IMPROVED PATTERN EXTRACTION SCHEME FOR CLUSTERING MULTIDIMENSIONAL DATA

AINA MUSDHOLIFAH

UNIVERSITI TEKNOLOGI MALAYSIA
IMPROVED PATTERN EXTRACTION SCHEME FOR CLUSTERING MULTIDIMENSIONAL DATA

AINA MUSDHOLIFAH

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

JULY 2013
To my beloved husband, mother, and farther.
ACKNOWLEDGEMENTS

In preparing this thesis, I was in contact with many people, researchers, and academics. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, PM. Dr. Siti Zaiton Mohd Hashim, for encouragement, guidance, and critics.

I am also indebted to Ministry of Science and Technology of Malaysia (MOSTI) and Universiti Teknologi Malaysia (UTM) for funding my Ph.D. study. Universitas Gadjah Mada (UGM) also deserves special thanks for generous support of my Ph.D. study.

My fellow Indonesian Students should also be recognized for their support. My sincere appreciation also extends to all my sisters, colleagues and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family members.
ABSTRACT

Multidimensional data refers to data that contains at least three attributes or dimensions. The availability of huge amount of multidimensional data that has been collected over the years has greatly challenged the ability to digest the data and to gain useful knowledge that would otherwise be lost. Clustering technique has enabled the manipulation of this knowledge to gain an interesting pattern analysis that could benefit the relevant parties. In this study, three crucial challenges in extracting the pattern of the multidimensional data are highlighted: the dimension of huge multidimensional data requires efficient exploration method for the pattern extraction, the need for better mechanisms to test and validate clustering results and the need for more informative visualization to interpret the “best” clusters. Density-based clustering algorithms such as density-based spatial clustering application with noise (DBSCAN), density clustering (DENCLUE) and kernel fuzzy C-means (KFCM) that use probabilistic similarity function have been introduced by previous works to determine the number of clusters automatically. However, they have difficulties in dealing with clusters of different densities, shapes and size. In addition, they require many parameter inputs that are difficult to determine. Kernel-nearest-neighbor (KNN)-density-based clustering including kernel-nearest-neighbor-based clustering (KNNClust) has been proposed to solve the problems of determining smoothing parameters for multidimensional data and to discover cluster with arbitrary shape and densities. However, KNNClust faces problem on clustering data with different size. Therefore, this research proposed a new pattern extraction scheme integrating triangular kernel function and local average density technique called TKC to improve KNN-density-based clustering algorithm. The improved scheme has been validated experimentally with two scenarios: using real multidimensional spatio-temporal data and using various classification datasets. Four different measurements were used to validate the clustering results; Dunn and Silhouette index to assess the quality, F-measure to evaluate the performance of approach in terms of accuracy, ANOVA test to analyze the cluster distribution, and processing time to measure the efficiency. The proposed scheme was benchmarked with other well-known clustering methods including KNNClust, Iterative Local Gaussian Clustering (ILGC), basic k-means, KFCM, DBSCAN and DENCLUE. The results on the classification dataset demonstrated that TKC produced clusters with higher accuracy and more efficient than other clustering methods. In addition, the analysis of the results showed that the proposed TKC scheme is capable of handling multidimensional data, validated by Silhouette and Dunn index which was close to one, indicating reliable results.
ABSTRAK

Data multidimensi merujuk kepada data yang mengandungi sekurang-kurangnya tiga atribut atau dimensi. Dengan adanya sejumlah besar data multidimensi yang telah dikumpul sejak bertahun-tahun, keupayaan mencerna data dan mendapatkan pengetahuan yang berguna semakin mencabar. Teknik penggugusan membolehkan proses manipulasi pengetahuan ini membolehkan analisis pola menarik yang boleh memberi manfaat kepada pihak yang berkenaan. Dalam kajian ini, tiga cabaran penting dalam mengekstrak pola data multidimensi diketengahkan: dimensi data multidimensi yang besar yang memerlukan kaedah penjelajahan yang cekap sebagai kaedah pengeluaran pola, keperluan bagi mekanisme yang lebih baik untuk menguji dan mengesahkan hasil penggugusan dan keperluan visualisasi yang lebih bermaklumat bagi mentafsir gugusan "terbaik". Algoritma penggugusan berdasarkan kepadatan seperti penggugusan ruang dengan hingar berasas kepadatan (DBSCAN), penggugusan kepadatan (DENCLUE) dan kernel fuzzy C-means (KFCM) yang menggunakan fungsi persamaan kebarangkalian telah diperkenalkan oleh kajian terdahulu untuk menentukan bilangan gugusan secara automatik. Walau bagaimanapun, masalah timbul apabila berhadapan dengan gugusan yang berbeza kepadatan, bentuk dan saiz. Di samping itu, algoritma-algoritma ini memerlukan banyak input parameter yang sukar untuk ditentukan. Penggugusan berasaskan kepadatan keranjang terdekat (KNN) termasuk penggugusan berasas keranjang-terdekat (KNNClust) telah dicadangkan untuk menyelesaikan masalah dalam menentukan parameter pelicinan bagi data multidimensi dan menemui gugusan dengan bentuk dan kepadatan yang sembarangan. Walau bagaimanapun, KNNClust berhadapan dengan masalah bagi data yang mengandungi gugusan yang mempunyai saiz berbeza. Oleh itu, penyelidikan ini mencadangkan satu skema baru pengekstrakan pola yang mengintegrasikan fungsi keranjang sejajar tiga dan keranjang kepadatan purata tempatan dalam algoritma penggugusan berdasarkan kepadatan KNN. Pembaikan skema telah disahkan melalui eksperimen dengan dua senario: menggunakan data ruang-masa multidimensi sebenar dan menggunakan pelbagai data pengelasa. Empat ukuran berbeza digunakan untuk mengesahkan hasil penggugusan; indeks Dunn dan Silhouette untuk menilai kualiti, F-measure untuk menilai prestasi pendekatan dari segi ketepatan, ujian ANOVA untuk menganalisis taburan gugusan, dan masa pemprosesan untuk menguji kecekapan. Skema yang dicadangkan diulas tanda dengan kaedah penggugusan lain yang terkenal termasuk KNNClust, penggugusan Gaussian tempatan lelaran (ILGC), asas k-means, KFCM, DBSCAN dan DENCLUE. Hasil bagi data pengelasaan menunjukkan bahawa TKC menghasilkan gugusan dengan ketepatan yang lebih tinggi dan lebih berkesan dari pada kaedah penggugusan lain. Di samping itu, hasil analisis menunjukkan bahwa skema TKC yang dicadangkan telah mampu menangani data multidimensi, disahkan dengan indeks Dunn dan Silhouette yang hampir dengan nilai satu, menunjukkan keputusan yang boleh dipercayai.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DECLARATION</td>
<td>ii</td>
</tr>
<tr>
<td></td>
<td>DEDICATION</td>
<td>iii</td>
</tr>
<tr>
<td></td>
<td>ACKNOWLEDGEMENTS</td>
<td>iv</td>
</tr>
<tr>
<td></td>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td></td>
<td>ABSTRAK</td>
<td>vi</td>
</tr>
<tr>
<td></td>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
</tr>
<tr>
<td></td>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>LIST OF FIGURES</td>
<td>xii</td>
</tr>
<tr>
<td></td>
<td>LIST OF ABBREVIATION</td>
<td>xv</td>
</tr>
<tr>
<td></td>
<td>LIST OF APPENDICES</td>
<td>xvi</td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.1 Overview</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1.2 Problem Background</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>1.3 Problem Statement</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.4 Research Goal</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>1.5 Research Objectives</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1.6 Research Scopes</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>1.7 Research Road Map</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>1.8 Importance of Research</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>1.9 Organization of Thesis</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>LITERATURE REVIEW</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2.1. Introduction</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>2.2. Multidimensional Data</td>
<td>16</td>
</tr>
</tbody>
</table>
2.3. Pattern Extraction of Multidimensional Data
   2.3.1. Data Mining Functionalities
   2.3.2. Type of Data Attributes in Data Mining

2.4. Clustering for Pattern Extraction of Multidimensional Data
   2.4.1. Clustering Algorithm
      2.4.1.1. Hierarchical Clustering
      2.4.1.2. Non-hierarchical (Partitioning) Clustering
      2.4.1.3. Density-based clustering
      2.4.1.4. K-Nearest-Neighbor Density Estimation-based Clustering
      2.4.1.5. Kernel Density Estimation-based Clustering
      2.4.1.6. \( k \)-NN Kernel Density Estimation-based Clustering
   2.4.2. Challenges of Clustering Multidimensional Data
   2.4.3. Preprocessing of clustering
   2.4.4. Clustering Validation
   2.4.5. Cluster Interpretation
      2.4.5.1. Scatter-Principal Component Analysis (Scatter-PCA)
      2.4.5.2. Parallel Coordinates Plot (PCP)
      2.4.5.3. Cartographic Map

2.5. Summary

3 RESEARCH METHODOLOGY
   3.1. Introduction
   3.2. Research Design
   3.3. Experimental Design
   3.4. Clustering using Triangular Kernel Clustering
   3.5. Data of Experiment
      3.5.1 Dummy Dataset
      3.5.2 UCI Dataset
      3.5.3 Spatio-Temporal Dataset
3.5.3.1. Crime Dataset 103
3.5.3.2. Fatal Crash Dataset 106
3.5.3.3. Spatio-Temporal Data Representation 109

3.6. Summary 116

4 PROPOSED PATTERN EXTRACTION SCHEME USING TRIANGULAR KERNEL CLUSTERING 117
4.1. Challenges of Density-based Clustering 117
4.2. Triangular Kernel Function for clustering 119
4.3. Local Average Density 121
4.4. Improved Triangular Kernel Clustering Scheme (TKC) 122
4.5. Algorithm of TKC 126
4.6. Experimental Result of Discovering True Clusters on Labeled Data 127
   4.6.1 Result and Analysis of Experiment on Dummy Dataset 128
   4.6.2 Result and analysis of experiment on UCI dataset 138
4.7. Experimental Result of Discovering Spatio-Temporal Clusters 152
   4.7.1 Experiment on Crime data 152
   4.7.2 Experiment on Crash data 173
4.8. Summary 193

5 CONCLUSION AND FUTURE WORKS 196
5.1 Thesis Summary 196
5.2 Research Findings and Contributions 198
5.3 Future Works 199

REFERENCES 202
Appendices A - H 213 - 293
## LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Related works</td>
<td>6</td>
</tr>
<tr>
<td>2.1</td>
<td>Univariate kernel function (Tran et al., 2006)</td>
<td>37</td>
</tr>
<tr>
<td>2.2</td>
<td>Raw table of crime data</td>
<td>49</td>
</tr>
<tr>
<td>2.3</td>
<td>Descriptive statistic of number of burglary, number of larceny, number of mvtheft and number of robbery in two crime clusters</td>
<td>60</td>
</tr>
<tr>
<td>2.4</td>
<td>A one-way ANOVA analysis of two crime cluster</td>
<td>61</td>
</tr>
<tr>
<td>2.5</td>
<td>A Tukey post-hoc test of two crime clusters</td>
<td>61</td>
</tr>
<tr>
<td>3.1</td>
<td>Three schemes experiment</td>
<td>83</td>
</tr>
<tr>
<td>3.2</td>
<td>Purposes and approaches of experiments 3 on clustering spatio-temporal data</td>
<td>87</td>
</tr>
<tr>
<td>3.4</td>
<td>Characteristics of dummy datasets used</td>
<td>98</td>
</tr>
<tr>
<td>3.5</td>
<td>Data characteristics of four UCI datasets used</td>
<td>103</td>
</tr>
<tr>
<td>3.6</td>
<td>The Distribution of precinct, beat plus and beat of Pittsburgh</td>
<td>106</td>
</tr>
<tr>
<td>3.7</td>
<td>The example of the crash data</td>
<td>109</td>
</tr>
<tr>
<td>4.1</td>
<td>Parameters of all algorithms used to discover true clusters on the nine dummy data</td>
<td>129</td>
</tr>
<tr>
<td>4.2</td>
<td>Percentage of F-measure for the experiment results of all algorithms on the dummy datasets</td>
<td>135</td>
</tr>
<tr>
<td>4.3</td>
<td>Parameters of all algorithms used to discover true clusters on the four UCI datasets</td>
<td>139</td>
</tr>
<tr>
<td>4.4</td>
<td>Percentage of F-measure for the experiment result of all algorithms on the four UCI data</td>
<td>140</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.5</td>
<td>One-way ANOVA analysis of three classes on Iris data</td>
<td>143</td>
</tr>
<tr>
<td>4.6</td>
<td>One-way ANOVA analysis of three clusters on Iris data</td>
<td>144</td>
</tr>
<tr>
<td>4.7</td>
<td>One-way ANOVA analysis of four clusters on Wine data</td>
<td>146</td>
</tr>
<tr>
<td>4.8</td>
<td>One-way ANOVA analysis of four clusters on Glass data</td>
<td>149</td>
</tr>
<tr>
<td>4.9</td>
<td>One-way ANOVA analysis of eight clusters on Ecoli data</td>
<td>151</td>
</tr>
<tr>
<td>4.10</td>
<td>Descriptive statistics of month years in the two clusters of burglary crime</td>
<td>157</td>
</tr>
<tr>
<td>4.11</td>
<td>One-way ANOVA analysis of two clusters within burglary crime data</td>
<td>158</td>
</tr>
<tr>
<td>4.12</td>
<td>Tukey <em>post-hoc</em> of variables on burglary crime data</td>
<td>159</td>
</tr>
<tr>
<td>4.13</td>
<td>The Distribution of precinct, beat plus and beat of Pittsburgh</td>
<td>161</td>
</tr>
<tr>
<td>4.14</td>
<td>The means and the pattern names of two clusters within burglary crime data</td>
<td>163</td>
</tr>
<tr>
<td>4.15</td>
<td>Comparison of clustering results of experiment 3 on burglary crime data</td>
<td>164</td>
</tr>
<tr>
<td>4.16</td>
<td>Description of each cluster produced by TKC algorithm for cluster I</td>
<td>171</td>
</tr>
<tr>
<td>4.17</td>
<td>Description of each cluster produced by TKC algorithm for cluster II</td>
<td>173</td>
</tr>
<tr>
<td>4.19</td>
<td>One-way ANOVA analysis of two clusters within fatal accident data</td>
<td>178</td>
</tr>
<tr>
<td>4.20</td>
<td>Tukey <em>post-hoc</em> of variables on fatal accident data</td>
<td>179</td>
</tr>
<tr>
<td>4.21</td>
<td>Means and the pattern names of two clusters within fatal accident data</td>
<td>183</td>
</tr>
<tr>
<td>4.23</td>
<td>Comparison of clustering results of experiment on fatal accident data</td>
<td>184</td>
</tr>
<tr>
<td>4.24</td>
<td>Description of each cluster found by TKC algorithm for cluster II</td>
<td>193</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparison of clustering algorithms</td>
<td>197</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Research Road Map</td>
<td>10</td>
</tr>
<tr>
<td>2.1</td>
<td>The changing process of crime event in Pittsburgh</td>
<td>17</td>
</tr>
<tr>
<td>2.2</td>
<td>Spatio-temporal data structure</td>
<td>19</td>
</tr>
<tr>
<td>2.3</td>
<td>Basic $k$–means algorithm</td>
<td>30</td>
</tr>
<tr>
<td>2.4</td>
<td>Points in DBSCAN algorithm</td>
<td>38</td>
</tr>
<tr>
<td>2.5</td>
<td>DBSCAN algorithm (Tan et al., 2006)</td>
<td>39</td>
</tr>
<tr>
<td>2.6</td>
<td>DENCLUE algorithm (Hinneburg and Keim, 2003)</td>
<td>42</td>
</tr>
<tr>
<td>2.7</td>
<td>Gaussian density estimate plot for each variable in iris data</td>
<td>51</td>
</tr>
<tr>
<td>2.8</td>
<td>Scatter plot for iris data</td>
<td>52</td>
</tr>
<tr>
<td>2.9</td>
<td>Illustration of component of silhouette index.</td>
<td>55</td>
</tr>
<tr>
<td>2.10</td>
<td>Scatter-PCA approach to visualize clustering results</td>
<td>68</td>
</tr>
<tr>
<td>2.11</td>
<td>Parallel coordinates plot (Yuan et al., 2009)</td>
<td>70</td>
</tr>
<tr>
<td>3.1</td>
<td>Research design</td>
<td>77</td>
</tr>
<tr>
<td>3.2</td>
<td>Experimental Setting</td>
<td>78</td>
</tr>
<tr>
<td>3.3</td>
<td>An example of clustering results</td>
<td>82</td>
</tr>
<tr>
<td>3.4</td>
<td>Design of experiment 1 on the dummy datasets</td>
<td>84</td>
</tr>
<tr>
<td>3.5</td>
<td>Design of experiment 2 on the UCI datasets</td>
<td>86</td>
</tr>
<tr>
<td>3.6</td>
<td>Design of experiment 3 on single non-spatio-temporal attribute</td>
<td>89</td>
</tr>
<tr>
<td>3.7</td>
<td>Design of experiment 3 of PCA-TKC-based clustering approach</td>
<td>91</td>
</tr>
<tr>
<td>3.8</td>
<td>Two approaches of spatio-temporal clustering</td>
<td>94</td>
</tr>
<tr>
<td>3.9</td>
<td>KNN-Kernel Density Estimation scheme</td>
<td>96</td>
</tr>
<tr>
<td>3.10</td>
<td>Distribution of synthetic data</td>
<td>99</td>
</tr>
</tbody>
</table>
3.11 Distribution of eight dummy data
3.12 Multidimensional representation of spatio-temporal data
3.13 Multidimensional representation of crime data
4.1 Improved pattern extraction scheme triangular kernel clustering (TKC)
4.2 Clustering produced by TKC
4.3 Clustering results of experiment on the dummy data produced by TKC
4.4 Clusters produced by DBSCAN within Flame data
4.5 Comparison of precision of clustering results produced by TKC, KNNClust, ILGC, DBSCAN, DENCLUE, k-means, and KFCM
4.6 Comparison of recall of clustering result produced by TKC, KNNClust, ILGC, k-means, and KFCM
4.7 Time processing for clustering on benchmarked classification dataset
4.8 Scatter-PCA visualization of Iris data
4.9 PCP of four Iris attributes
4.10 Scatter-PCA visualization of Wine data
4.11 A PCP of 13 Wine attributes
4.12 Scatter-PCA visualization of Glass data
4.13 A PCP of nine Glass attributes
4.14 Scatter-PCA visualization of Ecoli data
4.15 A PCP of seven Ecoli attributes
4.17 Distribution of clusters on burglary crime data produced by TKC
4.18 A cartographic map of burglary crime data in Pittsburgh.
4.19 A PCP of burglary crime attributes
4.20 A time series plot of the burglary crime data
4.21 Clustering structure on projected property crime data produced by PCA-TKC
4.22 A cartographic map of property crime data in Pittsburgh.
4.23 A time series plot of the property crime data.
4.24 Dunn index versus number of cluster for the property crime data using ILGC and TKC.

4.25 Silhouette index versus number of cluster for the crime property data using ILGC and TKC.

4.26 The structure of crime property clusters found by LTKC.

4.27 Average total number of crime property within each cluster.

4.28 Single Gaussian scatter diagram for fatal accident data.

4.29 Scatter-PCA of fatal accident clusters.

4.30 A cartographic map of fatal accident clusters in USA obtained by TKC.

4.31 A cartographic map of fatal accident clusters in USA obtained by ILGC.

4.32 A PCP of Fatal accident attributes.

4.33 A time series plot of the fatal accident data.

4.34 Clustering structure on projected fatal crash data produced by PCA-TKC.

4.35 A cartographic map of fatal crash data in USA.

4.36 A time series plot of the fatal crash data.

4.37 Dunn index versus number of cluster for the fatal crash data using ILGC and TKC.

4.38 Silhouette index versus number of cluster for the fatal crash data using ILGC and TKC.

4.39 The structure of fatal crash clusters produced by TKC algorithms.

4.40 Average number of fatal crashes within each cluster.

4.41 A PCP of four Iris attributes.

4.42 A PCP of four Iris attributes with the attributes reordered to emphasizes similarities and dissimilarities of groups.
LIST OF ABBREVIATION

TKC - Triangular Kernel Clustering
ILGC - Iterative Local Gaussian Clustering
KNNClust - Kernel-Nearest-Neighbor-based Clustering
DBSCAN - Density Based Spatial Clustering and Application of Noise
DENCLUE - DENsity based CLUstEring
KFCM - Kernel Fuzzy C-Means
SGS - Single Gaussian Scatter
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>List of publications</td>
<td>214</td>
</tr>
<tr>
<td>B</td>
<td>Flowchart of other clustering algorithms for comparison analysis</td>
<td>217</td>
</tr>
<tr>
<td>C</td>
<td>Experimental results of dummy data sets</td>
<td>219</td>
</tr>
<tr>
<td>D</td>
<td>SGS diagram of Iris data</td>
<td>225</td>
</tr>
<tr>
<td>E</td>
<td>Experimental results on UCI datasets</td>
<td>226</td>
</tr>
<tr>
<td>F</td>
<td>SGS diagram of Burglary crime data</td>
<td>235</td>
</tr>
<tr>
<td>G</td>
<td>Cartographic map of spatio-temporal clusters</td>
<td>241</td>
</tr>
<tr>
<td>H</td>
<td>Data of experiment</td>
<td>244</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Overview

For decades, large quantity of raw data has been collected from different application domains, such as business, science, telecommunication and health care systems. The volume of available data has increased exponentially because of the extensive use of electronic data gathering devices and technological sophistication (Golmah and Parvizian, 2010). The increasing volume, variety and velocity of data available from new digital sources such as social networks and point-of-sale remote sensing devices; in addition to traditional sources such as sales data and market research. The data are of course not only to be collected but also are necessary to analyze the content therein, whether there is useful information hidden in the data or vice versa. Thus, the difficulty is how to analyze these vast quantities of data to extract the meaningful insights, and use them effectively to support business purpose such as improvement of government services, increasing products and advanced customer experience.

The current data which are commonly stored in the database or data warehouse and mostly resulted by technology advances in many researchers usually contains many attributes indeed the single attribute that refers to multidimensional data (Seo, 2005). Some examples of collected multidimensional data are clinical data that stored the histories of clinical activities in a hospital (Tsumoto et al., 2012),
video data that stored human motion (Zhou et al., 2013) and UCI data (Frank and Asuncion, 2010).

A multidimensional data type which has also received much attention is the spatio-temporal data. This is due to the emergence of new applications, such as traffic control systems and monitoring systems security conditions of a particular area that use and capture the multi-dimensional data containing space and time aspect. Spatial-temporal database that contains time and space aspects is a database which stores the temporal, spatial and relevant attribute geographic entities (Ping et al., 2005). The spatial-temporal database was developed with the requirement of historical information to monitor and analyze changes over time. It specifically deals with geometry changing over time.

However, the recent advancement in positioning technology and location-based services has led to the rapid accumulation of multidimensional data (Hsu et al., 2008). Furthermore, various fields of applications have utilized collected multidimensional data, not only for storing data but also for the purpose of supporting decision making. For instance, in security-related application (Townsley et al., 2000), it is a requirement for the police department to firstly identify the pattern of crimes offense for formulating more effective preventive strategies. Another example is, to improve traffic safety, traffic agencies and public at large need to have knowledge of accident patterns since accidents are not randomly scattered along the road net (Beshah and Hill, 2010). Therefore, extracting required patterns from huge amount of multidimensional data is now crucial and an area of interest to the research (Dermoudy et al., 2009; Tabakov and Duffy, 2010).

Data mining provides a solution to generate compact and rich semantics representations of raw data, called patterns (Golmah and Parvizian, 2010). Patterns are compact since they summarize in some degree of the amount of information contained in the original raw data. Meanwhile, patterns are contently rich in semantics by means they reveal new knowledge hidden in the huge amount of raw
Data mining techniques have been applied by many researchers to extract useful patterns in a wide variety of application, such as planning and scheduling (Tsumoto et al., 2012), sales and marketing (Lawrence et al., 2001), and finance (Hadavandi et al., 2010).

However, many data mining tasks differ when applied for different purposes for different types of data. Clustering is the process of grouping unlabelled large data sets according to their similarity. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Conversely, clustering is a powerful exploratory technique for extracting the patterns hidden in the multidimensional data (Dermoudy et al., 2009).

The general clustering methods have difficulties for analyze the hidden pattern in the multidimensional data due to the number of dimension which is high (Hu et al., 2007). The multidimensionality of the data has been challenging to researchers in many disciplines due to the difficulty in comprehending more than three dimensions to discover clusters (Sembiring et al., 2011b). This difficulty is so well recognized that it has a provocative name: “the curse of high dimensionality.” (Seo, 2005).

In addition, due to the explorative and descriptive nature, intelligible representation and visualization of the found patterns is essential for the successful mining process (Ayramo et al., 2009). Visualization methods can be used to help analysts pick out complex patterns visually, propose explanations and generate hypotheses for further analysis, and present patterns in an easy-to-understand form (Guo et al., 2005). Data with huge in size and high in dimension, such as multidimensional data, are also a big challenge for researchers in the visualization field, in effort to provide powerful algorithms and tools. Various techniques and approaches have been recommended to explore the visualization of multidimensional data, such as parallel coordinate plot (Inselberg and Dimsdale, 1990).
1.2 Problem Background

Clustering techniques for pattern extraction of multidimensional data is a very promising subfield of data mining since increasing large volumes of multidimensional data are collected and need to be analyzed (Birant and Kut, 2007). The pattern extraction process for multidimensional data is more complex than conventional data, because the use of multidimensional data will result in more noise, complex data, and the possibility of unconnected data entities (Sembiring et al., 2011b). Besides that, the current collected multidimensional data are generally large in size and high in dimension. This, however, comes with two disadvantages, a reasonable response time and memory space.

There are several major challenges that are commonly associated with multidimensional data. Firstly, the dimension of the multidimensional data can cause serious problems for most analysis methods due to the curse dimensionality problem. One typical problem to address it is that it is unlikely for all variables to interrelate meaningfully. Analysts need to find interesting subspaces (subsets of variables) out of a combinatorial explosive number of possible subspaces in a high-dimensional dataset. Secondly, even when a selected multivariate data space is given as the starting point for analysis (which may be a subspace from a higher-dimensional dataset), it is still a challenge to discover the hidden relationships among those variables, as potential patterns may take various forms, linear or non-linear. Thirdly, the attribution of meaning to discover patterns typically requires the input from experts who have domain knowledge and the subsequent presentation of the patterns identified to a broader audience (e.g., other experts who will try to replicate the results, or policy makers who need to act on the results). Fourthly, large and high-dimensional datasets demand that all analysis methods are computationally efficient in terms of execution time (Guo et al., 2005).

However, this study focused on two crucial challenges in multidimensional data clustering: firstly, the exploration of efficient methods with minimal
requirement of input parameters due to the large amount of multidimensional data and the complexity of multidimensional data types, data representation (Yao, 2003) since clusters within multidimensional generally have different size, shape and densities; and secondly, developing the mechanisms to test, validate and interpret the clustering results to reconcile discrepancies in the data.

Recently, several clustering techniques have been introduced and applied to the multidimensional data, such as $k$-means (Anderson, 2009; Golob and Recker, 2004; Shekhar et al., 2001), hierarchical clustering (Skyving et al., 2009) and Support Vector Machine-based approach (SVM) (Chang et al., 2005). Since there are lacks of valid statistical evaluation methods meanwhile the results of the hierarchical cluster analysis are subject to interpretation by the investigator. $K$-means technique is probably the most popular and simplest solution for clustering the spatio-temporal data, but still, it has a problem on determining the proper number of clusters and random issue.

Using probabilistic similarity function, density based clustering algorithms determine the number of clusters automatically (Hammouche and Postaire, 2008). Density-based clustering algorithm uses local cluster criterion in which the clusters are defined as region in data space whose objects are dense, and clusters are separated from one another with low-density region (Zhang et al., 2013).

Previous researchers have been introduced and proposed many density-based clustering algorithms, such as density based spatial clustering of application with noise DBSCAN-based methods (Birant and Kut, 2007), density-based clustering DENCLUE (Hinneburg and Keim, 1998), and nearest neighbor-based approach (Ertoz et al., 2002; Steinbach et al., 2003; Wang et al., 2006). Although DBSCAN capable to find clusters of arbitrary shapes, it has difficulties to determine two input parameters, Eps and MinPts addition has problem on finding cluster of varying densities (Birant and Kut, 2007). DENCLUE covered the problem of DBSCAN to discover clusters of arbitrary shapes, densities, and size. However, DENCLUE faced
the problem on data with high-dimension and had less efficiency even computational expensive. Furthermore, SNN could handle the data with high-dimension; nevertheless SNN requires large memory space and determining the difficult density threshold.

$k$NN-Kernel density-based ($k$NN-Kernel-based) clustering algorithms combined $k$-nearest neighbor density estimation and kernel density estimation to cluster the multidimensional data. For instance, KNNClust introduced by (Tran et al., 2006) was proposed to solve the problems of existing density-based clustering algorithms on determining smoothing parameters. It has the capability to determine the number of clusters automatically within multidimensional data and to discover cluster with arbitrary shape and densities.

However, KNNClust is faced problem on data that contained cluster with different size. In addition, it is necessary to investigate the proper kernel function since in $k$NN-Kernel-based clustering algorithms, there are various kernel functions that can be used for kernel density estimation (Webb, 2002), such as, Gaussian, triangular, and rectangular although the most commonly used kernel functions are triangular and Gaussian (Tran et al., 2006). Therefore, Table 1.1 summarized the previous works that related to this study.

<table>
<thead>
<tr>
<th>Clustering Techniques (Author / Year)</th>
<th>Summary</th>
<th>Limitation / Future Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBSCAN - Density Based Spatial Clustering of Application of Noise (Ester et al., 1996)</td>
<td>Capable to find cluster of arbitrary shapes and contains noise.</td>
<td>Problem on finding clusters of varying densities and determining two input parameters.</td>
</tr>
<tr>
<td>DENCLUE - DEnsity CLUstEring (Hinneburg and Gabriel, 2007)</td>
<td>Proficient to find cluster of arbitrary shapes, densities and sizes.</td>
<td>Less efficiency and computational expensive, in addition has problem on data with high-dimension.</td>
</tr>
</tbody>
</table>

Table 1.1: Related works
<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNN Shared Nearest Neighbor (Ertoz et al., 2003)</td>
<td>Able to find cluster in high-dimension data</td>
<td>Computational expensive and required large memory space, in addition has problem on determining require density threshold.</td>
</tr>
<tr>
<td>KNNClust – Kernel Nearest Neighbor Clustering (Tran et al., 2006)</td>
<td>Pioneer to combine K-nearest neighbor density estimation and kernel density estimation.</td>
<td>Absence of clustering results validation process. Needed further investigation, such as: comparison to other clustering approaches and applying high dimensional datasets such as spatio-temporal data.</td>
</tr>
<tr>
<td>ILGC – Iterative Local Gaussian Clustering (Wasito et al., 2007)</td>
<td>Using a non-parametric density-estimation-based approach called iterative local Gaussian clustering (ILGC) to identify clusters of expressed genes.</td>
<td>Absence of clustering results validation process. Needed further investigation, such as: comparison to other clustering approaches, applying high dimensional datasets such as spatio-temporal data, or using other kernel functions.</td>
</tr>
<tr>
<td>Hybrid- SOM (Guo et al., 2008)</td>
<td>They proposed integrated approaches including SOM multivariate analysis, multidimensional visualization, multivariate mapping and human interaction, for detecting spatial patterns.</td>
<td>The addition of new components is needed for further development, such as considering other clustering methods, temporal analysis component and validation measurement of clustering results.</td>
</tr>
</tbody>
</table>
1.3 Problem Statement

Two main problems that will be addressed in this research are:

1. Need of improved clustering technique to extract patterns of multidimensional data.

   Existing density-based clustering algorithms have been proven outperformed other non-density-based clustering algorithms in clustering homogenous multidimensional data, but not arbitrary multidimensional data dealt in this research. However, it is necessary to address large size of multidimensional dataset.

2. Absence of technique for validating and interpreting the clustering results.

   Furthermore, in order to achieve clustering results with high-quality, it is required to validate the clustering results using appropriate validation measurements. In addition, it is essential to provide efficient visualization approaches to interpret the clustering results in informative graphical display for further analysis.

1.4 Research Goal

The aim of the research is to propose effective and efficient pattern extraction approach for mining multidimensional data using improved \( k \text{-NN-Kernel} \) density-based clustering algorithm compared to other widely-known clustering algorithms, including \( k \)-Means, KFCM, KNNClust, ILGC, DBSCAN and DENCLUE algorithms.
1.5 Research Objectives

The objectives of the research are:

1. To propose triangular kernel nearest neighbor based clustering scheme for better pattern extraction of multidimensional data.
2. To validate the pattern extraction scheme using index validation and 2-D visualization techniques onto various multidimensional datasets.

1.6 Research Scopes

The scopes of the research include:

1. Noise-less multidimensional data, such as crime data, and fatal crash data are used in this research.
2. Benchmarked classification data is utilized for evaluating the accuracy of proposed scheme.
3. Data preprocessing is applied to clean the data from its class label information.
4. The proposed algorithm is compared against other existing well-known clustering techniques such as $k$-Means, KFCM, KNNClust, ILGC, DBSCAN and DENCLUE algorithms.
5. Four validation techniques are used, namely Silhouette index, Dunn index, $F$-measure, and statistical analysis ANOVA, are to be used to analyze the clustering results.
6. 2-D visualization techniques are used for interpreting the pattern extracted.
1.7 Research Road Map

In order to discover patterns within multidimensional data, some problems were faced, such as problem on finding patterns with arbitrary densities. However, this study proposed an improved pattern extraction scheme using triangular kernel clustering. Figure 1.1 shows the research road map of this study.

First, developing a clustering algorithm based on density estimation approach since density-based clustering algorithms has capability to determine number of cluster automatically using local criterion. Second, a general framework for pattern extraction of multidimensional data is proposed to handle the difficulties of multidimensional data that contained cluster with arbitrary shape, size and densities. In addition, the density-based clustering algorithms are necessary to improved (for instance by using triangular kernel function) to provide the demand of clustering method that is efficient at execution time. Therefore, an improved density-based clustering algorithm is needed to be evaluated by utilizing a number of benchmarked datasets. Furthermore, the clustering results are crucial to be validated using some validation measurement and represented in informative view.

![Figure 1.1 Research Road Map](image-url)
1.8 Importance of Research

With enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis, and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Due to the availability gigantic of multidimensional data and the interest to extract knowledge further for decision making in various fields, clustering multidimensional data has great challenges in data mining research. It is important to develop efficient and reliable clustering method, especially to handle difficulties in determining the proper number of clusters, therefore, in this study; kernel nearest neighbor based clustering was proposed. Even if the database technology plays a central role in the development and deployment of application for multidimensional data, data mining capabilities will become increasingly important to discover and extract the information from multidimensional data (Hsu et al., 2008).

Another issue in clustering multidimensional data is the complexity of the data structure. The clusters involved in multidimensional data have different and arbitrary shapes and sizes. It would certainly be challenged to develop powerful clustering algorithm which will be solved using triangular kernel clustering approach.

In the mean time, such approach creates a look-up table save the distances between all pairs of data points. With the aid of the look-up table, the distances between all pairs of data points need to be evaluated only once throughout the clustering process. It can reduce time-consumption caused by repeated computation of the distance between every data point.

In this study, the performance between the proposed algorithms, KNNClust, ILGC, K-Means clustering, DBSCAN, DENCLUE, and KFCM, was analyzed, such
that we could determine which method is better for clustering spatio-temporal data. It is important to identify most appropriate technique for future research, which can be implemented in real world situation.

1.9 Organization of Thesis

The thesis consists of five chapters. The structure of thesis was given as follow:

Chapter 1: Introduction - Explains an overview of the background of the study, development of techniques and methods used in clustering multidimensional data and the common problems that are usually encountered in clustering multidimensional data. Also include the aim, problem statement, the objective, research scope, and general methodology.

Chapter 2: Literature review - This chapter explored the concept of multidimensional data, clustering technique, cluster validation, cluster visualization and cluster interpretation. It also contained the reviews of related previous works for clustering multidimensional data.

Chapter 3: Research Methodology - This chapter described the approach taken to solve extracting pattern of multidimensional data through clustering and detail description of the proposed approach. In addition, the experimental procedure and schemas were also discussed in this chapter.

Chapter 4: Proposed Pattern Extraction Scheme using Triangular Kernel Clustering - This chapter explained the basic theory and detailed description of the proposed algorithm, triangular kernel clustering, called TKC. Also included was the
evaluation of the accuracy of the proposed algorithm through discovering the clusters within the classification datasets. This chapter also explained the application of TKC on two multi-dimensional geospatio-temporal datasets. The detail of the experimental result on each data was also discussed in this chapter.

Chapter 5: Conclusions and Future Work - This chapter provided the summary of the research, the contribution of the work and recommendation for future studies.
REFERENCES


Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making.


Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6, 2(11), 559-572.


Speech and Image Processing Unit. (2012). Clustering datasets (Publication., from School of Computing, University of Eastern Finland: cs.jouensuu.fi


