DENSITY BASED SUBSPACE CLUSTERING: A CASE STUDY ON PERCEPTION OF THE REQUIRED SKILL

RAHMAT WIDIA SEMBIRING

Thesis submitted in fulfilment of the requirements
for the award of the degree of
Doctor of Philosophy in Computer Science

Faculty of Computer Systems & Software Engineering
UNIVERSITI MALAYSIA PAHANG

JANUARY 2014

ABSTRACT

This research aims to develop an improved model for subspace clustering based on density connection. The researches started with the problem were there are hidden data in a different space. Meanwhile the dimensionality increases, the farthest neighbour of data point expected to be almost as close as nearest neighbour for a wide range of data distributions and distance functions. In this case avoid the curse of dimensionality in multidimensional data and identify cluster in different subspace in multidimensional data are identified problem. However develop an improved model for subspace clustering based on density connection is important, also how to elaborate and testing subspace clustering based on density connection in educational data, especially how to ensure subspace clustering based on density connection can be used to justify higher learning institution required skill. Subspace clustering is projected as a search technique for grouping data or attributes in different clusters. Grouping done to identify the level of data density and to identify outliers or irrelevant data that will create each to cluster exist in a separate subset. This thesis proposed subspace clustering based on density connection, named DAta MIning subspace clusteRing Approach (DAMIRA), an improve of subspace clustering algorithm based on density connection. The main idea based on the density in each cluster is that any data has the minimum number of neighbouring data, where data density must be more than a certain threshold. In the early stage, the present research estimates density dimensions and the results are used as input data to determine the initial cluster based on density connection, using DBSCAN algorithm. Each dimension will be tested to investigate whether having a relationship with the data on another cluster, using proposed subspace clustering algorithms. If the data have a relationship, it will be classified as a subspace. Any data on the subspace clusters will then be tested again with DBSCAN algorithms, to look back on its density until a pure subspace cluster is finally found. The study used multidimensional data, such as benchmark datasets and real datasets. Real datasets are from education, particularly regarding the perception of students' industrial training and from industries due to required skill. To verify the quality of the clustering obtained through proposed technique, we do DBSCAN, FIRES, INSCY, and SUBCLU. DAMIRA has successfully established very large number of clusters for each dataset while FIRES and INSCY have a high failure tendency to produce clusters in each subspace. SUBCLU and DAMIRA have no un-clustered real datasets; thus the perception of the results from the cluster will produce more accurate information. The clustering time for glass dataset and liver dataset using DAMIRA method is more than 20 times longer than the FIRES, INSCY and SUBCLU, meanwhile for job satisfaction dataset, DAMIRA has the shortest time compare to SUBCLU and INSCY methods. For larger and more complex data, the DAMIRA performance is more efficient than SUBCLU, but, still lower than the FIRES, INSCY, and DBSCAN. DAMIRA successfully clustered all of the data, while INSCY method has a lower coverage than FIRES method. For F1 Measure, SUBCLU method is better than FIRES, INSCY, and DAMIRA. This study present improved model for subspace clustering based on density connection, to cope with the challenges clustering in educational data mining, named as DAMIRA. This method can be used to justify perception of the required skill for higher learning institution.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATION

Abbreviation	Description
ACM	Association for Computing Machinery
ANN	Artificial Neural Network
ASCLU	Alternative Subspace Clustering
CCA	Canonical Correlation Analysis
CI	Cluster Initialization
CKNN	Continuous kernel Neural Network
CLIQUE	Clustering in Quest
CSV	Comma Separated Values
DAMIRA	DAta MIning subspace clusteRing Approach
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNF	Disjunctive Normal Form
DSM	Data Stream Mining
EM	Expectation–Maximization
EDM	Educational Data Mining
Eps	Epsilon
FastICA	Fast Independent Component Analysis
FIRES	FIlter REfinement Subspace clustering
FSKKP	Fakulti Sains Komputer dan Kejuruteraan Perisian
FSMKNN	Fuzzy Similarity Measure and kernel Neural Network
GKM	Generalized k-Mean
HLI	Higher Learning Institution
HMM	Hidden Markov Models
ICA	Independent Component Analysis
ID3	Iterative Dichotomiser 3
IEEE	The Institute of Electrical and Electronics Engineers
INSCY	Indexing Subspace Clusters with In-Process-Removal of Redundancy
ISODATA	Iterative Self Organizing Data Analysis Technique
IT	Information Technology
KDD	Knowledge Discovery from Data

Abbreviation	Description
k-NN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LDR	Local Dimension Reduction
LOF	Local Outlier Factor
LSI	Latent Semantic Indexing
MinPts	Minimal Points
MrCC	Multi-resolution Correlation Cluster detection
NN	Neural Network
OPTICS	Ordering Points To Identify the Clustering Structure
OSCLU	Orthogonal Subspace CLUstering
PCA	Principal Component Analysis
PDI	Predetermined Decision Itemset
PIPA	Protect IP Act
PLS	Partial Least Square
PLSA	Probabilistic Latent Semantic Analysis
PROCLUS	PROjected CLUStering
RFM	Recency, Frequency, and Monetary
SC2D	Subspace Clustering with Dimensional Density
SIT	Student Industrial Training
SOM	Self Organizing Map
SOPA	Stop Online Piracy Act
SRM	Structural Risk Minimization
SSDR	Semi-Supervised Dimension Reduction
SUBCLU	density connected SUBspace CLUstering
SVD	Singular Value Decomposition
UCI	University California Irvine
UMP	Universiti Malaysia Pahang
VB	Vapnik–Chervonenkis-Bound

CHAPTER 1

INTRODUCTION

This part describes the background of the study, followed by a brief description of multidimensional data mining, cluster analysis, dimension reduction, outlier's detection, and subspace clustering. Furthermore, the problem statement, purpose, and scope of research also described in this chapter. Thesis statement and the contribution of this research explained, and finalized with the structure and a summary of the thesis.

1. 1 BACKGROUND

As an important part in information technology, data has been generated on a daily basis. The amount of data that has been generated and has improved rapidly. Not only may the quantity that needs to be handled carefully, the complexity of generated data also cause a number of difficulties for people that will work on it. The obvious reason of such complexity is that the number of dimensions of data has increased many folds. In other words, due to such multidimensional data—we call it high dimensional data—the information retrieval from it becomes very challenging.

It is often the case that high dimensional data will generate similar values or attribute. High dimensional data will normally be produced when there is non linear mapping of a point as D variable $y_1, y_2, y_3, \dots, y_D$ into x as output target (Figure 1.1).

Every point is in D dimension vector, each axes variable divided in 4 then should be $y_1, 1; y_1, 2\ ;\ \ldots;\ y_1, 4$, same state notion in y_2 and y_3.

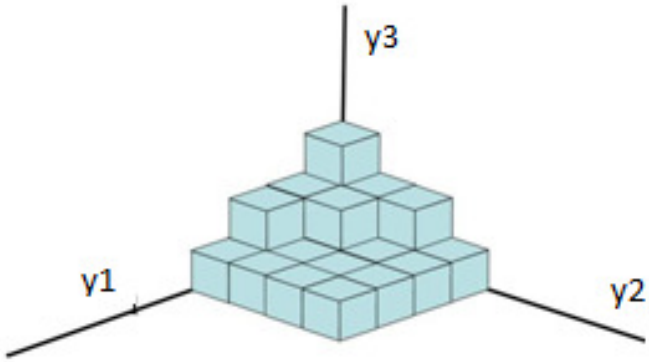

Figure 1.1. The high dimensional data

If there is data with a number of dimension D, and each dimension divided into I intervals, then there are I^Delement of data become $D+1$, then element of data become $I^{D*I} = I^{D+1}$. This shows that the amount of data increase exponentially with the increase of the dimension of the data. In fact, in the era of massive automatic data collection for digital libraries, image, medical record, growth of computational biology, e-commerce applications and the World Wide Web, the amount of data continues to grow exponentially.

The benefit of high dimensional data are: finding objects having particular feature values, pairs of objects from the same set or different sets that are sufficiently similar or closest to each other. It has become much cheaper to gather data than to worry much about what data to gather.

Expanding dimensions of used data will increase needs for data mining. The process of running an interactive data mining is needed because most of the results did not match with analyst expectations, hence, resulting in the need to redesign the process. The main idea in data mining is improving processes data through using of tools, automation the composition of data mining operations and building a methodology

(Yang and Xindong, 2006). Three fundamental elements in data mining are classification, clustering, and outlier detection.

Clustering is one of the data mining tools widely used has long been used in psychology, and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, but significant issues still remain (Steinbach et al., 2003). This tools to divide data into meaningful or useful clusters; most of the common algorithms fail to generate meaningful results because of the inherent of the objects. High dimensional data, spread data distributions, and density problem tend to fail in traditional data clustering. Algorithms often fail to detect meaningful clusters and have difficulties in finding clusters and handling outlier. In such way, that object in the same cluster are very similar, and the other objects in a different cluster are quite dissimilar (Han and Micheline, 2006).

These new dimensions can prove difficult to interpret, making the results hard to understand. Projected clustering, or subspace clustering, addresses the high dimensional challenge by restricting the search in subspaces of the original data space (Xu and Donald, 2009). Subspace clustering is an extension of traditional clustering, based on the observation that different cluster, group of data points, and may exist in the subspace within datasets.

1. 2 HIGH DIMENSIONAL DATA

Due to growth of using data in the real world, many variables and attribute needs to be explored. Ddata became higher dimensional and should have many space, cluster are often hidden in a subspace of the attribute.

Some recent works discusses clustering around higher-dimensional data. Practitioners of cluster analysis usually are not able to use the approach that suits their purpose best, but only those approaches that are available inconveniently accessible statistical or data mining software systems (Kriegel et al., 2009a), the density of the points in the 3-dimensional space are too low to obtain good clustering (Agrawal et al., 2005).

Another research discussed an approach for future work is the development of an efficient index structure for partial range queries (Kailing et al., 2004) and projected clustering for discovering interesting patterns in subspaces of high dimensional data spaces (Aggarwal et al., 1999). The use of rank-bases similarity measure can result in more stable performance than their associated primary distance measures (Houle et al. 2010). High-dimensional data input will increase the size of the search exponentially, in general classification will increase the likelihood of finding false or invalid (Maimon and Lior, 2005).

1. 3 THE CHALLENGES

The difficulty of deciding what constitutes a cluster, often allow clusters to be nested (Steinbach et al., 2003), the most reasonable interpretation of the structure of these points are that there are two clusters, each of which has three sub-clusters, while large databases are required to store massive amounts of data that are continuously inserted and queried (Khalilian and Musthapa, 2012).

The curse of dimensionality in genomic research can be grouped into three categories: filtering, wrapper and embedded methods (Liang and Klemen, 2008) . The curse of dimensionality is the apparent intractability of integrating a high dimensional function (Donoho, 2000), while the expected gap between the Euclidean distance to the closest neighbour shrinks as the dimensionality grows (Wang, 1999). High-dimensionality has significantly challenged traditional statistical theory. Many new insights need to be unveiled and many new phenomena need to be discovered (Fan and Li, 2006). Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty in distinguishing distances between data points (Tomasev et.al. 2013). The provided data with many dimensions it is important to let an analyst see and compare all these different dimensions, narrow down and investigate specific offices and the computers within those offices (Chen et.al. 2012).

There are some challenges in high dimensional data mining. There are the curse of dimensionality, multi cluster membership, noise tolerance and the potential implementation in educational data mining.

1. 3. 1 The Curse of Dimensionality

The curse of dimensionality (Bellman, 1957) referred to the impossibility of optimizing a function of many variables by a brute force search on a discrete discrete multidimensional grid (Steinbach et al., 2003). Figure 1.2 shows how the curse of dimensionality appears, in (A) number of subset = 6 is analyzed as single variant. Then, in (B) 12 subset was analyzed, where a single genetic variant (in A) and a single nutritional factor variable also have been analyzed. After that, two genetic variants and a single nutritional factor are analyzed in 36 number of a subset in (C).

Contingency table of two genetic variants and another two factors is analyzed too in 72 subsets (Cocozza, 2007). Due to this fact, dimensionality constitutes a serious obstacle to efficient data mining algorithms, while the number of records goes beyond a modest size of 10 attributes cannot provide any meaningful results (Maimon and Lior, 2005).

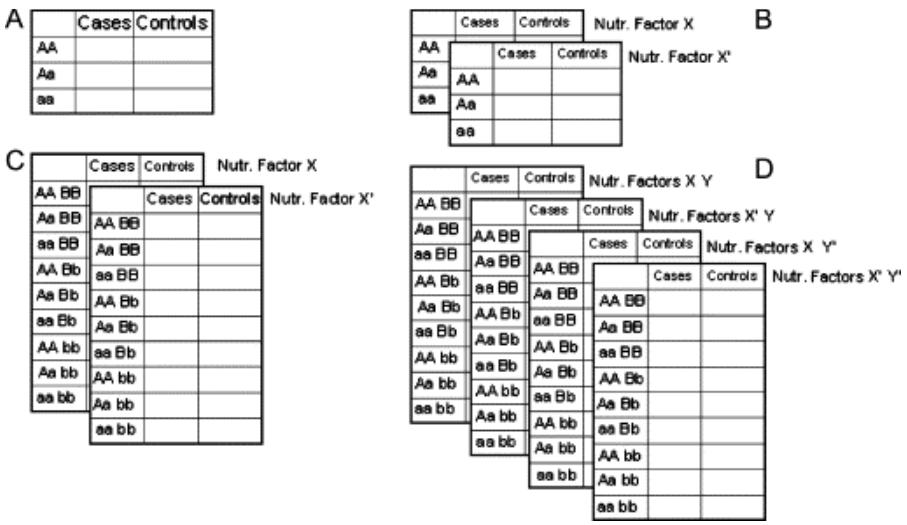

Figure 1.2. The curse of dimensionality

Figure 1.3 show another view of curse of dimensionality. In 1-dimension there are 10 position, while have 2-axis, will have 2-dimension with 100 positions. In 3-axis the position will have 1000 positions, so this called as the curse of dimensionality.

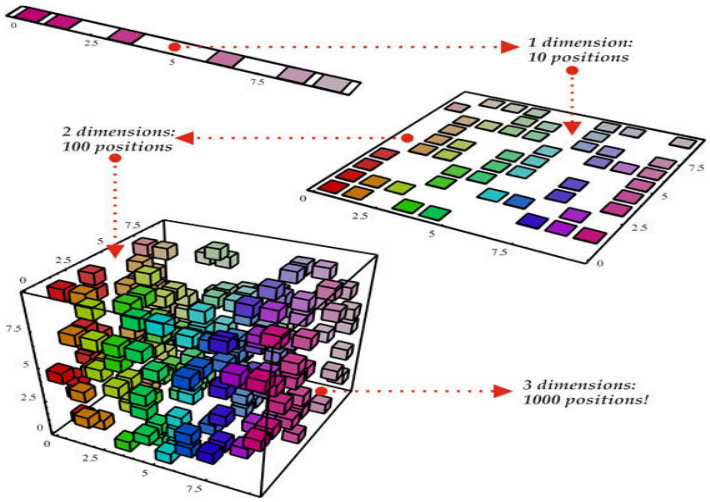

Figure 1.3. Another curse of dimensionality

1. 3. 2 Multi Cluster Membership

Various clustering methods have been studied; several detailed surveys of clustering methods also have been carried out. Most clustering algorithms do not work efficiently for a high dimensional space. In higher dimensions, the possibilities of some points are far apart from one another. Therefore, a feature selection process often precedes clustering algorithms. Objective feature selection is to find the dimensions in which the dots could be correlated. Pruning or removing the remaining dimension can reduce the confusion of data. The use of a traditional feature selection algorithm is to choose a particular dimension first before it can cause loss of information. However, cut too many dimensions will cause loss of information.

For example, describe two different cross-sections are projected to a set point in the space of three dimensions. There are two patterns in the data. The first pattern corresponds to a set of points close to each other on the xy plane (Figure 1.4), while the second pattern corresponds to a set of points close to each other on the field of xz

(Figure 1.5). Features a traditional selection does not work in this case, every dimension relevant to at least one cluster. At the same time, full-dimensional clustering space will no`t find two patterns, because each and every one of them are spread along one-dimensional.

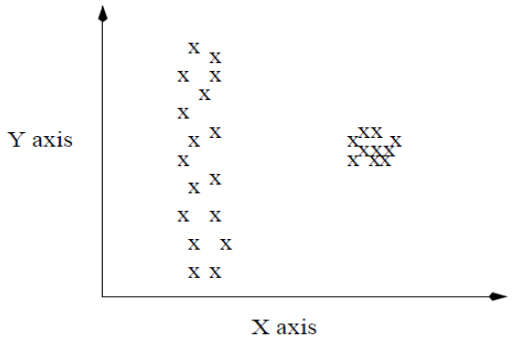

Figure 1.4. Cross section on X-Y axis

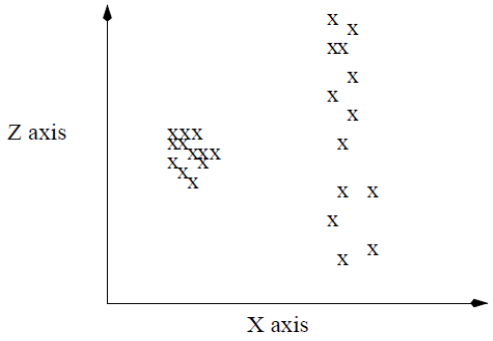

Figure 1.5. Cross section on X-Z axis

1. 3. 3 Noise Tolerance

Clustering can applied to several problems, like patient segmentation, customer classification, stock prediction, and analysis of trends. However, existing algorithms often do not achieve maximum results in overlap dimensions.

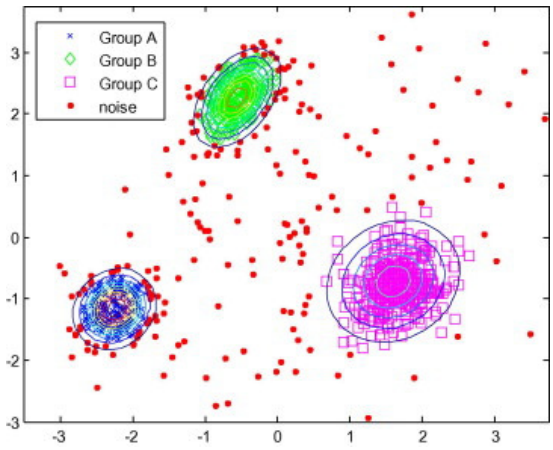

Figure 1.6. Noise in clustering

Source : Lee and Chi, 2013

On many dimensions of data, not every dimension has a relation with the resulting clusters, as shown in Figure 1.6, the clustering result for one case of randomly generated objects, red dots indicate objects declared as noises While blue, green, and pink marks are objects declared as belonging to groups A, B and C (Lee and Chi, 2013).

The most effective ways to solve this problem is determining the nearest cluster that has a relationship with these dimensions. The main concern of clustering is a setup into multidimensional space, and then determines the points into certain groups, so that each point is close to its cluster. It could also happen there are some points that do not include any group, referred as outliers.

1. 3. 4 Applicability to Educational Application

The use of data mining currently used in various fields. In Higher Learning Institution (HLI), data mining is used extensively to see the potential success of a student; data mining is a powerful tool for academic issues (Luan, 2002). Data Mining can be applied to the business of education, for example to find out which alumni are likely to make larger donations.

In the recent years, an increasing use of data mining specialization for research in the field of education, called as Educational Data Mining (EDM). EDM is defined as the field of data mining investigations in the field of education, which is devoted to developing methods or implements existing algorithms in the educational setting, such as understanding the background of students and to predict their success. (Baker, 2005).

The use of large scale data and development of information technology, create use of huge size of information and knowledge. Knowledge is very important asset in most sectors of human life, educational markets are becoming global as HLI attempt to internationalisation of the curriculum. HLI has to adjust and develop strategies to respond to changes in technologies and increasing demands of stakeholders (Baker and Yacef, 2009).

Educational data mining (EDM) is an emerging research, often used at universities, with many aspects of educational object. The process of admission of new students, students behaviour, and determine the appropriate course of study for students are frequent task in EDM. Another issue is how to precise the mapping of the competencies of students and college graduates. Competency mapping will provide a greater opportunity to encourage graduates to get jobs faster according to their competence. Many researches include case studies to measure accuracy the legal obligation that universities have to provide students with the necessary support to evaluate their students (Dekker et.al., 2009.; Pechenizkiy et.al., 2008.; Hamalainen et.al., 2004.; Hanna et.al., 2004).

The applications of EDM also implement in e-learning and online courses by implementing a model to predict academic dismissal and also GPA of graduated students (Nasiri, et.al., 2012). Some researchers have begun to study various data mining methods for helping instructors and administrators to improve the quality of e-earning systems (Romero and Ventura, 2007). There are also many other web based education such as well-known learning management systems (Pahl and Donnellan, 2003), web-based adaptive hypermedia systems (Koutri et.al., 2005) and intelligent tutoring systems (Mostow and Beck, 2006).

1. 4 PROBLEM STATEMENT

There are four main problems for clustering in high-dimensional data (Kriegel et al., 2009a): curse of dimensionality, distance of dimensions grows, different clusters will be found in different subspaces, and some attributes are correlated. Clustering algorithms measure the similarity between data points by considering all features/attributes of a data set in high dimensional data sets tend to break down both in terms of accuracy, as well as efficiency. Meanwhile, when the dimensionality increases, the farthest neighbour of data point expected to be almost as close as nearest neighbour for a wide range of data distributions and distance functions.

Due to this effect, the concept of proximity, and subsequently the concept of a cluster are seriously challenged in high dimensional spaces, thus an increasing number

of features/attributes can be automatically measured. However, not all of these attributes may be relevant for the clustering analysis. The irrelevant attributes may in fact "hide" the clusters by making two data points belong to the cluster look as dissimilar as an arbitrary pair of data points.

When rapid using of information technology, increasing capacity of storage media, huge network connections will lead to increase various data, and influence storing, processing, transmission and implement multidimensional data (Berka, 2009). Motivated by these observations, subspace is considered when subset of attribute or data points belongs to different clusters in different subspaces.

The purpose of this study to explore subspace clustering method, and define each item of subspace, values to design of data mining in multidimensional data. The research started with the question on how to cluster data hidden in a different space, and use it in educational data mining. This inquiry led toward understanding that in high dimension data, distance becomes less precise as the number of dimensions grows, different clusters might found in different subspaces, and given a large number of attributes likely that some attributes correlated. The particular focus was oriented toward values in density connection, because in density-based clustering can apply to calculate the distance to the nearest neighbour object on multidimensional data.

In high dimensional data, conventional algorithms often produce clusters that are not relevant. Conventional algorithms tend not to work to get the cluster with the maximum, even generate noise or outlier.

This research addressed four research questions:

a. How to avoid the curse of dimensionality in multidimensional data
b. How to identify cluster in different subspace in multidimensional data
c. How to develop an improved model for subspace clustering based on density connection
d. How to ensure subspace clustering based on density connection can be used to justify higher learning institution required skill.

1. 5 OBJECTIVE AND SCOPE

Objectives of the research are:

a. To develop an improved model for subspace clustering based on density connection
b. To develop clustering system for perception for skill required based on subspace cluster.
c. To develop online questionnaire for Higher Learning Institution (HLI) perception for skill required.

This research is limited in scope:

a. Using density based measurement for subspace clustering
b. Datasets input formed as numerical.

1. 6 CONTRIBUTION

This research, improved subspace clustering framework, under a well defined clustering goal, the high density cluster or its extension can be regarded as a parameter of interest for the underlying distribution. A clustering method, which can produce a cluster, can be regarded as a prediction. From a density connection, which is derived directly from clustering outcomes, a clustering distance measure is defined to assess the performance of different estimators. Some further techniques such as the subspace cluster are derived from the cluster family framework. These techniques can be used to increase accuracy and increase cluster significance, and reducing processing time. The work in this thesis provides an improved view of subspace clustering and has practical value in required skill planning.

This thesis contributes to the field of educational data mining, especially to the task of clustering, i.e. automatically grouping the objects of educational data into meaningful subclasses. The thesis includes: an improved model for multidimensional data mining, subspace clustering framework, and its application including a new subspace clustering model based on density connection. Also discuss the assessment of

clustering performance through this measure; the idea of forming a new clustering framework based on the concept of a density connection. Finally, additional techniques derived from the subspace clustering framework for generating new clustering methods.

1. 7 THESIS ORGANIZATION

The thesis is organized as follows:

Chapter 1 gives the reader a brief introduction the context of this thesis.

Chapter 2 provides a preview briefly of clustering, clustering high dimensional data, challenge in cluster detection, cluster analysis, subspace clustering, density based clustering, bottom up subspace clustering, top down subspace clustering, density based subspace clustering, cluster prediction, clustering paradigm, and lastly subspace measures.

Chapter 3 previews briefly of research methodology consist of research design, research framework, datasets, data collection, pre-processing data and validation, multidimensional data mining analysis.

Chapter 4 present the proposed technique to assess multidimensional data mining via for subspace clustering, which refer to multidimensional DAta MIning subspace clusteRing Approach (DAMIRA).

Chapter 5 present the implementation of DAMIRA. The experimental research are analysing, and comparison are done with the baseline technique, i.e., SUBCLU, FIRES, and INSCY based on three UCI benchmark datasets.

Chapter 6 present the implementation online questionnaire. The experimental research based on datasets from higher learning institution.

Chapter 7 summarizes and discusses the major contributions of the thesis. It concludes with pointing out some future research directions.

1. 8 SUMMARY

Expanding dimensions of used data in the era of massive automatic data and growth of dimensions will increase needs for data mining, improving processes data through using of tools, automation of data mining and implementation of the methodology. This methodology uses for classification, clustering, and outlier detection.