# ITERATIVE LOCAL GAUSSIAN CLUSTERING TO EXTRACT INTERESTING PATTERNS ON SPATIO-TEMPORAL DATABASE

TIRWANI BINTI AMAN

UNIVERSITI TEKNOLOGI MALAYSIA

# ITERATIVE LOCAL GAUSSIAN CLUSTERING TO EXTRACT INTERESTING PATTERNS ON SPATIO-TEMPORAL DATABASE

TIRWANI BINTI AMAN

A project report submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia

To my beloved husband, Mohd Yazid Husain
and children,
Nur Husna, Hanis Amirah & Muhammad Ariff Syahmi

And

My greatest parent, Mak & Ayah :-
Kalthum Sulaiman & Aman Mahmood

# ACKNOWLEDGEMENT

Alhamdulillah, praise be to Allah, only with His will and blessing that I manage to finish my project.

My first appreciation goes to my dedicated supervisor, Associate Prof. Dr. Ito Wasito for his guidance, motivation, suggestion and generous help in making this project a success. His advice, experience sharing and continuous teaching is much appreciated and was very helpful for me to complete this project. I would also like to thank my examiner, Prof. Dr. Robert Colomb, Assoc. Prof. Dr. Naomie Salim and Assoc. Prof. Dr. Ali Selamat for their positive comments and suggestions.

Thank you to my colleagues and friends, Rafidah Hanem, Aina Muzdalifah, Fakhrul, and others for their support, assistance and sharing. Not to forget all my lecturers and staff in Faculty of Computer Science and Information System for their support and help during the course of my study at UTM.

Finally, I would like to dedicate this thesis to my parents and family, my supportive husband and children for their patience, their endless love, and prayer for all this time. Thank you.

# ABSTRACT

The study of spatio-temporal data mining in extracting and analyzing interesting patterns from spatio-temporal database has attract great interest in diverse research field. Huge amount of research has been done in either spatial data mining or temporal data mining and numbers of clustering algorithms have been proposed. However, not much research has been done in the integration of both spatial and temporal data mining, which is spatio-temporal data mining. The focuses of this study is to analyses the Iterative Local Gaussian Clustering (ILGC) algorithm and implement the algorithm to the spatio-temporal data, which is crime data. . In ILGC approach, the K- nearest neighbor (KNN) density estimation is extended and combined with Gaussian kernel function, where KNN contribute in determining the best local data iteratively for Gaussian kernel density estimation. The local best is defined as the set of neighbors data that maximizes the Gaussian kernel function. ILGC used Bayesian rule in dealing with the problem of selecting best local data. To test and validate the ILGC approach, other clustering method, which is K-Means and Self Organizing Map (SOM) will be implemented on the same data sets.

# TABLE OF CONTENTS

## LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS

| | | |
|---|---|---|
| $\eta$ | - | learning rate |
| $\mu$ | - | mean |
| $\sigma^2$ | - | variance |
| $K$ | - | Kernel |
| $V$ | - | *v*olume |
| r | - | convergence rate |
| k | - | nearest neighbors / number of cluster |

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| CLARA | - | Clustering Large Applications |
| CLARANS- | | Clustering Large Applications based on RANDomized Search |
| DM | - | Data Mining |
| ILGC | - | Iterative Local Gaussian Clustering |
| KNN | - | K-Nearest Neighbor |
| PAM | - | Partition Around Mediods |
| PCA | - | Principal Component Analysis |
| SI | - | Silhouette index |
| SOMs | - | Self Organizing Map |
| SVD | - | Singular Value Decomposition |

## LIST OF APPENDICES

# CHAPTER 1

# PROJECT OVERVIEW

## 1.1 Introduction

The study of spatio-temporal data mining in extracting and analyzing interesting patterns from spatio-temporal database has attract great interest in diverse research field. The main reason for this interest is the availability of datasets containing both spatial and temporal data elements across wide applications ranging from public safety (such as in crime investigation), public health (such as disease reports), transportation systems to product lifecycle management.

Spatio-temporal databases are system that manages both time and space information. It embodies spatial, temporal and spatio-temporal database concepts and it captures simultaneously spatial and temporal aspects of data (K.Manolis et.al, 2003). In order to retrieve useful information and interesting patterns, knowledge discovery process and data mining are used. Spatio-temporal data mining refers to the process of discovering meaningful patterns, trends and correlation in spatio-temporal databases. The major tasks in data mining include clustering, classification, prediction and association-rule.

Clustering is a technique to discover clusters of similar characteristics and group them into homogeneous clusters form the given data. Since spatio-temporal database are different from both spatial and temporal databases, they need different approach for clustering. The main issue in clustering of spatio-temporal database is to handle space and time dimension simultaneously.

## 1.2 Problem Background

Finding useful patterns in data has been given many names, such as data mining, knowledge extraction, pattern recognition and information discovery. The term data mining has been mostly used by statistician, data analyst, and information system communities. The availability of huge volume of geospatial data that continuously updated , has greatly challenge our ability to digest the data and gain useful knowledge that would otherwise lost.

During the last three decades, there are great numbers of research that has been done on data mining, which only focus on either spatial aspects or temporal aspects of data, separately. However, in today's digital world, both time and space is needed to be present as first-class concepts in any information system (T.Sellis et al,2003). Both aspects are central to our understanding of geographic process and events. Spatio-temporal data mining refers to the extraction of implicit knowledge, spatial and temporal relationships or other patterns not explicitly stored in spatio-temporal database (X.Yao, 2000). The mining of spatio-temporal patterns can lead to important observations in any applications, such as environmental monitoring and crime studies. The huge amount of spatio-temporal data and the complexity of its data types, data representation, and spatial data structure cause spatio-temporal data mining process challenging. Since there is still not much research and studies on spatio temporal data mining, significant attention is needed. One important technique for mining spatio-temporal data is clustering.

Clustering is the process of grouping set of physical or abstracts into classes of similar objects. Data clustering is under strong development in areas of research including data mining and become highly active topic in data mining. Many clustering algorithms have been develop, such as k-means, k-mediods, density based method, squared error, fuzzy clustering and self-organizing maps (SOM). Unfortunately, most of the studies focused on either spatial or temporal database, while clustering in spatio-temporal database need more investigation.

## 1.3 Problem Statement

Spatio-temporal data is complex (Gahegan, 2001) and it is characterized by high volumes of data. Several different clustering techniques have been introduced by various authors and researchers. Most proposed algorithms for clustering are inherited from neighboring field such as machine learning and statistics. Many approaches in clustering concern simple data, composed of elements represented as single points in some multidimensional space (Kakkar, 2004). Solution under these restrictions could not be applied appropriately to more complex data such as spatio-temporal data.

Our work is in the area of implementing clustering techniques for spatio-temporal databases. The implementation of Iterative Local Gaussian Clustering

(ILGC) has shown achievement in bioinformatics. The algorithm has yet to be implemented in spatio-temporal database. In this study, the ILGC will be tested to the crime data, as crime data mining has a promising future for increasing the effectiveness and efficiency of criminal analysis. With respect to spatio-temporal domain, experimentation and implementation of this algorithm will be done in Matlab in order to determine the effectiveness of iterative local clustering for spatio-temporal database management.

The hypothesis of this study can be stated as
" How efficient is the Iterative Local Gaussian Clustering algorithm on spatio-temporal database compared to K-Means and Kohonen Self-Organizing Maps? "

## 1.4 Project Aim

This project aims is to determine how to do clustering on spatio-temporal database and the effectiveness of the ILGC techniques, compared to SOM and K-Means algorithm. The crime data set will be used in this study.

## 1.5 Project Objective

Objectives of the project are :

1. To survey the use of existing clustering techniques on spatio-temporal database

2. To explore and analyze the use of Iterative Local Clustering on spatio-temporal database
3. To analyze the effectiveness of Iterative Local Clustering on spatio-temporal database compared to K-Means and Self Organizing Maps (SOM) using Silhouette Index

## 1.6 Project Scope

The scopes of the project includes :

1. Experimental comparisons between Iterative Local Clustering with other existing clustering techniques which is K-Means and SOM on spatio-temporal data mining
2. Implementation of Iterative Local Gaussian Clustering algorithm using Matlab environment.

## 1.7 Important of Research Study

The performance between ILGC, SOM and K-Means clustering technique is analyzed, such that we can determine which method is better for clustering the spatio-temporal database. It is important to identify appropriate technique for future research and can be implemented in real world situation, especially in data crime analysis.

## 1.8 Organization of Report

The report consists of 4 chapters. Chapter 1 presents the introduction of the study, problem background, objective and project scope. Chapter 2 explores the literature review on clustering, spatio-temporal database, k-means, SOM and ILGC technique. Project methodology is discussed in Chapter 3 and Chapter 4 analyzes the initial findings of the study.

# REFERENCES

Abidi, S.S.R, Ong, J.(2000). *A Data Mining Strategy for Inductive Data Clustering: A Synergy Between Self-Organizing Neural Networks and K-Means Clustering Technique*. IEEE.

Abonyi, J., Feil, B. (2007). *Cluster Analysis For Data Mining and System Identification*. Birkhauser Verlag.

Abraham, T., Roddick, J.F.(1998). *Opportunities For Knowledge Discovery in Spatio-Temporal Information Systems*.

Abraham, T., Roddick, J.F.(1999). Survey of Spatio-temporal Databases. Geoinformatica 3..61-99. Boston: Kluwer Academic Publishers.

Aldridge, M. (2006). *Clustering: An Overview*. In Micheal W.Berry, Murray Browne (ed.), *Lecture Notes in Data Mining*, (pp 99-109). World Scientific Publishing Co. Pte. Ltd.

Awan, A.M., M.Sap M.N.(2005). Clustering Spatial Data using a Kernel-based Algorithm, *Proceedings of the Postgraduate Annual Research Seminar.*

Bandyopadhyay, S.and Maulik, U.(2005). *Knowledge Discovery and Data Mining*. In Sanghamitra et al(ed.). *Advanced Methods for Knowledge Discovery from Complex Data*. Springer.

Bell,D.A., Anand, S.S.& Shapcott,C.M. (1994), Database Mining in Spatial Databases, *Proceeding of the International Workshop on Spatio-Temporal Databases*.

Berkhin, P.(2002). *Survey of Clustering Data Mining Techniques*. CA: Accrue Software Inc.

Claudia, M.A., Arlindo, L.O., *Temporal Data Mining: An Overview*. In Lecture Notes in Computer Science.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). *Cluster analysis and display of genome-wide expression pattern*. PNA USA 95.

Ertoz, L., Steinbach, M., Kumar, V. (2002). A new shared nearest neighbor clustering algorithm and its application. In *Proceedings of the workshop on Clustering High Dimensional Data and its application*. pp.105-115. Arlington, VA, USA

Fayyad, M.U., Shapiro, G. P., Smuth,P., Uthurusamy, R.(1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press.

Gaffney, S.and Smyth, P.(1999). Trajectory Pattern Mining with mixture of regression models. *Proceeding of the 5th ASM SIGKDD International Conference on Knowkedge Discovery and Data Mining.*

Halkidi, M., Batistakis,Y., Vazirgiannis, M. (2001). On Clustering Validation Techniques, *Journal of Intelligent Information Systems.* 17:2/3, 107–145.

Heinrich, K.E. (2006). *Clustering: Partitional Algorithms*. In Micheal W.Berry, Murray Browne (ed.) ,*Lecture Notes in Data Mining*, (pp 121-132). World Scientific Publishing Co. Pte. Ltd.

Hirano, S. and Tsumoto, S.( 2005). *A Clustering Method For Spatio-Temporal Data And Its Application To Soccer Games Record*. pp. 612-621. Springer Verlag.

Jain, A.K. & Dubes, R.C. (1998). *Algorithms for Clustering Data*. Prentice Hall Advanced Reference Series.

Juha Vesanto(1997). Data Mining Techniques Based on the Self-Organizing Map

Kakkar, S.(2004). Methodology for Clustering Spatio-Temporal Databases. Master Thesis. University of Cincinnati.

Kalnis, P., Mamoulis, N. and Bakiras, S. (2005). On discovering Moving Clusters in Spatio-Temporal Data . *Proceeding of 9th International Sym. On Spatial and Temporal Database*.

Kohonen, T.(2001). *Self-Organizing Maps*. Springer Series in Information Sciences, Springer.

Koperski, K., Adhikary J., Han, J.(1996). Spatial Data Mining : Progress and Challenges Survey Paper, *Proceedings of 1996 ACM-SIGMOID Workshop on Research Issues on Data Mining and Knowledge Discovery*. Montreal, Canada.

Krzysztof J.Cios, Witolld Pedrycz, Roman W.Swiniarski, Lukasz A.Kurgan(2007). *Data Mining : A Knowledge Discovery Approach*. NY: Springer Sience+Business Media.

Larose, D.T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*: John Wiley & Sons.

Lu, W., Ooi, B.C.(1993). Discovery of General Knowledge in Large Spatial Databases, *Proceedings of the 1993 FarEast Workshop on GIS*, (pp21-23). Singapore.

Gahegan M. et al(2001). The Integration of Geographic Visualization with Knowledge Discovery in Databases and Geocomputation. Cartography and Geographic Information Systems, Special Issue on Research Challenges in Geovisualization.

Martinez, W.L., Martinez, A.R.(2004). *Exploratory Data Analysis with Matlab.* Chapman & Hall/CPC, UK.

Miller, N.J and Han, J.(2001). *Geographic data mining and knowledge discovery: An Overview*. In *Geographic Data Mining and Knowledge Discovery* (pp3-32). London, New York: Taylor and Francis.

Moore, A. (2001). K-means and hierarchical clustering, Course notes, http://www-2.cs.cmu.edu/~awm/~2004

Ng, R.and Han, J.(1994). Efficient and Effective Clustering methods for spatial data mining. *Proceeding of the 20th Conference on VLD.* 144-155. Santiago, Chile.

Ramanathan, K.and Guan,S.K. (2006). *Recursive Self Organizing Maps with Hybrid Clustering*. IEEE.

Roddick, J.F,.Egenhofer, M.J, Hoel, E., Papadia, D. *Spatial, Temporal and Spatio-Temporal Databases- Hot Issues and Directions for PhD Research*.

Roddick, J.F., Spilioupoulou, M.(2002), A survey of temporal knowledge discovery paradigms and methods. *IEEE transactions of Knowledge and Data Engineering*.

Sellis, T.et al(2003). *Spatio-temporal Database*, LNCS 2520, pp.1-8, Springer-Verlag Berlin Heidelberg.

Shlens, J. (2005). A Tutorial on Principal Component Analysis, California.

Sourina, O.and Liu, D.(2005). *Visual Interactive Clustering and Querying Spatio-Temporal Data*. In O.Gervasi et al. (Eds.). ICCSA 2005, LNCS, pp968-977, Springer Verlag Berlin Heidelberg.

Tran,T.N., Wehrens, R., Buydens, L.M.C.(2005). KNN-Kernel density based clustering for high dimensional multivariate data. *Computational Statistical & Data Analysis*, Elsevier.

Wachowicz, M. (2002). Uncovering Spatio-Temporal Patterns in Environmental Data, *Water Resources Management Journal*, Special Issue on geocomputation in water resources and environment , pp. 469-487.

Wang, K., Wang, B., Peng, L.(2009). CVAP: Validation for Cluster Analyses. *Data Science Journal*.Vol. 8, 88-93.

Wasito, I., Mohd Hashim, S.Z., Sukmaningrum, S.(2007). Iterative local Gaussian clustering for expressed genes identification linked to malignancy of human colorectal carcinoma. Bioinformation2. 175-181. Biomedical Informatics Publishing Group.

Watts, M.J. and Warner, S.P. (2008). *Estimating the risk of insect species invasion : Kohonen Self Organizing Maps versus K-Means Clustering*. Elsevier B.V.

Webb, A.(2002). Statistical Pattern Recognition. pp.81-122. Malvern, UK:Wiley.

Wen, X.and Somogyi, R.E.(1998*). Large scale gene expression mapping of central nervous system development*. pp.334-339. PNAS USA.

Yao, X.(2003). Research Issues in Spatio-temporal Data Mining, In *Workshop on Geospatial Visualization and Knowledge Discovery.*18-20. Lansdowne, Virginia.