SOM. Both approaches are tested, benchmarked and analysed using three datasets which are Iris Flowers, Italian Olive Oils and Wine through case studies for dimension reduction, clustering and classification of complex and non-linear data. The study on AANN alone shows that this non-linear algorithm is able to efficiently reduce dimensions of the three datasets. This paves the way towards structurally hybridising AANN as dimension reduction method with SOM as clustering method (AANNSOM) for data clustering enhancement. This hybrid AANNSOM is then introduced and applied to cluster Iris Flowers, Italian Olive Oils and Wine datasets. The hybrid methodology proves to be able to improve data clustering accuracy, reduce quantisation errors and decrease computational time when compared to SOM in all case studies. However, the topographic errors showed inconsistency throughout the studies and it is still difficult for both AANNSOM and SOM to provide additional inherent information of the datasets such as the exact position of a data in a cluster. Therefore, M3-SOM, a novel methodology based on SOM training algorithm is proposed, developed and studied on the same datasets. M3-SOM was able to improve data clustering and classification accuracy for all three case studies when compared to conventional SOM. It is also able to obtain inherent information about the position of one data or "sub-cluster" towards other data or sub-cluster within the same class in Iris Flowers and Wine datasets. Nevertheless, it faces difficulties in achieving the same level of performance when clustering Italian Olive Oils data due to high number of data classes. However, it can be concluded that both methodologies have been able to improve data clustering and classification performance as well as to discover inherent information inside multidimensional data.

ABSTRAK

Tesis ini membentangkan beberapa rangkaian penyelidikan yang membawa kepada pengenalan aplikasi baharu untuk algoritma pintar dalam bidang informatik dan analisis. Kajian ini bertujuan untuk membangunkan kaedah-kaedah baharu untuk mengurangkan dimensi dan pengugusan data berbilang dimensi dan bukan linear. Peningkatan prestasi kaedah sedia ada dibuat berdasarkan kepada dua pendekatan asas. Yang pertama ialah dengan memberi penekanan kepada kaedah baharu penyusunan semula struktur-struktur algoritma melalui penghibridan algoritma pintar konvensional iaitu Rangkaian Neural Automatik Bersekutu (*Auto-Associative Neural Networks* (AANN)) dan Peta Swaorganisasi (*Self Organizing Maps* (SOM)) bagi tujuan peningkatan prestasi pengugusan data. Yang kedua ialah bagi tujuan peningkatan prestasi pengugusan dan penkelasan data dengan memperkenalkan perubahan baharu terhadap kerangka pemprosesan data dan prosedur latihan di dalam algoritma konvensional SOM. Kedua-dua pendekatan ini diuji, ditanda aras dan dianalisis dengan menggunakan tiga set data bunga Iris, minyak zaitun Itali dan Wain melalui beberapa kes-kes kajian untuk mengurangkan dimensi, pengugusan dan klasifikasi data yang kompleks dan bukan linear. Kajian terhadap AANN sahaja menunjukkan algoritma bukan linear ini mempunyai keupayaan pada tahap ketepatan yang tinggi untuk mengurangkan dimensi ketiga-tiga set data tersebut. Ini membuka jalan ke arah penggabungan struktur AANN sebagai kaedah pengurangan dimensi dengan SOM sebagai kaedah pengugusan data (AANNSOM) bagi tujuan peningkatan pengelompokan data. Penghibridan algoritma AANN dan SOM (AANNSOM) diperkenalkan dan digunakan dalam kes-kes kajian pengelompokan data set-set data bunga Iris, minyak zaitun Itali dan Wain. Metodologi hibrid terbukti dapat meningkatkan ketepatan pengelompokan data, mengurangkan kesilapan pengkuantuman dan mengurangkan kerumitan pengiraan berbanding SOM dalam semua kajian kes. Walaubagaimanapun, kesilapan topografi didapati tidak konsisten disepanjang kajian dan ianya masih sukar bagi kedua-dua AANNSOM dan SOM untuk memberikan maklumat tambahan yang wujud di dalam set-set data seperti kedudukan sebenar sesuatu data di dalam kelompok. Oleh itu, metodologi M3-SOM yang berdasarkan algoritma latihan SOM adalah dicadangkan, dibangunkan dan dikaji pada set data bunga Iris, minyak zaitun Itali dan Wain. M3-SOM mampu meningkatkan prestasi ketepatan pengelompokan dan klasifikasi data untuk ketiga-tiga kajian kes berbanding konvensional SOM. Ia juga mampu untuk mendapatkan maklumat yang wujud mengenai kedudukan salah satu data atau "sub-kelompok" ke arah data lain atau sub-kelompok dalam kelas yang sama untuk set-set data bunga Iris dan Wain. Walau bagaimanapun, ia menghadapi kesukaran untuk mencapai tahap prestasi yang sama untuk kelompok data minyak zaitun Itali kerana bilangan kelas data yang tinggi. Walaubagaimanapun, boleh disimpulkan bahawa kedua-dua kaedah telah dapat meningkatkan prestasi pengugusan dan klasifikasi data serta menemui maklumat yang wujud di dalam data berbilang dimensi.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AANN	-	Auto-Associative Neural Networks
AGNES	-	Agglomerative Nesting
AI	-	Artificial Intelligence
ANN	-	Artificial Neural Networks
BMU	-	Best Matching Unit
BNN	-	Bottleneck Neural Networks
Cala	-	Calabria
CCA	-	Canonical Correlation Analysis
CCIA	-	Cluster Center Initialization Algorithm
CSar	-	Coastal Sardinia
CoA	-	Correspondence Analysis
DENCLUE	-	DENsity-based CLUstEring
DBSCAN	-	Density-based Spatial Clustering of Applications with Noise
DNA	-	Deoxyribonucleic acid
DRSC	-	Dimension Reduced Spectral Clustering
DBCLASD	-	Distributed based Clustering Algorithm For Mining Large Spatial Databases
DIANA	-	Divisive Analysis
DSOM	-	Double SOM
DGSOT	-	Dynamically Growing Self Organizing Tree
ELig	-	East Liguria
EM	-	Expectation-Maximization
FA	-	Factor Analysis
Fast-ICA	-	Fast Independent Component Analysis
FCM	-	Fuzzy C-Means

GDBSCAN	-	Generalized Density-based Spatial Clustering of Applications with Noise
GUI	-	Graphical User Interface
GHTSOM	-	Growing Hierarchical Tree SOM
AANNSOM	-	Hybrid AANN and SOM
ICA	-	Independent Component Analysis
IT	-	Information Technology
ISar	-	Inland Sardinia
KDD	-	Knowledge Discovery in Databases
LE	-	Laplacian Eigenmaps
LDA	-	Linear Discriminant Analysis
LTSA	-	Local Tangent Space Analysis
LPP	-	Locality Preserving Projection
LLC	-	Locally Linear Coordination
LLE	-	Locally Linear Embedded
Lowess	-	Locally Weighted Scatterplot Smoother
M3-SOM	-	M3-Self Organizing Maps
MVU	-	Maximum Variance Unfolding
MDS	-	Multi-Dimensional Scaling
NLPCA	-	Non Linear Principal Component Analysis
NMDS	-	Non-metric Multidimensional Scaling
NApu	-	North Apulia
OPTICS	-	Ordering Points to Identify the Clustering Structure
PLS	-	Partial Least Square
PO	-	Polar Ordination
PCA	-	Principal Component Analysis
qe	-	Quantization error
SOTA	-	Self Organizing Tree Algorithm
SOM	-	Self-Organizing Maps
Sici	-	Sicily
SVD	-	Singular Value Decomposition
SApu	-	South Apulia
STING	-	Statistical Information Grid Approach
SVM	-	Support Vector Machine
te	-	Topographic error

Umbr	-	Umbria
WLig	-	West Liguria

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction: Background and Motivation

Since the dawn of modern time, humans have usually been attracted in how nature functions, including themselves. This understanding has allowed mankind to reproduce certain forms of nature functions and to extend human limitation. An inspiring example is escaping gravitation; (in other words: flying), and the human race is currently increasingly fascinated in reproducing one of the most important features of nature: *intelligence*. Human has shown the ability to copy or learn from nature. Many existing inventions today are originated from the certain forms of nature or biological system. The flying ability of bird creates idea and path for human to invent aeroplane, robotic arm is replicated from human arm as well as barcode is devised from the uniqueness of finger thumb print and other inventions originated from features of nature. *Intelligence* however is one the features of nature that is not physically exist but can be studied and modelled. This unique feature inspires human to study about it, understand it and create machines or systems that could imitate it. This would become the artificial features of nature or artificial learning mechanisms for these machines or systems and is commonly known as Artificial Intelligence (AI). The learning ability of natural *intelligence* in clustering, classifying and recognizing objects or patterns that could reveal any form of inherent information motivates the author to initiate this research in intelligent machine learning algorithm in the field of informatics and analytics aiming to develop novel methodologies for reducing dimensions and clustering of highly non-linear multidimensional data.

Knowledge discovery has become one of the most challenging fields of study nowadays. Over the years, many work have started with the purpose of learning the machine to explore, discover and understand knowledge and information that could be beneficial for us. The need to extract information and knowledge from a huge pool of abundant and complicated data structures is enormous. It can be considered as one of the most important characteristic of the information age. The incredible development and advancement in Information Technology (IT), in particular the Internet, have led us into a technological situation that can be called "data explosion”. The aspect of data availability has been increased much more than assimilation capacity of any normal human being. According to a study conducted at Digital Universe Study [1], the amount of generated data in particular digital data have grown exponentially in the last decade and will continue to grow by 50-fold by the year 2020. This massive increase in both the volume and the variety of data demands for advances in methodology to understand, process, interpret and summarize the data. According to [2], we are overloaded by many types of data such as scientific data, medical data, financial data etc. These data require huge amount of time and demand vast attention. As a result, efforts must be taken by us to find ways to automatically: analyse the data, cluster or classify them, summarize them, discover and characterize trends or patterns in them, and identify abnormalities. Researches in statistics, visualization, artificial intelligence and machine learning are contributing to this one of the most active area of database research community.

From a more technical point of view, understanding the structure of multidimensional datasets arising from the data explosion is very important in data mining, data clustering, pattern recognition, and machine learning. This would enable us to obtain additional inherent information about the datasets. However, it would be difficult, costly and time consuming to acquire new knowledge from databases that are larger and complicated if it is done manually. It may even be not viable when the data exceed certain limits of size and complexity. Therefore, the automated analysis and visualisation of multidimensional datasets has become the focus of many scientific research with the objective is to find uniformities and relationships in the data. This could gain access to useful knowledge or inherent information that could be hidden inside the data. AI technique of Artificial Neural

Networks (ANN) can be considered as one of the promising tools in this field. Inspired by advances in biomedical research, ANN forms a class of algorithms aiming to simulate the biological neural networks. One of the well-known ANN algorithm, the Self-Organizing Maps (SOM) has become one of the most popular unsupervised learning algorithms. Created in the early 1980s by Finnish Professor Teuvo Kohonen, the work or research related to the algorithm, visualization and application of SOM has been published in more than 5000 research articles according to [3]. The algorithm comprehensively visualise natural groupings and relationships in the data and has been successfully applied in a wide variety of research areas such as image processing, speech recognition, life sciences, bio-informatics to financial analysis to name few.

SOM algorithm performs a non-linear projection of multidimensional data onto a two dimensional display. The mapping is topology-preserving which means that the more identical or similar two data are in the input space, the closer they will appear together on the final map. This allows the user to identify clusters such as large groupings of a certain type of input pattern. What features the members of a cluster have in common could also be revealed through additional examination. It is an efficient tool in information visualisation. The basic implementation of SOM is simple and the map representation is easy to understand. Furthermore, the results are reliable and the algorithm scales exceptionally well. In many studies and applications, SOM has proved to be excellent in helping in visualising and understanding the data.

Regardless of its popularity and ability to cluster multidimensional data, SOM has shown some issues and limitations with regards to its structure and visualization performances. The inherent information about one particular data with respect to other data, a sub-cluster, or an unknown data, have been often difficult to be interpreted and understood. Sever shortcomings in interpreting data clusters and the difficulties to distinguish different data classes using SOM particularly when the data are unknown, complex or non-linear have stirred the motivation for this research.

1.2 Aim and Objectives of the Research

The aim of this research is to devise a new paradigm of analytics using hybrid and complementary algorithmic methodologies for clustering data with complex, non-linear and unknown inherent structures.

The objectives of the research are as follows:

1. To review the strengths and weaknesses of various techniques for dimension reduction and clustering data with complex and non-linear characteristics.

2. To identify suitable datasets with non-linear characteristics for dimension reduction and clustering as Case-studies for implementation of proposed methodologies.

3. To improve the ability of the existing methodologies for dimension reduction techniques.

4. To develop new training methodologies for clustering, classifying and visualizing data with complex, non-linear and unknown structures.

1.3 Scope of the Research

The scope of the research is limited to the review of literature, study, development and analysis related to data mining, dimension reduction techniques, clustering algorithms and multidimensional datasets.

1.4 Research Framework and Contributions

1.4.1 Introduction

This research focuses on the studies, developments and enhancements of dimension reduction, clustering and classification of data with complex, non-linear and unknown structures. This research is carried out throughout the incremental steps which begin from the review of various dimension reduction and clustering techniques to the re-development and analysis of these techniques, and to the development and analysis of the proposed enhanced techniques for data clustering and classification.

1.4.2 Research Framework

The framework of the research is described as the following steps (Figure 1.1) and organized according to Chapter 2, Chapter 3, Chapter 4 and Chapter 5 of the thesis. The steps are:

1. The review of dimension reduction and clustering techniques is described in Chapter 2.
2. The re-development of AANN and the study and analysis on it's ability to reduce dimension of multidimensional data is described in Chapter 3.
3. The development of hybrid AANN and SOM and the study and analysis on it's clustering performances is presented in Chapter 4.
4. The development of M3-SOM and the study and analysis on it's clustering and classification performances is presented in Chapter 5.

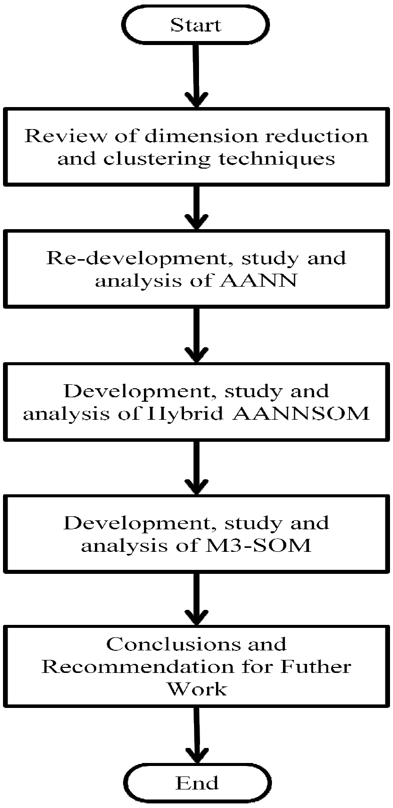
Start
Review of dimension reduction and clustering techniques
Re-development, study and analysis of AANN
Development, study and analysis of Hybrid AANNSOM
Development, study and analysis of M3-SOM
Conclusions and Recommendation for Futher Work
End

Figure 1.1 The framework of the research

1.4.3 Research Contributions

The contributions of the research are as follows:

1. The development and study of a new computational methodology based on SOM training algorithm. The detail investigations on SOM training algorithm specifically on it's neighbourhood functions, weights updates process and matrix of distances, lead to the creation of a new methodology in SOM data clustering, classification and visualization. This new methodology provides more inherent information about multidimensional data such as it's ability to identify the closest data, the farthest data or the group of data that may form a group of sub-clusters in the same class.

2. The development of hybrid AANNSOM algorithm for multidimensional data clustering. This algorithm is a combination of supervised learning and dimension reduction method of AANN with unsupervised learning and clustering method of SOM. The AANNSOM achieves higher clustering performance in clustering multidimensional datasets when compared to conventional SOM.

3. The development and study of AANN algorithm as dimension reduction method. The study focuses on the dimension reduction ability of this algorithm when dealing with datasets of different sizes, dimensions and clusters. AANN has demonstrated its ability to non linearly reduce the dimension of the datasets, but did not remove inherent characteristics of each dataset allowing them to be classified with high levels of accuracy. This algorithm has also demonstrated its ability to embed test data according to its class.

1.5 Thesis Structure

This thesis is divided into six chapters and is organized as follows. Chapter 1 provides an introduction to the research, aim, objectives and scope of the research, research contributions and also thesis outline. Chapter 2 covers the review of literature on the concepts of multidimensional data, data mining, dimension reduction, data clustering, clustering analysis and processes. Different methods of data clustering from K-means, hierarchical clustering, Principal Component Analysis (PCA) to AANN and SOM are presented in this chapter. The non-linear characteristics of multidimensional datasets used in this research are investigated, presented and discussed as well in Chapter 2. Chapter 3 is dedicated to the investigation of AANN's ability reducing dimensions and clustering of multidimensional data. Chapter 4 introduces combination of AANN and SOM algorithms as hybrid method for dimension reduction and data clustering. Chapter 5 presents a new methodology based on SOM training algorithm for data clustering, classification and visualization. Finally, the last chapter of the thesis (Chapter 6) presents overall conclusions of the thesis and recommendations for further work.

REFERENCES

1. J. G. a. D. Reinsel, "THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," IDC iview, 2012.
2. J. Han and M. Kamber, Data Mining: Concept and Techniques, Morgan Kauffman, 2001.
3. M. Cotrelle and M. Verleysen, "Advances in Self Organizing Maps," *Journal of Neural Networks,* vol. 19, no. 6-7, pp. 721-722, 2006.
4. P. C. Wong and R. D. Bergeron, "30 Years of Multidimensional Multivariate Visualization," in *Scientific Visualization, Overviews, Methodologies, and Techniques*, 1997.
5. D. Muhammad Khan and N. Mohamudally, "An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm," *International Journal of Information Processing and Management,* vol. 1, no. 1, pp. 104-113, 2010.
6. N. Varghese, V. Verghese, P. Gayathri and N. Jaisankar, "A Survey of Dimensionality Reduction and Classification Methods," *International Journal of Computer Science & Engineering Survey (IJCSES),* vol. 3, no. 3, pp. 45-54, 2012.
7. P. Andritsos, "Data Clustering Techniques," Toronto, 2002.
8. A. Jain, M. Murty and P. Flynn, "Data Clustering: A Review," *ACM Computing Surveys,* vol. 31, no. 3, pp. 264-323, 2000.
9. D. Napoleon and S. Pavalakodi, "A New Method For Dimension Reduction Using K-Means Clustering Algorithm For High Dimensional Dataset," *International Journal of Computer Applications (IJCA),* vol. 13, no. 8, pp. 41-46, 2011.
10. R. W. Sembiring, J. Mohamad Zain and A. Embong, "Dimension Reduction of Health Data Clustering," *International Journal on New Computer*

Architectures and Thier Applications (IJNCAA), pp. 1041-1050, 2011.

11. S. N. Kadir, D. F. Goodman and K. D. Harris, “High Dimensional Cluster Analysis with the Masked EM Algorithm,” *Journal of Neural Computation,* pp. 2379-2394, 2014.

12. A. Goh and R. Vidal, “Clustering and Dimensionality Reduction on Riemannian Manifolds,” in *IEEE Conference on Computer Vison and Pattern Recognition (CVPR)*, Anchorage, 2008.

13. D. Niu, J. G. Dy and M. I. Jordan, “Dimensionality Reduction for Spectral Clustering,” *Journal of Machine Learning Research (JMLR),* pp. 552-560, 2011.

14. T. Chandrasekhar, K. Thangavel and E. Elayaraj, “Effective Clustering Algorithms for Gene Expression Data,” *International Journal of Computer Applications (IJCA),* vol. 32, no. 4, pp. 25-29, October 2011.

15. J. Salome J and R. Suresh, “Efficient Clustering for Gene Expression Data,” *International Journal of Computer Applications (IJCA),* vol. 47, no. 5, pp. 30-35, June 2012.

16. C. Sorzano, J. Vargas and M. A. Pascual, “A Survey of Dimensionality Reduction Techniques,” Cornell University, 2014.

17. L. Van der Maaten, E. Postma and H. Van den Herik, “Dimensionality Reduction: A Comparative Review,” Maastricth University, Maastricth, 2008.

18. I. Nafornita and I. Buciu, “Linear and Nonlinear Dimensionality Reduction Techniques,” *Journal of Studies in Informatics and Control,* vol. 16, no. 4, pp. 431-444, 2007.

19. M. Belkin and P. Niyogi, “Laplacian Eigenmaps for Dimensionality Reduction and Data Representation,” *Journal of Neural Computation,* pp. 1373-1396, 2003.

20. A. Anaissi, P. J. Kenned and M. Goyal, “A Framework for High Dimensional Data Reduction in the Microarray Domain,” in *IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, 2010.

21. B. Scholkopf, A. Smola and K.-R. Muller, "Non Linear Component Analysis as a Kernel Eigenvalue Problem," *Journal of Neural Computation,* pp. 1299-1319, 1998.

22. M. A. Kramer, "Nonlinear Principle Component Analysis Using Autoassociative Neural Networks," *AIChE Journal,* vol. 37, pp. 233-243, 1991.

23. M. Daszykowski, B. Walczak and D. Massart, "A Journey Into Low-dimensional Spaces With Autoassociative Neural Networks," *International Journal of Pure and Applied Analytical Chemistry (Talanta),* vol. 59, pp. 1095-1105, 2003.

24. M. H. Asyali, D. Colak, O. Demirkaya and M. S. Inan, "Gene Expression Profile Classification: A Review," in *Current Bioinformatics*, vol. 1, 2006, pp. 55-73.

25. Y. Zhao and G. Karypis, "Data Clustering in Life Science," in *Methods in Molecular Biology*, vol. 224, New Jersey, Humana Press Incorporation, 2007, pp. 183-218.

26. J. C. Patra, E. L. Ang, P. K. Meher and Q. Zhen, "A New SOM-based Visualization Technique for DNA Microarray Data," in *International Joint Conference on Neural Networks*, Vancouver, Canada, 2006.

27. G. Kerr, H. Ruskin, M. Crane and P. Doolan, "Techniques for Clustering Gene Expressions Data," *Computers in Biology and Medecine,* vol. 38, no. 3, pp. 283-293, 2008.

28. S. Hautaniemi, O. Yli-Harja, J. Astola, P. Kauraniemi, A. Kallioniemi, M. Wolf, J. Ruiz, S. Mousses and O.-P. Kallioniemi, "Analysis and Visualization of Gene Expression Microarray Data in Human Cancer Using Self-Organizing Maps," *Journal of Machine Learning,* vol. 52, pp. 45-66, 2003.

29. T. Naenna, R. A. Bress and M. J. Embrechts, "A Modified Kohonen Network for DNA Splice Junction Classification," in *IEEE Region Ten Conference on Analog and Digital Techniques in Electrical Engineering*, Chiang Mai, Thailand, 2004.

30. M. K. Markey, J. Y. Lo, G. D. Tourassib and C. E. Floyd Jr., “Self Organizing Maps for Cluster Analysis of A Breast Cancer Database,” *Artificial Intelligence in Medecine,* vol. 27, pp. 113-127, 2003.

31. S. C. Madeira and A. L. Oliveira, “Biclustering Algorithms for Biological Data analysis: A Survey,” *IEEE/ACM Transaction on Computational Biology and Bioinformatics,* vol. 1, no. 1, pp. 24-45, 2004.

32. A. Ben-Dor, R. Shamir and Z. Yakhini, “Clustering Gene Expression Patterns,” *Journal of Computational Biology,* vol. 6, pp. 281-297, 1999.

33. A. K. Jain and R. C. Dubes, Algorithm for Clustering Data, New Jersey: Prentice Hall Incorporation, 1998.

34. M. Debnath, P. G. B.K.S. and P. S. Bisen, “Microarray,” in *Molecular Diagnostics: Promises and Possibilities*, 1st ed., Springer Science & Business Media, 2010, pp. 193-208.

35. G. J. McLachlan, R. W. Bean and S.-K. Ng, “Clustering,” in *Bioinformatics: Structure, Function and Application*, vol. 2, J. M. Keith, Ed., New Jersey, Humana Press, 2008, pp. 423-439.

36. T. D. Pham, C. Wells and D. I. Crane, “Analysis of Microarray Gene Expression Data,” in *Current Bioinformatics*, vol. 1, 2006, pp. 37-53.

37. S. Mocelin and C. R. Rossi, “Principles of Gene Microarray Data analysis,” in *Microarray Technology and Cancer Gene Profiling*, Padova, Italy, Landes Bioscience and Springer Science, 2007, pp. 19-30.

38. D. Jiang, C. Tang and A. Zhang, “Cluster Analysis for Gene Expression Data: A Survey,” in *IEEE Transactions on Knowledge and Data Engineering*, 2004.

39. M. B. Eisen, P. T. Spellman, P. O.Brown and D. Botstein, “Cluster Analysis And Display Of Genome Wide Expression Patterns,” in *Proceedings of National Academy of Science*, USA, 1998.

40. R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, 2nd Edition ed., Wiley Interscience, 2001.

41. T. Kohonen, “The Self-Organizing Maps,” in *Proceedings of the IEEE*, 1990.

42. T. Kohonen and H. Ritter, “Self-Organizing Semantic Maps,” in *Biological Cybernatics*, vol. 61, Springer-Verlag, 1989, pp. 241-254.

43. J. Vesanto and E. Alhoniemi, “Clustering of the Self Organizing Maps,” *IEEE Transaction on Neural Networks,* vol. 11, no. 3, pp. 586-600, 2000.

44. R. Xu and D. Wunsch II, “Survey of Clustering Algorithms,” *IEEE Transactions on Neural Networks,* vol. 16, no. 3, pp. 645-678, 2005.

45. P. Berkhin, A Survey of Clustering Data Mining Techniques, San Jose, CA: Springer Berlin Heidelberg, 2006, pp. 25-71.

46. J. MacQueen, “Some Methods For Classifications and Analysis Of Multivariate Observations,” in *The 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, California, 1967.

47. F. D. Smet, J. Mathys, K. Marchal, G. Thijs, B. D. Moor and Y. Moreau, “Adaptive Quality-based Clustering of Gene Expression Profiles,” *Journal of Bioinformatics,* vol. 18, pp. 735-746, 2002.

48. L. J. Heyer, S. Kruglyak and S. Yooseph, “Exploring Expression Data: Identification and Analysis of Co-expressed Genes,” *Journal of Geneome Research,* pp. 1106-1115, 1999.

49. D. D. Cohen, A. A. Melkman and S. Kasif, “Hierarchical Tree Snipping: Clustering Guided by Prior Knowledge,” *Journal of Bioinformatics,* vol. 23, pp. 3335-3342, 2007.

50. W. Pan, “Incorporating Gene Functions As Priors in Model-based Clustering of Microarray Gene Expression Data,” *Journal of Bioinformatics,* vol. 22, pp. 795-801, 2006.

51. L. I. Smith, 2002. [Online]. Available: http://neurobot.bio.auth.gr/2005/a-tutorial-on-principal-components-analysis/. [Accessed 05 February 2012].

52. J. Giraudel and L. Sovan, “A Comparison of Self Organizing Map Algorithm and Some Conventional Statistical Methods for Ecological Community Ordination,” *Journal of Ecological Modelling,* vol. 146, pp. 329-339, 2001.

53. J. Mistry, F. V. Nelwamondo and T. Marwala, “Missing Data Estimation Using Principle Component Analysis and Autoassociative Neural Networks,” *Journal of Systemics, Cybernatics and Informatics,* vol. 7, no. 3, pp. 72-79, 2009.

54. V. M. Stone, “The Autoassociative Neural Network-A Network Worth Considering,” in *World Automation Congress (WAC)*, Hawaii, 2008.

55. M. Lukk, M. Kapushesky, J. Nikkil, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen and A. Brazma, “A Global Map of Human Gene Expression,” *Nature Biotechnology,* pp. 322-324, 2010.

56. R. A. Fisher, “UCI Machine Learning Repository,” 2007. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Iris. [Accessed 31 January 2011].

57. K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery and W. L. Ruzzo, “Model-Based Clustering and Data Transformations for Gene Expression Data,” *Journal of Bioinformatics,* vol. 17, pp. 977-987, 2001.

58. G. McLachlan, R. W. Bean and D. Peel, “A Mixture Model-based Approach to the Clustering of Microarray Expression Data,” *Journal of Bioinformatics,* vol. 18, no. 3, pp. 413-422, 2001.

59. M. Medvedovic, K. Y. Yeung and R. E. Bumgarner, “Bayesian Mixture Model-based clustering of Replicated Microarray Data,” *Journal of Bioinformatics,* vol. 20, pp. 1222-1232, 2004.

60. G. C. Tseng and W. H. Wong, “Tight Clustering: A Resampling-based Approach for Identifying Stable and Tight Patterns in Data,” *Journal of Biometrics,* vol. 61, pp. 10-16, 2005.

61. C. Fraley and A. E. Raftery, “Model based Clustering, Discriminant Analysis And Density Estimation,” *Journal of the American Statistical Association,* vol. 97, pp. 611-631, 2002.

62. A. V. Lukashin and R. Fuchs, “Analysis of Temporal Gene Expression Profile: Clustering by Simulated Annealing and Determining the Optimal Number of Clusters,” *Journal of Bioinformatics,* vol. 17, no. 5, pp. 405-414, 2001.

63. K. Bryan, P. Cunningham and N. Bolshakova, “Application of Simulated Annealing to the Biclustering of GeneExpression Data,” in *IEEE Transaction on Information Technology in Biomedecine*, 2006.

64. J. Sander, M. Ester, H.-P. Kriegel and X. Xu, “Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications,” *Journal*

of Data Mining and Knowledge Discovery, vol. 2, no. 1, pp. 169-194, 1998.

65. A. Hinneburg and D. A. Keim, “An Efficient Approach to Clustering in Large Multimedia Databases with Noise,” in *The 4th International Conference on Knowledge Discovery and Datamining (KDD'98)*, New York, 1998.

66. W. Wang, J. Yang and R. Muntz, “STING: A Statistical Information Grid Approach to Spatial Data Mining,” in *The 23rd International Conference on Very Large Data Bases*, San Francisco, 1997.

67. G. Sheikholeslami, S. Chatterjee and A. Zang, “WaveCluster: A Wavelet-based Clustering Approach for Spatial Data in Very Large Databases,” *The International Journal on Very Large Data Bases,* vol. 8, no. 3-4, pp. 289-304, 2000.

68. S. Haykin, Neural Networks, a Comprehensive Foundations, Prentice Hall, 1999.

69. “MathWorks,” Mathworks Incorporation, 1994-2015. [Online]. Available: http://www.mathworks.com/products/neural-network/index.html. [Accessed 10 March 2012].

70. J. Vesanto, “SOMToolbox Homepage,” 2008. [Online]. Available: http://www.cis.hut.fi/somtoolbox/. [Accessed 20 January 2011].

71. M. Scholz, F. Kaplan, C. L. Guy, J. Kopka and J. Selbig, “Non-linear PCA: A Missing Data Approach,” *Journal of Bioinformatics,* vol. 21, no. 20, pp. 3887-3895, 2005.

72. M. Abidi, S. Yasuki and P. Crilly, “Image Compression Using Hybrid Neural Networks Combining The Auto-Associative Multilayer Perceptron and The Self Organizing Feature Map,” *IEEE Transaction on Consumer Electronics,* vol. 40, no. 4, pp. 796-811, 1994.

73. P. Baldi, “Autoencoders, Unsupervised Learning and Deep Architecture,” *Journal of Machine Learning and Pattern Recognition,* vol. 27, pp. 37-50, 2012.

74. A. Flexer, “On the Use of Self Organizing Maps for Clustering and Visualization,” in *International Conference on Principle on Data Mining and Knowledge Discovery*, Prague, Czech Republic, 1999.

75. J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas, “SOM Toolbox for Matlab 5,” 2000.

76. E. Domany, “Cluster Analysis of Gene Expression Data,” *Journal of Statistical Physics,* vol. 110, pp. 3-6, March 2003.

77. X. Xiao, E. R. Dow, R. Eberhart, Z. B. Miled and R. J. Oppelt, “Gene Clustering Using Self Organizing Maps and Particle Swarm Optimization,” in *IEEE International Symposium on Parallel and Distributed Processing*, France, 2003.

78. T.-S. Chon, “Self Organizing Maps Applied to Ecological Sciences,” *Ecological Informatics,* vol. 6, no. 1, pp. 50-61, 2011.

79. H. Jin, W.-H. Shum and K.-S. Leung, “Expanding Self Organizing Maps for Data Visualization and Cluster Analysis,” *Journal of Information Sciences,* vol. 163, pp. 157-173, 2004.

80. H. Yin, “The Self-Organizing Maps: Background,Theories, Extensions and Applications,” in *Computational Intelligence: A Conpendium Studies in Computational Inteligence*, vol. 115, Springer, 2008, pp. 715-762.

81. A. Forti and G. L. Foresti, “Growing Hierarchical Tree SOM: An Unsupervised Neural Network With Dynamic Topology,” *Journal of Neural Networks,* vol. 19, pp. 1568-1580, 2006.

82. J. Gorrichaa and V. Loboa, “Improvements on the Visualization of Clusters in Geo-referenced Data Using Self-Organizing Maps,” in *Computers and Geosciences*, vol. 37, Elsevier, 2011.

83. K. Fujimura, K. Masuda and Y. Fuku, “A Consideration on the Multi-dimensional Topology in Self Organizing Maps,” in *International Symposium on Intelligent Signal Processing and Communications (ISPACS)*, Totori, Japan, 2006.

84. P. Stefanovic and O. Kurasova, “Visual Analysis of Self Organizing Maps,” in *Non Linear Analysis: Modelling and Control*, vol. 16, Vilnius, Vilnius University, 2011, pp. 488-504.

85. J. Herrero, A. Valencia and J. Dopazo, “A Hierarchical Unsupervised Growing Neural Network For Clustering Expression Patterns,” *Journal of Bioinformatics,* vol. 17, no. 2, pp. 126-136, 2001.

86. F. Luo, L. Khan, F. Bastani, I.-L. Yen and J. Zhou, “A Dynamically Growing Self-Organizing Tree(DGSOT) for Hierarchical Clustering Gene Expression Profiles,” *Journal of Bioinformatics,* vol. 20, no. 16, pp. 2605-2617, 2004.

87. J. M. Z. James S. Kirk, “A Two-stage Algorithm For Improved Topography Preservation in Self Organizing Maps,” in *IEEE International Conference on Systems, Man, and Cybernetics*, 2000.

88. E. A. Uriaite and F. D. Martin, “Topology Preservation in SOM,” *International Journal of Mathematical and Computer Sciences,* vol. 1, no. 1, pp. 19-22, 2005.

89. M.-C. Su and H.-T. Chang, “A New Model of Self-Organizing Neural Networks and Its Application in Data Projection,” *IEEE Transactions on Neural Networks,* vol. 12, no. 1, pp. 153-158, 2001.

90. A. Sugiyama and M. Kotani, “Analysis of Gene Expression Data by Using Self Organizing Maps and K-means Clustering,” in *IEEE International Joint Conference on Neural Networks*, Honolulu, USA, 2002.

91. M. A. Kraaijveld, J. Mao and A. K. Jain, “A Nonlinear Projection Method Based on Kohonen's Topology Preserving Maps,” *IEEE Transaction on Neural Networks,* vol. 6, no. 3, pp. 548-559, 1995.

92. G. Polzlbauer, “Survey and Comparisons of Quality Measures for Self Organizing Maps,” in *Fifth Workshop on Data Analysis (WDA04)*, Vysoke Tatry, Slovakia, 2014.

93. “Minitab,” Minitab Incorporation, 2014. [Online]. Available: http://www.minitab.com/en-us/. [Accessed 28 August 2014].

94. J. Zupan and J. Gasteiger, “Neural Network in Chemistry and Drug Design,” 2006. [Online]. Available: http://www2.ccc.uni-erlangen.de/publications/ANN-book/datasets/. [Accessed 16 Feb 2011].

95. M. e. a. Forina, “UCI Machine Learning Repository,” 2007. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Wine. [Accessed 23 December 2013].

96. P. Prabhu and N. Anbazhagan, “Improving the Performance of K-Means Clustering For High Dimensional Data Set,” *International Journal on Computer Science and Engineering,* vol. 3, no. 6, pp. 2317-2322, 2011.

97. M. Y. Kiang, “Extending the Kohonen Self-Organizing Map Networks forClustering Analysis,” *Computational Statistics & Data Analysis,* vol. 38, pp. 161-180, 2001.

98. M. J. Embrechts, B. J. Hargis and J. D. Linton, “Augmented Efficient BackProp for Backpropagation Learning in Deep Autoassociative Neural Networks,” in *International Joint Conference on Neural Networks (IJCNN)*, Barcelona, 2010.

99. G. Kerschen and J.-C. Golinval, “Feature Extraction Using Auto-Associative Neural Networks,” *Smart Materials and Structures,* vol. 13, no. 1, pp. 211-219, 2004.

100. M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems 2nd Edition, Pearson Education Limited, 2005.

101. M.-C. Su, T.-K. Liu and H.-T. Chang, “Improving the Self-Organizing Feature Map Algorithm Using an Efficient Initialization Scheme,” *Tamkang Journal of Science and Engineering,* vol. 5, no. 1, pp. 35-48, 2002.

102. E. Mesbahi, “Cryptic codes in non-coding DNA: Autoassociative Neural Networks and multidimensional Self Organising Maps (SOM) mediated prediction of positional significance of cis-elements in co-regulated expression systems,” 29 March 2012. [Online]. Available: http://www.ncl.ac.uk/marine/research/project/1997. [Accessed 15 July 2012].