

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



**DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE)  
UNIVERSITI MALAYSIA PAHANG**

**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

## UNIVERSITI MALAYSIA PAHANG

## DECLARATION OF THESIS AND COPYRIGHT

Author's full name : Rahmat Widia Sembiring  
 Date of birth : 23 May 1965  
 Title : Density Based Subspace Clustering:  
 A Case Study on Perception of the Required Skill  
 Academic Session : 2013/2014

I declare that this thesis is classified as:

- CONFIDENTIAL (Contain confidential information under the Official Secret Act 1972)\*
- RESTRICTED (Contain restricted information as specified by the Organization where the research was done)\*
- OPEN ACCESS I agree that my thesis to be published as online open access (Full text)

I acknowledge that Universiti Malaysia Pahang reserve the right as follows:

1. The Thesis is the Property of Universiti Malaysia Pahang.
2. The Library of Universiti Malaysia Pahang has the right to make copies for the purpose of research only.
3. The Library has the right to make copies of the thesis for academic exchange.

Certified By:

-----  
 Rahmat Widia Sembiring

PCC09001/T632398


Date : 28 January 2014

-----  
 Prof. Dr. Jasni Mohamad Zain

Date : 28 January 2014

## SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis, and in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of "Doctor of Philosophy in Computer Science"

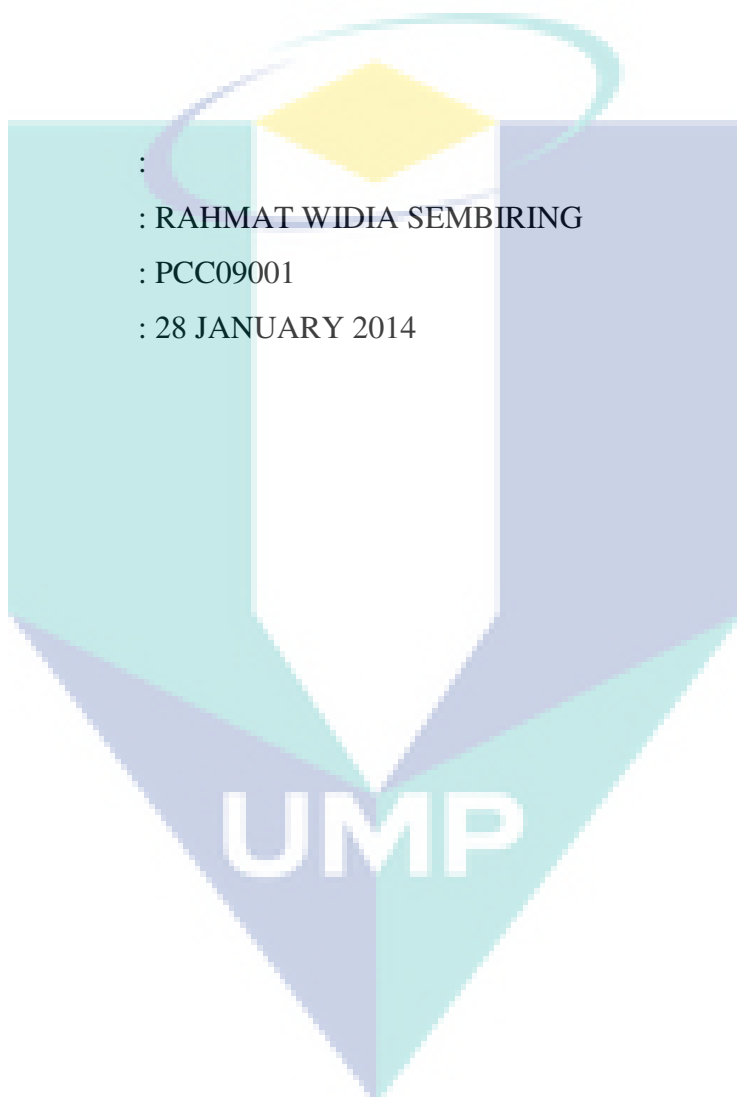


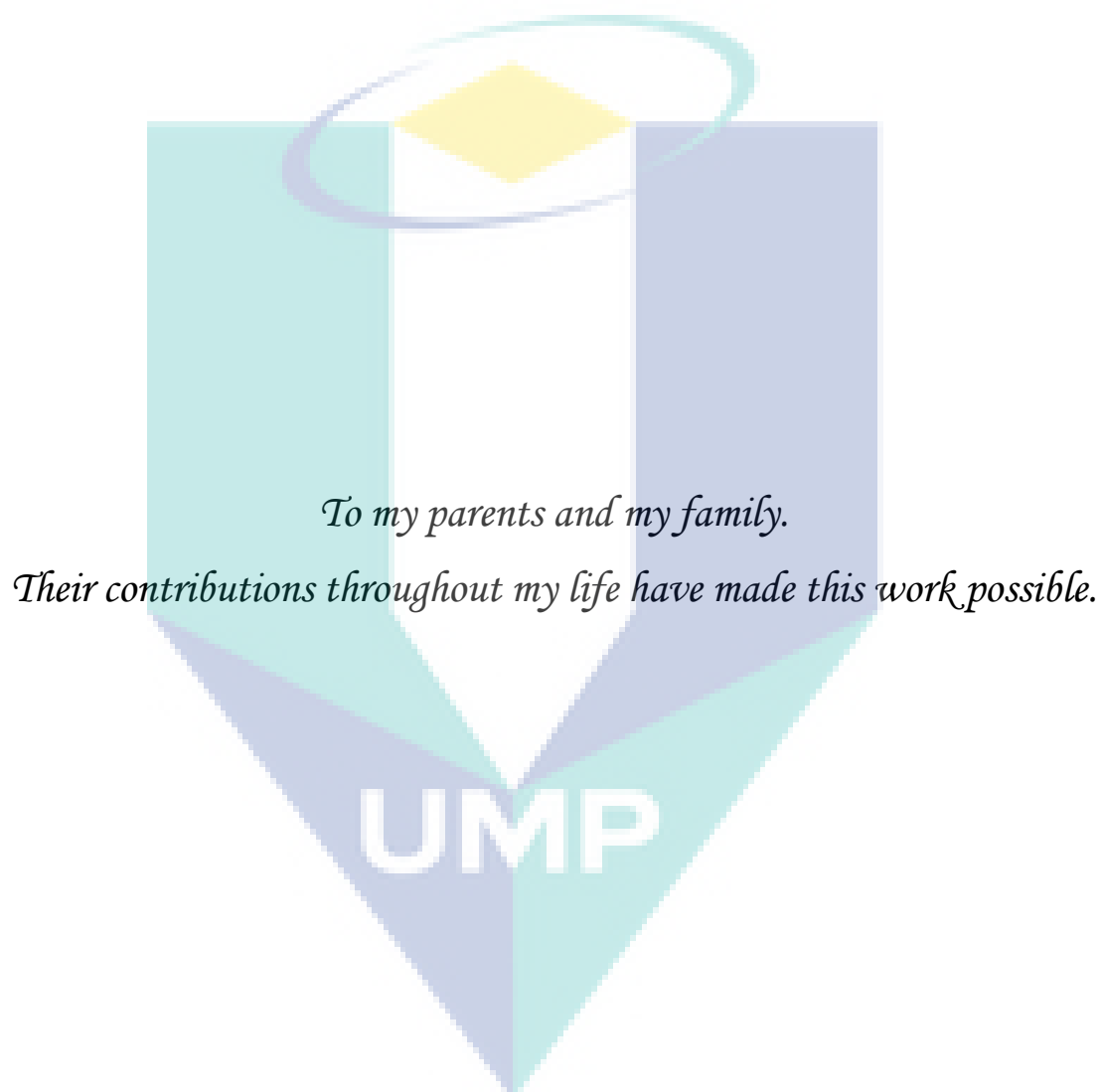
Signature :  
Name of Supervisor : PROFESSOR DR. JASNI MOHAMAD ZAIN  
Position : DEAN OF FACULTY OF COMPUTER SYSTEMS &  
SOFTWARE ENGINEERING  
Date : 28 JANUARY 2014

## STUDENT'S DECLARATION

I hereby declare that work in this thesis is my own except for quotations and summaries that have been duly acknowledged. The thesis has not been accepted for any degree and is not concurrently submitted for award of another degree.

Signature :  
Name : RAHMAT WIDIA SEMBIRING  
ID Number : PCC09001  
Date : 28 JANUARY 2014





## ACKNOWLEDGEMENTS

I humbly thank Allah Swt. for the unlimited help and mercy I have received during my life and study in Malaysia.

I would like to express my sincere gratitude to my supervisor Professor Dr. Jasni Mohamad Zain for her great ideas, invaluable guidance, continuous encouragement, and constant support in making this research possible. She has always impressed me with her outstanding professional conduct, and her tolerance of my naive mistakes. I would like to express my deep appreciation and respect to her as she always stand by me, in particular when I was in a very difficult situation, till the completion of this thesis, and these concluding moments. I also would like to express very special thanks to my previous supervisor Professor Dr. Abdullah Embong for his motivation to encourage Ph.D, his guidance, suggestions, and co-operation throughout the study.

A special appreciation I sincerely give to the external and internal examiner for valuable and constructive feedback to my thesis. All of the suggestions would enhance the quality of my thesis. I sincerely thank to the internal examiner for the interesting comment, to see the lost part in my thesis that I should proceed, with his/her suggestion I can finish my thesis with satisfactory result.

I would also like to thank Chancellor Universiti Malaysia Pahang, Vice Chancellor Universiti Malaysia Pahang, Dean and staff of Centre of Graduate Studies, Dean and staff of the Faculty of Computer Systems & Software Engineering, especially Professor Dr. Kamal Zuhairi Zamli, Associate Professor Ruzaini Abdullah Arshah, Dr. Azhar Kamaluddin, Dr. Mazlina Abdul Majid, Puan Darwina Kastam Tan, Puan Fauziah Sabli, friends, and the proofreaders, for their support throughout my study at this university.

My sincere thanks go to the Minister of Education and Culture Republic of Indonesia, Director General of Higher Education Ministry of Education and Culture Republic of Indonesia, Director of Politeknik Negeri Medan, Professor Dr. H.Muhammad Zarlis, H.Maulia Ahmad Ridwansyah Putra, Dedy Hartama M.Kom, Dr. Benny Benjamin Nasution, and Dr. Tutut Herawan who helped me in many ways.

I acknowledge my sincere indebtedness and gratitude to my parents for their love, dream, and sacrifice throughout my life, my parents-in-law, who consistently encouraged me to carry on my higher studies in Malaysia. I am also grateful to my wife (Hj. Dian Asmayuni Barus, S.Kom), my daughter (Amalia Nadhilah Sembiring) and my sons (Abidin Luthfi Sembiring and Alfian Ramadhan Sembiring) for their sacrifice, patience, and understanding that were inevitable to make this work possible, especially during the time when I had to be away from them. I cannot find the appropriate words that could properly describe my appreciation for their devotion, support, and faith in my ability to attain my goals.

## 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.



## ABSTRAK

Pada dekad ini banyak penelitian memperlihatkan peningkatan dimensi data, data yang berdekatan dapat menghilangkan informasi sebenar. Algoritma kluster biasa yang digunakan mengukur kesamaan data ataupun atribut berdimensi tinggi kerap tidak mencapai hasil yang diinginkan. Hal ini kerana atribut dari satu set data tidak berhubungan, atau sebaliknya terlalu berdekatan. Data yang terlalu berdekatan diperkirakan dapat membentuk kelompok yang saling bertumpang tindih dan akan membentuk kluster yang rapat, data mungkin terdapat pada kluster yang berbeda dan juga dalam sub-ruang yang berbeza. Kluster sub-ruang diproyeksikan sebagai teknik pencarian untuk mengelompokkan data atau atribut pada kluster yang berbeza. Pengelompokan dilakukan dengan menentukan tingkat kerapatan data dan juga mengidentifikasi outlier atau data yang tidak relevan, sehingga masing-masing cluster ada dalam subset tersendiri. Tesis ini mengusulkan inovasi algoritma kluster sub-ruang berdasarkan kerapatan dimensi. Pada tahap awal akan dihitung kerapatan dimensi, hasil kerapatan dimensi akan dijadikan data masukan untuk menentukan kluster awal, yakni menggunakan algoritma DBSCAN. Data pada setiap kluster kemudian diuji apakah memiliki hubungan dengan data pada kluster yang lain, yakni dengan menggunakan usulan algoritma kluster sub-ruang. Jika data memiliki hubungan dengan data di kluster yang lain maka akan dikelompokkan sebagai sebuah sub-ruang. Setiap data pada kluster sub-ruang kemudian akan diuji kembali dengan algoritma DBSCAN untuk melihat kembali kerapatannya, sampai akhirnya ditemukan kluster sub-ruang yang sebenar. Pada eksperimen akan digunakan *benchmark dataset* dan juga *real dataset*. *Real dataset* yang digunakan adalah dari bidang pendidikan, khususnya tentang persepsi *student industrial learning* dan dari industri. Data dikumpulkan melalui soal selidik. DAMIRA berhasil menghasilkan kluster dalam setiap sub-ruang. Untuk verifikasi kualiti dari metode kluster yang dikembangkan, kami menguji DBSCAN, FIRES, INSCY, SUBCLU. DAMIRA berjaya menghasilkan jumlah kluster yang besar pada setiap datasets. Sementara itu FIRES dan INSCY, tidak berjaya menghasilkan kluster pada semua datasets. DAMIRA dan SUBCLU tidak menghasilkan data yang tidak terkluster, hal ini menjadi asas bahwa kluster yang dihasilkan akan lebih akurat. Pada klusterisasi *glass dataset* dan *liver dataset*, metode DAMIRA memerlukan masa lebih lama 20 kali dari metode FIRES, INSCY and SUBCLU. Namun demikian pada *job satisfaction dataset* DAMIRA memerlukan waktu yang lebih sedikit. Untuk data yang lebih besar hasil DAMIRA lebih efisien dari SUBCLU, namun masih tetap lebih rendah dari FIRES, INSCY dan DBSCAN. DAMIRA berjaya memproses semua data menjadi kluster, sementara metod INSCY memproses lebih rendah dari FIRES. Untuk F1 Measure SUBCLU lebih baik dari FIRES, INSCY, dan DAMIRA. Penyelidikan ini menawarkan peningkatan model kluster subruang berasaskan kerapatan hubungan, menghadapi cabaran penambangan data pada kajian pendidikan. Model yang ditawarkan diberi nama DAMIRA, yang boleh digunakan untuk menduga persepsi kemahiran yang diperlukan untuk institusi pengajian tinggi.

## TABLE OF CONTENTS

	Page
<b>SUPERVISOR'S DECLARATION</b>	<b>ii</b>
<b>STUDENT'S DECLARATION</b>	<b>iii</b>
<b>ACKNOWLEDGEMENTS</b>	<b>iv</b>
<b>ABSTRACT</b>	<b>v</b>
<b>ABSTRAK</b>	<b>vi</b>
<b>TABLE OF CONTENTS</b>	<b>vii</b>
<b>LIST OF TABLES</b>	<b>x</b>
<b>LIST OF FIGURES</b>	<b>xi</b>
<b>LIST OF ABBREVIATION</b>	<b>xv</b>
<b>CHAPTER 1 INTRODUCTION</b>	<b>1</b>
1. 1 Background	1
1. 2 High Dimensional Data	3
1. 3 The Challenges	4
1. 3. 1 The Curse of Dimensionality	5
1. 3. 2 Multi Cluster Membership	6
1. 3. 3 Noise Tolerance	7
1. 3. 4 Applicability to Educational Application	8
1. 4 Problem Statement	9
1. 5 Objective and Scope	11
1. 6 Contribution	11
1. 7 Thesis Organization	12
1. 8 Summary	12
<b>CHAPTER 2 DATA MINING</b>	<b>14</b>
2. 1 Introduction	14
2. 2 Data Mining Technique and Its Application	21
2. 3 Clustering	23

2. 4	Clustering High Dimensional Data	25
2. 4. 1	Cluster Analysis	26
2. 4. 2	Dimension Reduction	28
2. 5	Density Based Clustering	30
2. 6	Subspace Based Clustering	33
2. 6. 1	Bottom up Subspace Clustering	39
2. 6. 2	Top down Subspace Clustering	40
2. 6. 3	Density Based Subspace Clustering Concept	40
2. 7	Educational Data Mining	41
2. 8	Performance Evaluation	44
2. 9	Summary	48
<b>CHAPTER 3 METHODOLOGY</b>		<b>49</b>
3. 1	Research Design	49
3. 1. 1	Pre-Research / Awareness of Problem	51
3. 1. 2	Suggestion	52
3. 1. 3	Development	52
3. 1. 4	Evaluation	53
3. 1. 5	Summary	53
3. 2	Research Framework	54
3. 2. 1	Data Collection	54
3. 2. 2	Pre-processing Data and Validation	56
3. 2. 3	Initial Data	58
3. 2. 4	Clustering Strategy	59
3. 2. 5	Strategy 1 of Data Analysis – Clustering Analysis	60
3. 2. 6	Strategy 2 of Data Analysis – Subspace Cluster Analysis	62
3. 2. 7	Strategy 3 of Data Analysis – Subspace Cluster Based on Density Connection	62
3. 3	Summary	62
<b>CHAPTER 4 SUBSPACE CLUSTER BASED ON DENSITY CONNECTION</b>		<b>64</b>
4. 1	Clustering Strategy	64
4. 2	Subspace Clustering	69
4. 3	Subspace Clustering Based On Density Connection	73
4. 4	Summary	82

<b>CHAPTER 5 RESEARCH FINDING AND DISCUSSION</b>	<b>83</b>
5.1 Introduction	83
5.2 Dataset Properties	87
5.3 Experimental Result	93
5.4 Performance Evaluation	111
5.4.1 Efficiency	111
5.4.2 Accurate	113
5.4.3 Coverage	114
5.4.4 F1-Measure	115
5.5 Summary	116
<b>CHAPTER 6 ONLINE QUESTIONNAIRE FOR EDUCATIONAL DATA MINING</b>	<b>117</b>
6.1 Platform of Online Questionnaire	117
6.2 Summary	131
<b>CHAPTER 7 CONCLUSION AND FUTURE WORK</b>	<b>132</b>
7.1 Conclusion	132
7.2 Future Works	134
<b>REFERENCES</b>	<b>136</b>
<b>APPENDIXES</b>	<b>152</b>
<b>LIST OF PUBLICATION</b>	<b>157</b>

## LIST OF TABLES

Table No.	Title	Page
2.1	Summarizes using of clustering for Educational Data Mining (EDM)	44
3.1	Benchmark Real World Data Set	54
3.2	Number of sample in each respondent group	55
3.3	Student Industrial Training Dataset	57
3.4	Industrial Dataset	58
5.1	Example of Initial Data	84
5.2	Multidimensional separate into 1-dimension	85
5.3	Clustering result based on DBSCAN	85
5.4	Result of Generate Subspace Cluster	86
5.5	Result of Group of Subspace Cluster	86
5.6	The Property of Dataset	88

The image features a large, semi-transparent watermark of the UMP logo. The logo is a shield-like shape composed of four colored triangles (teal, light blue, yellow, and purple) meeting at a central point. The letters 'UMP' are printed in white, bold, sans-serif font across the bottom of the shield.

UMP

## LIST OF FIGURES

Figure No.	Title	Page
1.1.	The high dimensional data	2
1.2.	The curse of dimensionality	5
1.3.	Another curse of dimensionality	6
1.4.	Cross section on X-Y axis	7
1.5.	Cross section on X-Z axis	7
1.6.	Noise in clustering	7
2.1.	Structure of informatics	15
2.2.	Knowledge Discovery Process	16
2.3.	Supervised Learning	18
2.4.	Flow to solve a given problem of supervised learning	19
2.5.	Unsupervised Learning	20
2.6.	Data Mining Taxonomy	22
2.7.	Data mining technique	23
2.8.	Rare and common cases in unlabelled data	24
2.9.	Matrix	27
2.10.	Dissimilarity matrix	27
2.11.	Core distance of OPTICS	32
2.12.	Pseudo code of basic OPTICS	32
2.13.	Procedure ExpandClusterOrder of OPTICS	33
2.14.	Illustration of the two general problems of clustering high-dimensional data	34
2.15.	Data with 11 object in one bin	35
2.16.	Data with 6 objects in one bin	35
2.17.	Data with 4 objects in one bin.	35
2.18.	Cluster overlap each other	36
2.19.	Sample data plot in 2 dimension (a and b).	37
2.20.	Sample data plot in 2 dimension (b and c).	37
2.21.	Sample data visible in 4 cluster.	37
2.22.	Subspace Clustering	41
2.23.	Purity as an external evaluation criterion for cluster quality.	46

2.24.	Precision (P) and Recall (R)	47
2.25.	Illustration of true false of expectation	47
3.1.	Research Design	50
3.2.	Path of literature review	51
3.3.	Research Framework	54
3.4.	The strategy of multidimensional data mining analysis.	60
3.5.	Cluster initialization	61
4.1.	The cluster of points and also identify outliers	65
4.2.	Border and core point	66
4.3.	Density reachable	66
4.4.	Another density reachable	67
4.5.	Density connected	67
4.6.	Define cluster	68
4.7.	Normalization	70
4.8.	Define point and border point	70
4.9.	Density reachable	71
4.10.	Define connection of each other point	72
4.11.	A procedure of data sets usages.	73
4.12.	Pseudocode of DAta MIning subspace clusteRing Approach (DAMIRA)	75
4.13.	Change n-dimension to 1-dimension	75
4.14.	Flowchart to find first cluster and first subspace	76
4.15.	Initial data (database)	77
4.16.	1-dimension of cluster	77
4.17.	Flowchart to determine candidate subspace	78
4.18.	1-dimension of cluster	78
4.19.	1-dimension of cluster	78
4.20.	Detail of candidate subspace	79
4.21.	Flowchart to determine best subspace	80
4.22.	Determine best subspace	81
4.23.	Script to determine best subspace	81
5.1.	Separate multidimensional into 1-dimension	85
5.2.	Separate multidimensional into 1-dimension	87
5.3.	Data distribution of glass datasets	88

5.4.	Data distribution of liver datasets	89
5.5.	Data distribution of job satisfaction datasets	89
5.6.	Data distribution of Ump_student_ b1_b4 datasets	90
5.7.	Data distribution of Ump_student_ c1_c11 datasets	90
5.8.	Data distribution of Ump_student_d1_d6 datasets	91
5.9.	Data distribution of Ump_industry_ b1_b4 datasets	91
5.10.	Data distribution of Ump_industry_ c1_c11 datasets	92
5.11.	Data distribution of Ump_industry_d1_d6 datasets	92
5.12.	Number of cluster real datasets	93
5.13.	Algorithm Capability	94
5.14.	Application Programs	95
5.15.	Computer Programming	96
5.16.	Hardware and Device	97
5.17.	Human Computer Interaction	98
5.18.	Information System	99
5.19.	Information Management (Database)	99
5.20.	IT Resource Planning	100
5.21.	Intelligent System	101
5.22.	Networking and Communication	102
5.23.	System Development through Integration	103
5.24.	Resource Management	104
5.25.	Communication and Interpersonal	105
5.26.	Leadership	106
5.27.	Information Management	107
5.28.	Systems Thinking	108
5.29.	Technical/Functional Competence	109
5.30.	Number of cluster Higher Learning Institution datasets	110
5.31.	Un-cluster data of real datasets	110
5.32.	Un-cluster data of higher learning institution datasets	111
5.33.	Time processing of clustering of real datasets	112
5.34.	Time processing of clustering of higher learning institution datasets	113
5.35.	Accuracy of real datasets	114
5.36.	Coverage of real datasets	114



5.37.	Coverage of higher learning institution datasets	115
5.38.	F1 measure of real datasets	116
6.1.	Web architecture	119
6.2.	Web homepage	120
6.3.	Homepage of data access	121
6.4.	Architecture of online questionnaire	122
6.5.	Key in for add new HLI	123
6.6.	Dashboard of HLI detail	123
6.7.	Editing online questionnaire	124
6.8.	Manage questionnaire	124
6.9.	Industrial Respondent Database	125
6.10.	Result of student respondent	125
6.11.	Export Student Respondent Result	126
6.12.	University and study program choose	126
6.13.	Flowchart query of questionnaire	127
6.14.	Student details form	128
6.15.	Online Questionnaire section	129
6.16.	Question structure for frequency of course implemented	129
6.17.	Question structure for important knowledge competence	130
6.18.	Question structure for importance of soft skill competence	130
7.1.	Flowchart of online questionnaire	156

The image contains a large, semi-transparent watermark of the UMP logo. The logo is a shield-like shape divided into four quadrants of different colors: top-left is light blue, top-right is light purple, bottom-left is light green, and bottom-right is light blue. The letters 'UMP' are written in white, bold, sans-serif font across the center of the shield. The watermark is positioned behind the table of contents.

## 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  $y_{1,1}; y_{1,2}; \dots; y_{1,4}$ , same state notion in  $y_2$  and  $y_3$ .

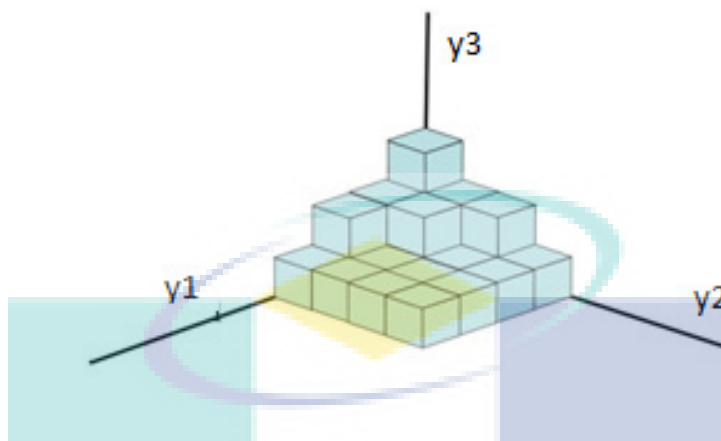


Figure 1.1. The high dimensional data

If there is data with a number of dimension  $D$ , and each dimension divided into  $I$  intervals, then there are  $I^D$  element of data become  $D+1$ , then element of data become  $I^{D+1} = I^{D+1}$ . 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 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).

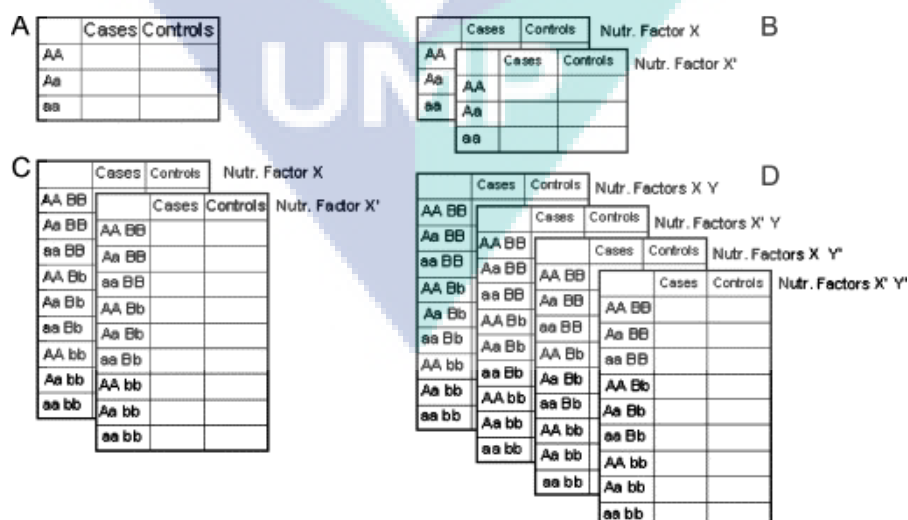


Figure 1.2. The curse of dimensionality



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.

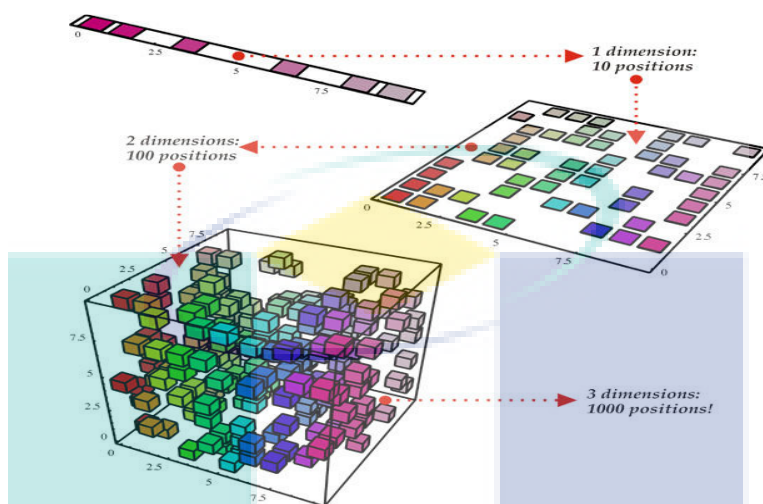


Figure 1.3. Another curse of dimensionality

### 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.

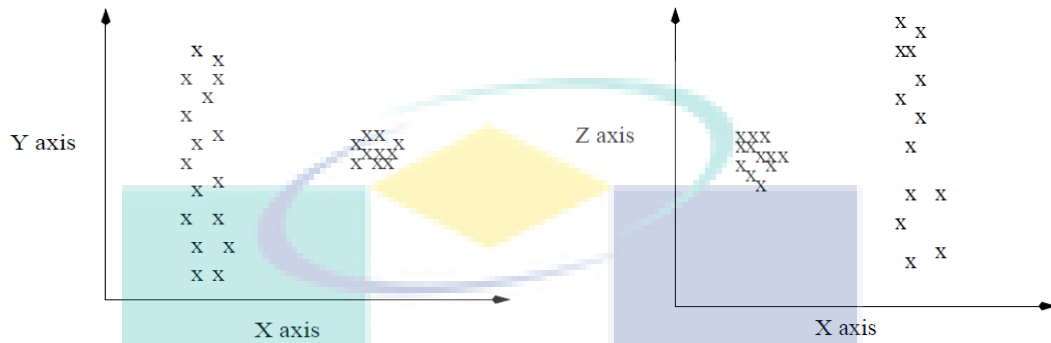


Figure 1.4. Cross section on X-Y axis

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.

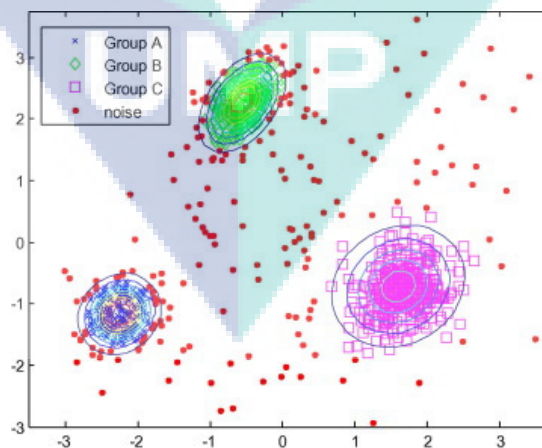


Figure 1.6. Noise in clustering

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-learning 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

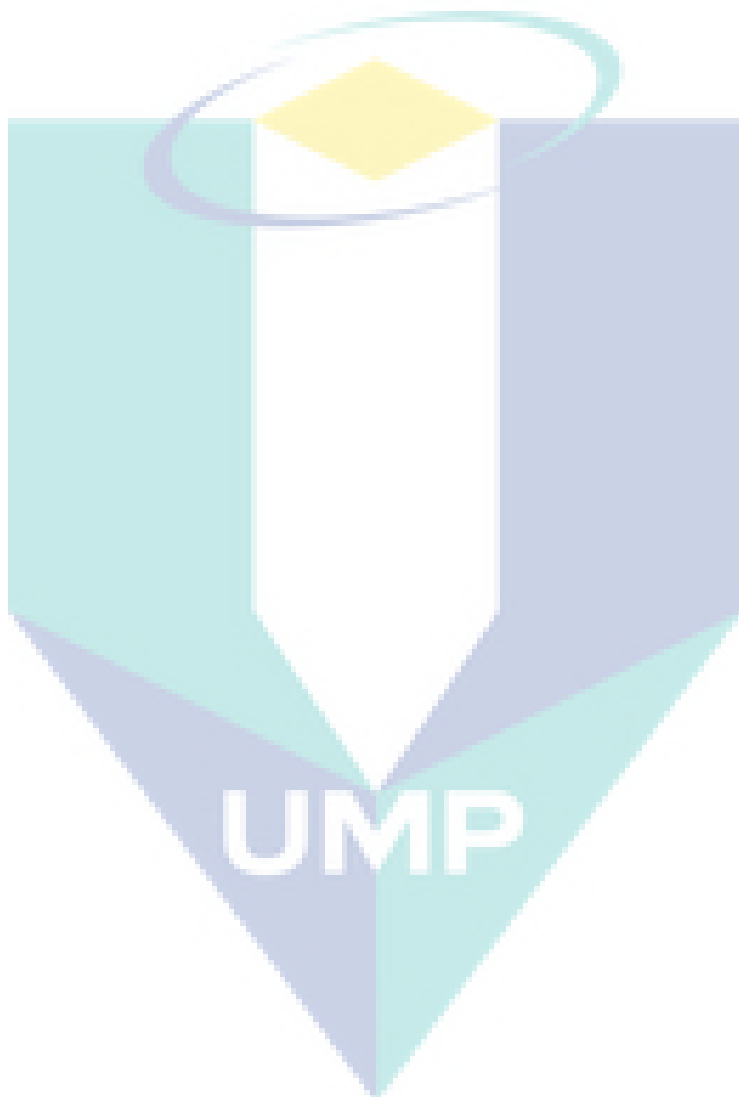
The thesis is organized as follows:

- 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.

Some challenges in high dimensional data mining are the curse of dimensionality, multi cluster membership and noise tolerance. The concept of a cluster is seriously challenged in high dimensional spaces, where subspace is considered when subset of attribute or data points belongs to different clusters in different subspaces. This study try to develop an improved model for subspace clustering based on density connection, implement in multidimensional data, and educational data.





## CHAPTER 2

### DATA MINING

This chapter introduces data mining, machine learning, data mining technique and its application. This thesis focuses on clustering multidimensional data, overcome challenge in cluster detection, and cluster analysis. For cluster subspace, cluster was chosen as an important topic to discuss, especially for density based clustering. Two kinds of subspace clustering methods were discussed; bottom up subspace clustering and top down subspace clustering. Overall, as a main reference for thesis novelty, this chapter discusses density based subspace clustering concept with reference to subspace measures.

#### 2.1 INTRODUCTION

Rapid development of information technology, increasing capacity of storage media and big computer network connections will lead to the increase use of digital data processing. The larger amount of data and the diversity of the course cause it to become more difficult to process it into useful information. The structure information in general is influenced by several factors including storing, processing and transmission, as shown in Figure 2.1 (Berka et al., 2009). Good data processing is needed to generate useful information. Data processing is influenced by the theory of computation, programming, databases, and knowledge bases. In fact, it is identified as rich data but

with poor information, the data only produces very little information. In this condition, knowledge and discovery become necessary to be implemented.

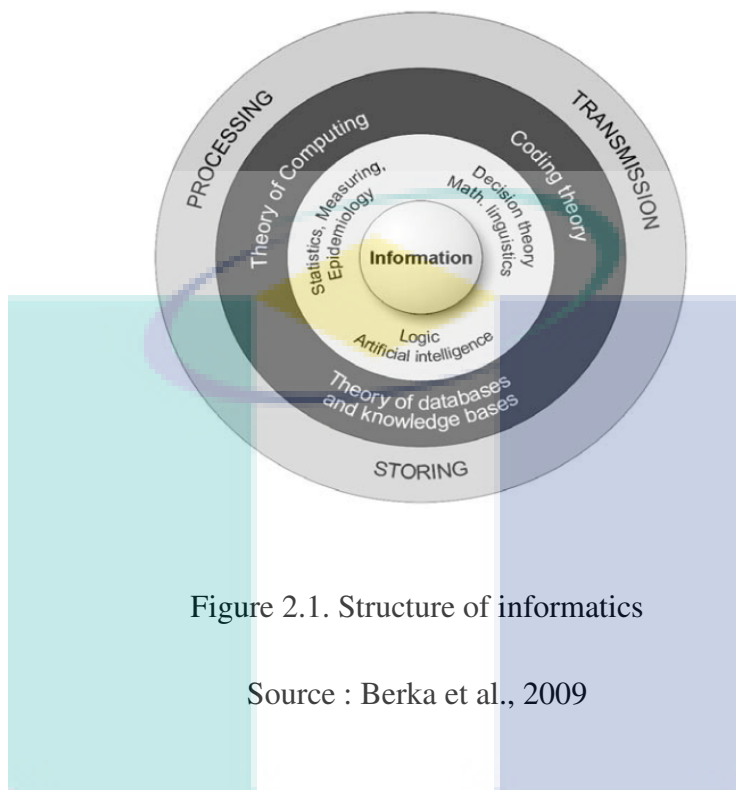


Figure 2.1. Structure of informatics

Source : Berka et al., 2009

Knowledge Discovery from Data (KDD) is developed out of the data mining domain, and is closely related in terms of methodology and terminology (Fayyad et al., 1996). KDD is defined as the extraction of a set of information previously unknown, but potentially useful. The activities start from data collection and integrated into the storage media. The collected data will be selected and pre-processing will be carried out. The results will be presented in the prepared data. This data will be processed through data mining. The process of data mining will result in a pattern and with interpretation and assimilation, these patterns will generate knowledge. Figure 2.2 shows the stages of the process of KDD by Brachman (Abonyi and Balazs, 2007).

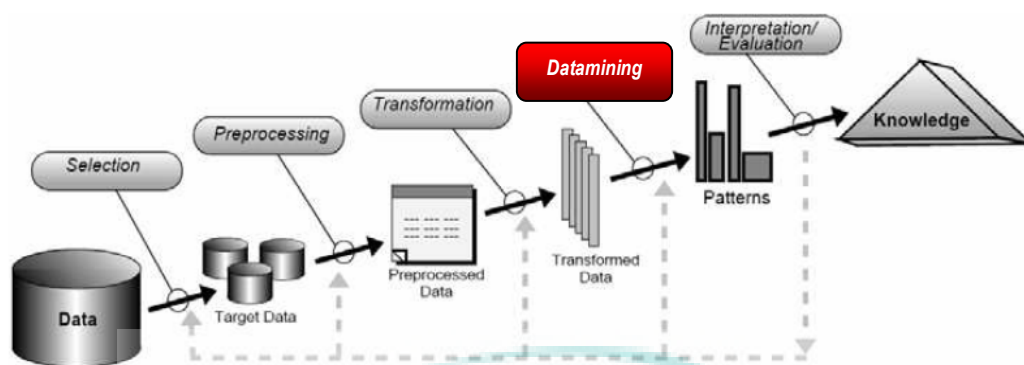


Figure 2.2. Knowledge Discovery Process

Source : Abonyi and Balazs, 2007

From Figure 2.2, we can see data mining as part of knowledge discovery. Data mining techniques have been applied to industrial problem solving, engineering science, both in private, and government industries. Data mining techniques have increased over the last decade, followed by the progress of data mining tools, which can be used for various applications. Several applications often require data mining, such as health, medical, marketing, sales, medical, financial, e-commerce, multimedia, security, and lately developed for educational purposes too. The consequence of this fact resulted in the rapid growth of data mining needs. Development of data mining is also a important issue that is always studied in HLI.

In data mining we know machine learning can perform the learning instead of human beings. With the advantage of technology, currently we can store large amounts of data that can even be accessed from different places. The data can be recorded in digital form or other data. Such data will only be useful if it is processed into information and used to exercise prediction. Suppose the database has stored data about the date of consumer spending, the type of goods purchased, cost per item of goods, and also certain specification of the goods, we can predict the next purchase of the consumer. This is the niche of research in machine learning. Application of machine learned methods of large databases is called data mining (Alpaydin, 2010). In relation with computer science, machine learning lies in the process of getting results. Computer

science focuses on the results from a computer programmed manually, while machine learning is used so that computers can obtain its own. Machine learning is, by and large, a direct descendant of an older discipline, statistical model fitting (Baldi and Soren, 2001), and usually, programmed computers should be used in order to optimize a performance criterion using example data or experience.

Currently, computer science has developed into a larger area, where the various sub-disciplines that have been developed are part of the data analysis. Sub-disciplines are largely overlapping with each other as well as with the statistics. Sub-disciplines are machine learning, pattern recognition, data mining, image processing, neural networks, and perhaps computational learning theory and other areas as well (Boulicaut et al., 2004). Machine learning is an intersection between computer science and statistic.

Machine learning will have some parameters, to predict or to make predictions in the future, or descriptive to gain knowledge from data. The goal in machine learning is to extract useful information from a corpus of data by building good probabilistic models (Baldi and Soren, 2001). Implementation of machine learning is classified as supervised (directed) learning and unsupervised (in-directed) learning. Generally unsupervised learning created the probability for inputs undefined, while in unsupervised learning create observations as assumed of latent variables. With unsupervised learning, it is possible to learn larger and more complex models than with supervised learning.

Supervised learning assumes that training examples are classified whereas unsupervised learning concerns the analysis of unclassified examples. This approach is referred to as supervised (hypothesis development and testing). However, there is more to data mining than the technical tools used. Data mining involves a spirit of knowledge discovery, learning new and useful things, referred to as unsupervised.

Supervised learning is a part of machine learning task to infer data or information from supervised training data and consists of a set of training examples. Each example consists of an input object and a desired output value. The supervised learning algorithms (Figure 2.3) assume that the training data has a fixed set of

predicting attributes. This algorithm analyses the training data and produces an inferred function as classifier or a regression.

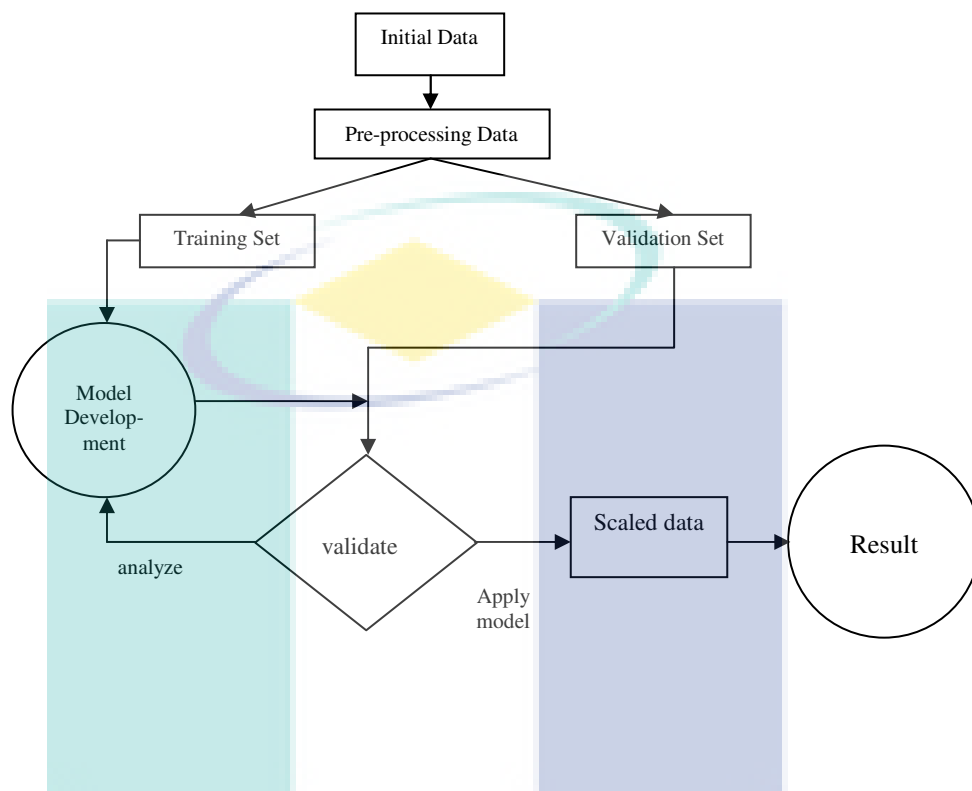


Figure 2.3. Supervised Learning

Supervised learning is common in classification; the final is often forgetting the computer to learn a classification of datasets. The inferred function must be predicting the correct output value for any valid input object. The most common technique of supervised learning is to neural networks, decision trees, regression, decision trees, bayesian networks, rule induction and support vector machines.

Many supervised learning algorithms were developed to create classification. Examples include back propagation algorithm, Naive Bayes Classifier, C4.5, Info Fuzzy Network (Last, 2004), Analytical Learning (Zhang and Jeffrey, 2005), Artificial Neural Network by Yu (Boulicaut et al., 2004), Support Vector Machine by Farquad (Olivas et al. 2010), but Artificial Neural Networks can be used for both supervised and unsupervised learning. The flow to solve a given problem of supervised learning is shown in Figure 2.4.

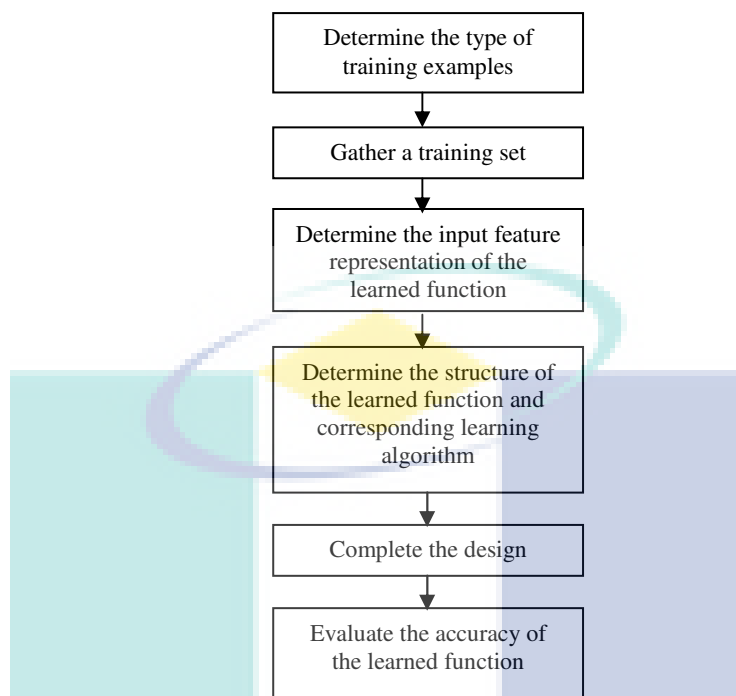


Figure 2.4. Flow to solve a given problem of supervised learning

While unsupervised learning is closely related to the density estimation in statistics. Unsupervised is also used to summarize and explain the key features of the data (Figure 2.5). Performed unsupervised learning discovers the hidden structure of data that is without label as predefined information. Many methods used in unsupervised learning are based on data mining methods for pre-processing data. The advantage of unsupervised method is that there is no need to have any information about the data beforehand.

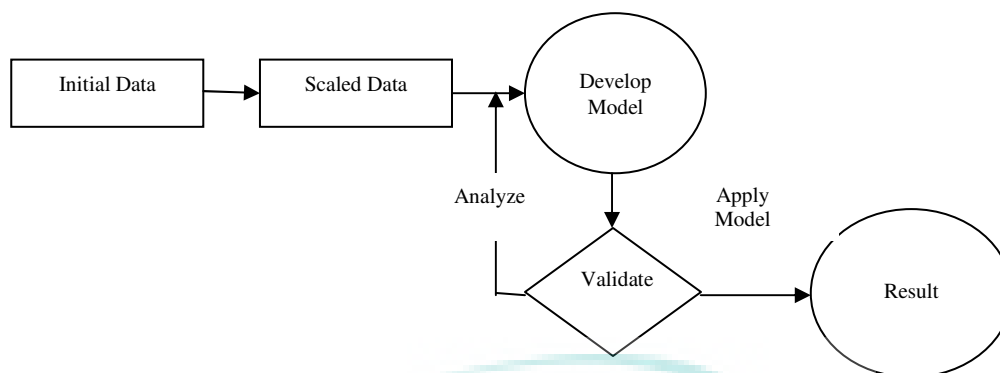


Figure 2.5. Unsupervised Learning

Unsupervised methods, such as Principal Component Analysis (PCA) and Latent Semantic Indexing (LSI) are obviously directly applicable to multi-label data. For example, in (Gao, 2005) the authors directly apply LSI based on singular value decomposition in order to reduce the dimensionality of the text categorization problem.

Unsupervised learning is considered important because it tends to present the data tends to consist of various types, and there is no pattern. (Dayan, 2008). In unsupervised learning the machine simply receives inputs  $x_1, x_2, \dots, x_n$ , but hard to imagine that the machine can give the correct response, because it is not obtained preliminary data or feedback as preliminary information. (Ghahraman, 2004). Any kind of unsupervised methods are clustering, association rules, link analysis, and visualization.

A problem that we face in clustering is to decide the optimal number of clusters into which data can be partitioned. In most algorithms, through two-dimensional experimental results to visualize cluster can easily, and verify it, but it becomes uneasy when the number of dimensions increases. It is clear to justify that result visualization cluster to be very important, especially to verify the correctness of the results (Maimon and Lior, 2005). Unsupervised learning is more exploratory in nature and tries to find patterns of interest that are intrinsic to the data and not related to some imposed label. For example, clustering and segmentation is a form of unsupervised. Some of

unsupervised techniques are association rules, K-means clustering, and self-organizing maps.

Moreover, the perception of clusters using available visualization tools is a difficult task for humans that are not accustomed to higher dimensional spaces. As a consequence, if the determination of the parameters is not appropriate clustering algorithm, caused cluster result is not optimal, even leading to a wrong decision. Determine the exact number of clusters results are discussed and become a major issue in research (Dave, 1996.; Gath and Geva, 1989.; Theodoridis and Konstantinos, 2009.; Xie et al., 2009.; Maimon and Lior, 2005).

Unsupervised learning presented with unclassified instances aims to identifying groups of instances with similar attribute values. In addition, the numerical discretization is carried out, normalization or standardization of numerical attributes, making the indicator, merging attribute values. Nominal change the binary value, exchange value, remove the attribute, replacing missing values, is another thing that can be done on unsupervised learning.

Semi-supervised learning is crossing the borders between traditional unsupervised clustering without external knowledge and classification, which is the classical task within supervised learning.

## 2.2 DATA MINING TECHNIQUE AND ITS APPLICATION

Data mining technique consists of six common classes of activities: anomaly detection, association rule learning, clustering, classification, regression, and summarization. Data mining is an interdisciplinary intersection of artificial intelligence, machine learning, statistics, and database systems. Recent years the trends of data mining include distributed data mining, hypertext/hypermedia mining, ubiquitous data mining, as well as multimedia, spatial, time series, and sequential data mining (Hsu, 2002).



The main endeavour in data mining is to extract knowledge from data. This knowledge is captured in order for it to be understood by a human. Basically, this application depends on the specific problem at hand. It is well known that Data Mining methods can be classified into exploratory, descriptive (or unsupervised), predictive (or supervised) and local (Hand et al., 2010; Maimon and Lior, 2005), while data mining taxonomy shown in Figure 2.6.

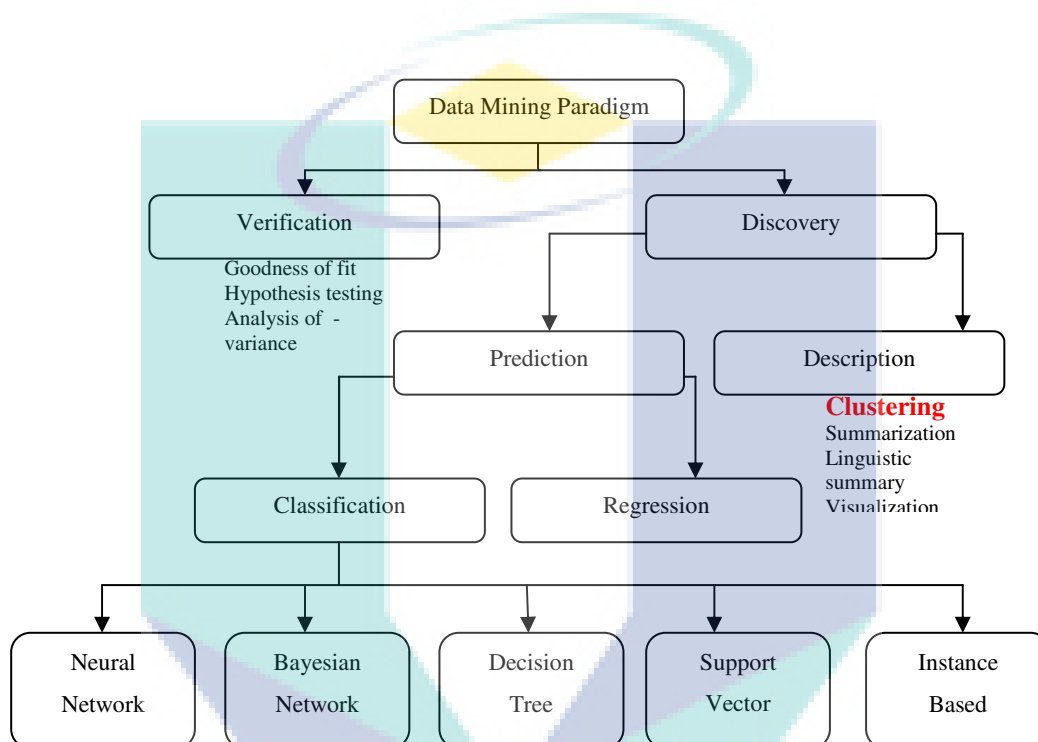


Figure 2.6. Data Mining Taxonomy

Generally, data mining techniques (Figure 2.7) are based on inductive logic, statistical reasoning, programming, fuzzy sets, machine learning and neural network techniques. Based on the hypotheses, information from dataset will be extracted and observed. The patterns that emerge will be observed to answer or study the discovered rule to partition the data into a particular group and make associations between data, or find customized data rules.

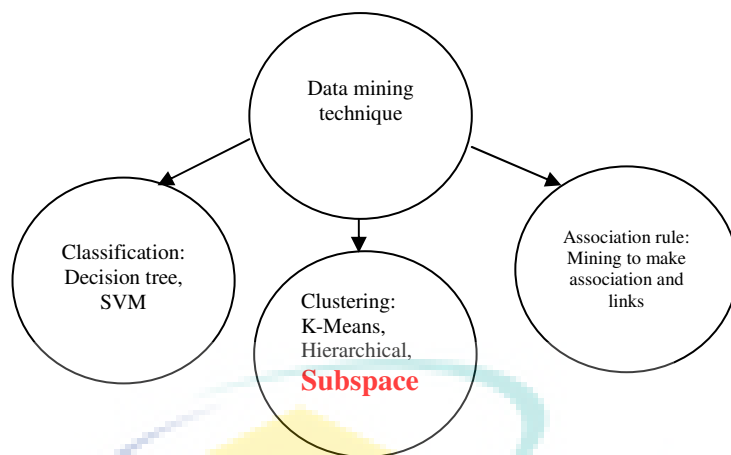


Figure 2.7. Data mining technique

In the future, data mining will become more powerful. The expansive use of data through internet, wireless, and gadget will exploit huge number of data. Pre-processing will become a necessary part of data mining, faster and transparent (Kriegel et al., 2007). Most important areas of data mining need to focus on distributed data mining (Hsu, 2002), while statistical data mining and efficient algorithm will time consume and increase result quality.

### 2.3 CLUSTERING

Classification and clustering are two major machine learning tools and two of the most important areas of data mining because of the vast challenge it contributes to the field of research such as scalability, ability to deal with different types of attributes, discovery of cluster with arbitrary shape, noisy data, interpretability and usability (Han and Micheline, 2006). This method can be applied in medical diagnosis, fraud detection, education, and predicting financial trends.

One of the most effective methods for analysing very large amounts of data, collected in a dataset, is to use a clustering method. The process of clustering is considered as an unsupervised learning portion, because there is no predefined class determination, and without initial information for a data. (Maimon and Lior, 2005). The main purpose of clustering method is to identify groups of unlabelled data (Figure 2.8),

by defining objects based on similarity, and organize them into several different clusters. Distance measurement can be adopted as a measure of the difference.

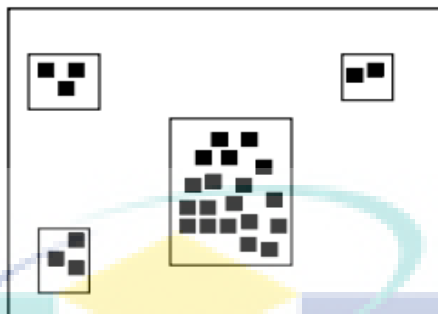


Figure 2.8. Rare and common cases in unlabelled data

Source : Maimon and Lior, 2005

The main purpose of clustering is to find a natural grouping of the data without requiring any background knowledge. Without background knowledge, it is often very difficult to specify appropriate parameter settings.

Some traditional clustering algorithms have been proven to determine the cluster of low-dimensional data, but often fail when high dimensional data. For high dimensional data, there will be tens or even hundreds of attributes. One important data mining task associated with unsupervised learning is clustering, which involves the grouping of entities into categories. A second important unsupervised learning task is association rule mining, which looks for associations between items (Aggarwal et al., 1999; Maimon and Lior, 2005).

Clustering (Kantardzic, 2011) is the process of grouping similar objects/records. It is the main strategy for unsupervised learning. Clustering is used in data set description and as a preliminary step in some predictive learning tasks for better understanding of the data to be analysed (Maimon and Lior, 2005). Clustering is a form of unsupervised classification, which means that the categories into which the collection

must be partitioned are not known, and so the clustering process involves the discovering of these categories.

Clustering is similar to classification which categorizes data into groups. Clustering is usually done by determining the similarity between the data in the predefined attributes. Most similar data are grouped into one group, and the group usually specifically different from the others. Because the cluster is not a standard, required in-depth interpretation. Generally, there are two methods of clustering, ie clustering partitioned time series clustering and hierarchical clustering (Maimon and Lior, 2005).

#### 2.4 CLUSTERING HIGH DIMENSIONAL DATA

Clustering is a process of grouping data items based on the values of their attributes. Data items generally interpreted as points in multidimensional features space. Item within the same cluster should be similar according. Clustering high dimensional data is a current problem in many area of science; complexities of data increase its dimension.

Clustering high dimensional data is analyses of the data have dozens or even thousands of dimensions. One method that many clustering observed in the high dimensional data is subspace clustering. Subspace clustering is widely used in many fields, such as medicine, where DNA microarray technology can produce a large number of data dimensions. It is also widely used in clustering text documents, climatology, and education.

If the dimension of data is too large it will be difficult to be calculated, it will also be difficult visualized, as well as case dimensions exponentially increase. This problem is known as the curse of dimension. In this case the concept of distance becomes a major challenge, because it will be so dense data. Discrimination of the nearest and farthest point the data is blurred or even meaningless.

A cluster is intended to group objects that are related, based on observations of their attribute's values. However, given a large number of attributes some of the attributes will usually not be meaningful for a given cluster. For example, in newborn screening a cluster of samples might identify newborns that share similar blood values, which might lead to insights about the relevance of certain blood values for a disease. But for different diseases, different blood values might form a cluster, and other values might be uncorrelated. This is known as the local feature relevance problem.

Different clusters might be found in different subspaces, so a global filtering of attributes is not sufficient. Given a large number of attributes, it is likely that some attributes are correlated. Hence, clusters might exist in arbitrarily oriented affine subspaces. Recent research (Houle et al. 2010), indicates that the discrimination problems only occur when there are a high number of irrelevant dimensions, and that shared-nearest-neighbour approaches can improve results.

#### 2.4.1 Cluster Analysis

Analysis cluster takes ungrouped data and uses automatic techniques to put this data into groups. Clustering is unsupervised, and does not require a learning set. It shared a common methodological ground with classification. Most of the mathematical models can be applied to cluster analysis as well, dividing object into groups, as clustering and assigning its particular to describe an object as classification.

Typically cluster analysis operates on data matrix (Figure 2.9) either in dissimilarity matrix (Figure 2.10) (Hand, 2010). Data matrix presents  $m$  object, such as number of respondent, with  $n$  variable such as, course, knowledge skill, and soft skill. The structure is in the form or relational table, or  $m$ -by- $n$  matrix ( $m$  objects  $\times$   $p$  variables). Dissimilarity matrix stores a number of proximities available for all pairs of  $n$  objects in table  $n$ -by- $n$  where  $d(i,j)$  is the measured difference or dissimilarity between objects  $i$  and  $j$ .

$$\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$$

Figure 2.9. Matrix

$$\begin{bmatrix} 0 & & & & & \\ d(2,1) & 0 & & & & \\ d(3,1) & d(3,2) & 0 & & & \\ \vdots & \vdots & \vdots & & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 & \end{bmatrix}$$

Figure 2.10. Dissimilarity matrix

Source : Hand, 2010

The reason for cluster analysis is to utilize result and characterize cluster as unique prototype representative of objects. This prototype can be used as a basis for the number of data for data processing technique. Clearly defined clustering is a subjective process and depends on precludes an absolute judgment relative to efficacy of all clustering techniques (Baraldi and Alpaydin, 2002.; Jain et al., 1999).

Basic procedures of cluster analysis are (Olivas, 2010):

- a. Feature selection and feature extraction.  
Feature selection chooses distinguishing feature from a set of candidates; while feature extraction is utilize some transformation to generate useful and novel features from the original ones.
- b. Clustering algorithm design or selection  
Usually, related to determine of an appropriate similarity or instance measure, and construction of criterion function. Data objects are clustered based on similarity of each other.
- c. Cluster validation  
Partition member or not in a group usually lead to a different cluster of data. Effective evaluation standard and criteria are critically important to provide user with confidence for the result.
- d. Result interpretation  
Important goal of clustering is provide with meaningful and clear understanding of the data, and solve the problem effectively.

## 2. 4. 2 Dimension Reduction

Applications related to multidimensional data continue to grow. Techniques to support more efficient query becomes an important research issue at this time. This technique is needed to open the multimedia content, data exploration in areas of health, population issues, decision-making in education, as well as to analyse the time-series. Processes such as data pre-processing, data cleaning, data integration and transformation, and reduction of dimension can be applied to improve the quality of the results.

Real data are often incomplete (Magnani, et al., 2004), lack of attributes, noisy, contains outlier, and also inconsistent, thus requiring the data pre-processing. Pre-processing of data is to improve algorithm (Orfanidis, et al., 2008), accuracy, completeness, consistency, timeliness, value added, interpretation, and better accessibility.

Pre-processing is the process of transforming data into simpler, more effective, and in accordance with user needs. More accurate results and shorter computation time can be used as indicators. The data also becomes smaller without changing the information in it. Some pre-processing method is done by selecting a subset of a large population sample of data, referred to as denoising. This will be followed by normalization and feature extraction.

Dimensional reduction becomes a fundamental problem in most of the data mining process. It benefits not only for computational efficiency, but can be improving the accuracy of the analysis (Cunningham, 2007). Dimension reduction techniques are often used to overcome “the curse of dimensionality”, as part of pre-processing in addition to simplify the data model.

Dimension reduction techniques can be grouped into feature selection and feature extraction. Feature selection is the process of finding a subset of the original variables, with the aim to reduce and eliminate the noise dimension. It can improve the performance of data mining, including improving the speed and accuracy. In some

cases, regression or classification analysis can be done to reduce the dimension, which produces more accurate dimensions. Several algorithms have been proposed such as ReliefF (Sikonja and Kononenko, 2003), Focus, Support Vector Machine Recursive Feature Elimination (SVM RFE) and Feature Subset Selection using Expectation Maximization (FSSEM).

Feature extraction is a technique to transform high-dimensional data into lower dimensions. Several supervised learning algorithms have been proposed, namely Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA), Partial Least Square (PLS), Latent Semantic Indexing (LSI), and Singular Value Decomposition (SVD). While for unsupervised learning, algorithms such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), FastICA (extension of ICA) can be used as a basic component analysis.

Dimension reduction methods associated with regression, additive models, neural network models, and methods of Hessian (Fodor, et al. 2003). Local Dimension Reduction (LDR) looks for relationships in the dataset and reduces the dimensions of each individual using a multidimensional index structure (Chakrabarti, et al., 2000). Nonlinear algorithm gives better performance than PCA for sound and image data (Kambhatla, et al. 1994). Principal Component Analysis (PCA) which is based on dimension reduction and texture classification scheme can be applied to manifold statistical framework (Sang, et al., 2007). The semantics of linear algebra is significantly simpler than Probabilistic Latent Semantic Analysis (PLSA) and LDA, while PLSA is much simpler than the LDA (Chua and Chong, 2009).

Most applications, dimension reduction is performed as pre-processing step (Ding and Tao, 2007), performed with traditional statistical methods that will parse an increasing number of observations (Fodor, 2002). Dimension reduction creates a more effective domain characterization (Bi et al., 2003). Sufficient Dimension Reduction (SDR) is a generalization of nonlinear regression problems, where the extraction of features is as important as the matrix factorization (Globerson and Naftali. 2003), while SDR (Semi-Supervised Dimension Reduction) is used to maintain the original structure of high dimensional data (Zhang, et al., 2008).



Outlier, which often also be interpreted as an anomaly, is a set of data that is considered to have different properties compared with other data. Outlier analysis is also known as anomaly analysis or anomaly detection, or deviation detection (object attribute values they will, significantly different from other object attribute values). Outlier test approach based on density-based approach, where the outlier is a point located in areas with low density. To find outliers can use the formula:

$$density(x, k) = \left( \frac{\sum_{y \in N(x, k)} dist(x, y)}{|N(x, k)|} \right)^{-1}$$

Where  $N(x, k)$  is the set containing the  $k$  nearest neighbours  $x$ ,  $y$  is the nearest neighbour of  $x$  and  $|N(x, k)|$  is the number of members of the set  $N(x, k)$ . Meanwhile, to calculate the LOF (Local Outlier Factor) can be done with the approach:

$$average\_relative\_density(x, k) = \frac{density(x, k)}{\sum_{y \in N(x, k)} density(y, k) / |N(x, k)|}$$

## 2.5 DENSITY BASED CLUSTERING

Cluster analysis is a quite popular method of discretizing the data (Han and Micheline, 2006). Cluster analysis performed with multivariate statistics, identifies objects that have similarities and separate from the other object, so the variation between objects in a group smaller than the variation with objects in other groups.

Cluster analysis consists of several stages, beginning with the separation of objects into a cluster or group, followed by appropriate to interpret each characteristic value contained within their objects, and labelled of each group. The next stage is to validate the results of the cluster, using discriminant function.

Density-based clustering method calculating the distance to the nearest neighbour object, object measured with the objects of the local neighbourhood, if inter-object close relative with its neighbour said as normal object, and vice versa.

Density-based methods assume that the points that belong to each cluster are drawn from a specific probability distribution (Banfield and Raftery, 1993). The overall distribution of the data is assumed to be a mixture of several distributions (Maimon, 2005).

Density based clustering differentiates regions which have higher density than its neighbourhood and does not need the number of clusters as an input parameter. Regarding a termination condition, two parameters indicate when the expansion of clusters should terminate: given the radius of the volume of data points to look for,  $\epsilon$ , a minimum number of points for the density calculations,  $\rho$ , has to be exceeded (Bicici and Deniz, 2007). For a broad range of data distribution and distance measure, the relative contrast does diminish as the dimensionality increase (Houle et al., 2010).

DBSCAN is most popular algorithm based on density based clustering, relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape (Ester et al., 1996). DBSCAN has some advantages, one of them is it does not require knowledge of the number of clusters in the data a priori and can find arbitrarily shaped clusters. Another advantage is that DBSCAN has a notion of noise, and requires just two parameters. However there are disadvantages of DBSCAN; it can only result in a good clustering as good as its distance measure is in the function  $regionQuery(P, \epsilon)$ , and for high-dimensional data, it is difficult to find an appropriate value for  $\epsilon$ . Furthermore, it cannot cluster data sets well with large differences in densities.

Based on DBSCAN algorithm to assign cluster memberships, Ankerst proposed OPTICS (Ordering Points To Identify the Clustering Structure). OPTICS algorithm creates an augmented ordering of the database representing its density-based clustering structure (Aggarwal et al., 1999). OPTICS use DBSCAN algorithm to assign cluster memberships, and produce information of core distance a reachability-distance

reachability-distances of OPTICS, with  $r(p_1,o)$ ,  $r(p_2,o)$  for  $MinPts=4$ , as shown in Figure 2.11.

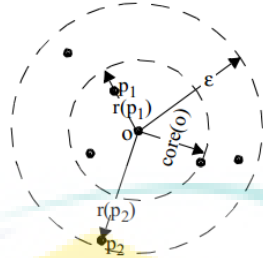


Figure 2.11. Core distance of OPTICS

Source : Aggarwal et al., 1999

Pseudocode of basic OPTICS shown in Figure 2.12 and Procedure *ExpandClusterOrder* shown in Figure 2.13.

```

OPTICS (SetOfObjects,  $\epsilon$ , MinPts, OrderedFile)
OrderedFile.open();
FOR i FROM 1 TO SetOfObjects.size DO
  Object := SetOfObjects.get(i);
  IF NOT Object.Processed THEN
    ExpandClusterOrder(SetOfObjects, Object,  $\epsilon$ ,
      MinPts, OrderedFile)
  OrderedFile.close();
END; // OPTICS

```

Figure 2.12. Pseudo code of basic OPTICS

Source : Aggarwal et al., 1999

```

ExpandClusterOrder(SetOfObjects, Object,  $\epsilon$ , MinPts,
OrderedFile);
neighbors := SetOfObjects.neighbors(Object,  $\epsilon$ );
Object.Processed := TRUE;
Object.reachability_distance := UNDEFINED;
Object.setCoreDistance(neighbors,  $\epsilon$ , MinPts);
OrderedFile.write(Object);
IF Object.core_distance  $\neq$  UNDEFINED THEN
  OrderSeeds.update(neighbors, Object);
  WHILE NOT OrderSeeds.empty() DO
    currentObject := OrderSeeds.next();
    neighbors:=SetOfObjects.neighbors(currentObject,  $\epsilon$ );
    currentObject.Processed := TRUE;
    currentObject.setCoreDistance(neighbors,  $\epsilon$ , MinPts);
    OrderedFile.write(currentObject);
    IF currentObject.core_distance $\neq$ UNDEFINED THEN
      OrderSeeds.update(neighbors, currentObject);
  END; // ExpandClusterOrder

```

Figure 2.13. Procedure ExpandClusterOrder of OPTICS

Source : Aggarwal et al., 1999

## 2.6 SUBSPACE BASED CLUSTERING

Clustering has been used extensively as a primary tool for data mining, but do not scale well to cluster high dimensional data sets in terms of effectiveness and efficiency, because of the inherent sparsity of high dimensional data. Problem arises when the distance between any two data points becomes almost the same (Moise and Sander, 2008), therefore it is difficult to differentiate similar data points from dissimilar ones. Secondly, clusters are embedded in the subspaces of the high dimensional data space, and different clusters may exist in different subspaces of

different dimensions (Gan et al., 2006). Techniques for clustering high dimensional data have included both feature transformation and feature selection techniques (Parson et al., 2004).

When clustering high dimensional data there are two main problems (Kriegel et al., 2009a), first is how to search for the relevant subspace and detection of the final cluster. Figure 2.14 illustrate the heuristic for both problems to develop solution.

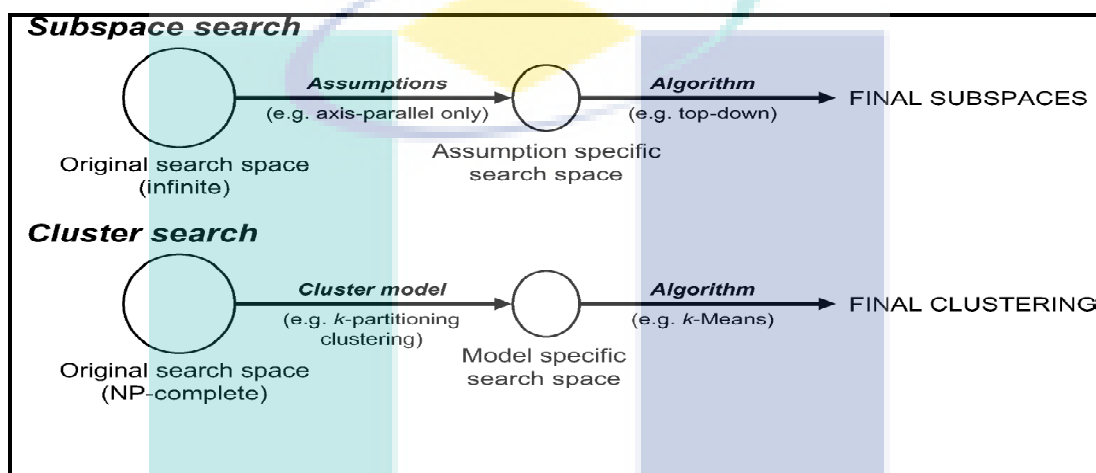


Figure 2.14. Illustration of the two general problems of clustering high-dimensional data

Source : Kriegel et al., 2009a

As known, no meaningful cluster analysis is possible unless a meaningful measure of distance or proximity between pairs of data points have been established. Most of the clusters can be identified by their location or density characters (Wang et al., 2007). There is a general categorization for high dimensional data set clustering: dimension reduction, parsimonious models, and subspace clustering. A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present (Steinbach et al., 2003).

Distance functions have been used in various dimensional clustering algorithms, depending on the particular problem being solved. Manhattan segmental distance is used in PROCLUS that is defined relative to some set of dimension (Moise and Sander, 2008). Employing the segmental distance as opposed to the traditional Manhattan distance is useful when comparing points in two different clusters that have varying number of dimension, because the number of dimension has been normalized.

Existing projected clustering algorithms are either based on the computation of  $k$  initial clusters in full dimensional space, or leverage the idea that clusters with as many relevant attributes as possible are preferable. Consequently, these algorithms are likely to be less effective in the practically most interesting case of projected clusters with very few relevant attributes, because the members of such clusters are likely to have low similarity in full dimensional space (Moise and Sander, 2008).

Generally, objects are represented as vectors or points contained in one or more dimensions. Cluster analysis performed to find groups (Parson et al, 2004), or patterns that are similar (Figure 2.15). Due to the increase of data needed to process 2 dimensions, by adding a dimension into 2 dimension in this phase, clustering process resulted outlier or noise (Figure 2.16). However, 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 (Figure 2.17), in this high dimensional datasets (3 or more dimension) will produced more outlier or noise, cluster should have intersection. Such clustering problem called "curse of dimensionality".

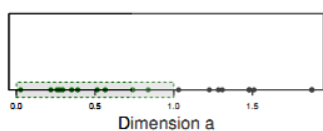


Figure 2.15. Data with 11 object in one bin

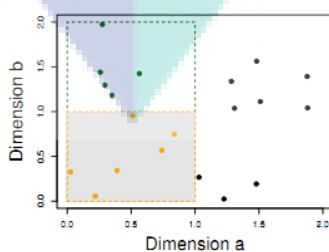


Figure 2.16. Data with 6 objects in one bin

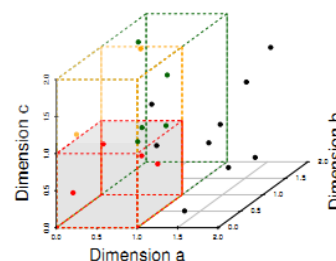


Figure 2.17. Data with 4 objects in one bin.

Source : Parson et.al, 2004

The distance between objects in high dimensional datasets are generally similar to each other. This fact will produce clusters that tend to be very tight, even coincide or overlap. Similarity judgments of objects are usually performed to detect clusters. Resemblance or similarity between objects is often determined by measuring the distance between objects in various dimensions. Subspace method is ideal to use for the case of high dimensional datasets. Subspace clustering is an extension of conventional clustering (Parson et al., 2004), is used to find second cluster, third and so on of the datasets, which are in different domains.

Figure 2.18 illustrates the need for subspace clustering. Subspace clustering is a method of detecting all groups in all subspaces (Stearn et al., 2011). It is possible, one point as a member of several groups that are on in a different subspace. This term is commonly used in high dimensional clustering.

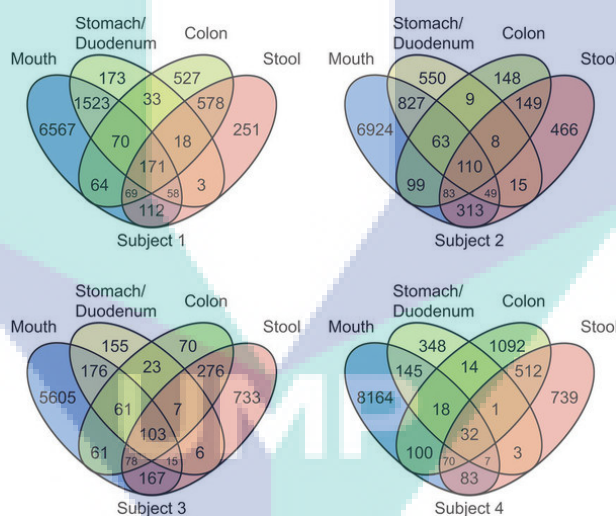


Figure 2.18. Cluster overlap each other

Source : Stearn et al., 2011

In datasets with high dimensions, it will cause problems, the cluster object in each cluster would be very difficult to recognize. If the dimension will be the elimination of multiple identities, then it can result in loss of data. So dimension reduction becomes irrelevant to do.

In two dimensions, three clusters can be formed. As in Figure 2.19, sample plot data in 2 dimension (a and b), two clusters properly separated, but a cluster remain mixed. It is possible sample plot data in 2 dimensions (b and c), two clusters properly separated, but still a mixed cluster (Figure 2.20). Besides it can also be produced with a clear separation of clusters (Figure 2.21), but still object overlap, and uneasy to separate using conventional clustering algorithms.

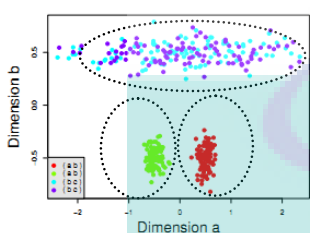


Figure 2.19. Sample data plot in 2 dimension (a and b).

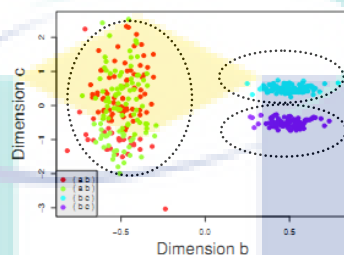


Figure 2.20. Sample data plot in 2 dimension (b and c).

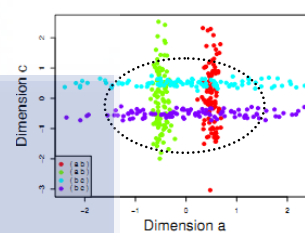


Figure 2.21. Sample data visible in 4 cluster.

Source : Agrawal et al., 2005

In reference to multidimensional data problem, let  $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$  be a set of finite bounded, totally ordered domains and  $\mathcal{S} = A_1 \times A_2 \times \dots \times A_n$  is  $n$ -dimensional numerical space. We will refer to  $A_1, A_2, \dots, A_n$  as the dimensions (attributes) of  $\mathcal{S}$ .

$$\mathcal{S} = \begin{bmatrix} A_1 \\ A_2 \\ \dots \\ A_m \end{bmatrix}, \mathcal{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \dots & \dots & \dots & \dots \\ A_{m1} & A_{m2} & \dots & A_{mn} \end{bmatrix}$$

Each unit  $u$  is the intersection of one interval from each attribute. A unit  $u$  has the form  $\{u_1, u_2, \dots, u_d\}$  where  $u_i = l_i h_i$  is a right-open interval in the partitioning of  $A_i$ . We say that a point  $v = \langle v_1, v_2, \dots, v_d \rangle$  is contained in a unit  $u = \{u_1, u_2, \dots, u_3\}$  if  $l_i \leq v_i \leq h_i$  for all  $u_i$ . The selectivity of a unit is defined to be the fraction of total data



points contained in the unit. We call a unit  $u$  dense if selectivity ( $u$ ) is greater than  $\tau$ , where the density threshold  $\tau$  is another input parameter. Consider a projection of the data set  $V$  into  $A_{t_1} \times A_{t_2} \times \dots \times A_{t_k}$ , where  $k < d$  and  $t_i < t_j$  if  $i < j$ . A unit in the subspace is the intersection of an interval from each of the  $k$  attributes.

A cluster is a maximal set of connected dense units in  $k$ -dimensions. Two  $k$ -dimensional unit's  $u_1, u_2$  are connected if they have a common face or if there exists another  $k$ -dimensional unit  $u_3$  such that  $u_1$  is connected to  $u_3$  and  $u_2$  is connected to  $u_3$ . Units  $u_1 = \{r_{t_1}, r_{t_2}, \dots, r_{t_k}\}$  and  $u_2 = \{r'_{t_1}, r'_{t_2}, \dots, r'_{t_k}\}$  have a common face if there are  $k - 1$  dimensions, assume dimensions  $A_{t_1}, A_{t_2}, \dots, A_{t_{k-1}}$ , such that  $r_{t_j} = r'_{t_j}$  and either  $h_{t_k} = l'_{t_k}$  or  $h'_{t_k} = l_{t_k}$ , for  $j \in \{1, 2, \dots, k - 1\}$ .

A region in  $k$  dimensions is an axis-parallel rectangular  $k$ -dimensional set. We are only interested in those regions that can be expressed as unions of units, henceforth all references to a region mean such unions. A region can be expressed as a Disjunctive Normal Form (DNF) expression on intervals of the domains  $A_i$ .

We say that a region  $R$  is contained in a cluster  $C$  if  $R \cap C = R$ . A region  $\mathcal{R}$  contained in a cluster  $C$  is said to be maximal if no proper superset of  $\mathcal{R}$  is contained in  $C$ . A minimal description of a cluster is a non-redundant covering of the cluster with maximal regions. That is, a minimal description of a cluster  $C$  is a set  $\mathcal{R}$  of maximal regions such that their union equals  $C$  but the union of any proper subset of  $\mathcal{R}$  does not equal  $C$ . While given a set of data points and the input parameters,  $\xi$  and  $\tau$ , find clusters in all subspaces of the original data space and present a minimal description of each cluster in the form of a DNF expression. To overcome this, it is necessary to place each object datasets into different subspaces. This becomes a challenge in data mining research, how to explore the data that has a multidimensional, and put every object into a separate cluster.

Number of possible axis-parallel subspaces where clusters could reside in the same dimensionality of the data space, the main task of research in the field is to develop of appropriate subspace search heuristics. There are two opposite basic

techniques for searching subspaces, namely bottom up subspace clustering and top down subspace clustering (Kriegel et al., 2009).

### 2. 6. 1 Bottom up Subspace Clustering

Bottom up subspace clustering starts from all one-dimensional subspaces that accommodate at least one cluster by employing a search strategy similar to frequent item set mining algorithms. CLIQUE is representative of bottom up subspace clustering.

CLIQUE (Kailing et al., 2004) identifies dense clusters in subspaces of maximum dimensionality. Once the appropriate subspaces are found, the task is to find clusters in the corresponding projections. The data points are separated according to the valleys of the density function. The clusters are unions of connected high density units within a subspace. It generates cluster descriptions in the form of DNF expressions that are minimized for ease of comprehension. It produces identical results irrespective of the order in which input records are presented and does not presume any specific mathematical form for data distribution.

CLIQUE starts by identifying subspaces that contain clusters. In this phase, it can find dense unit by initially determining 1-dimensional dense units by making a pass over the data. Having determined  $(k-1)$ -dimensional dense units, the candidate  $k$ -dimensional units are determined using the candidate generation procedure given below. While the procedure just described dramatically reduces the number of units that are tested for being dense, we still may have a computationally infeasible task at hand for high dimensional data. As the dimensionality of the subspaces considered increases, there is an explosion in the number of dense units, and so we need to prune the pool of candidates. The pruned set of dense units is then used to form the candidate units in the next level of the dense unit generation algorithm. After identifying subspace containing cluster, this is then followed by identification of clusters and generation of minimal description for the clusters.

SUBCLU (density connected SUBspace CLUstering) uses the concept of density-connectivity underlying the algorithm DBSCAN, SUBCLU is based on a formal

clustering notion (Kailing et al., 2004). In contrast to existing grid-based approaches, SUBCLU is able to detect arbitrarily shaped and positioned clusters in subspaces. The monotonicity of density-connectivity is used to efficiently prune subspaces in the process of generating all clusters in a bottom up way.

### 2. 6. 2 Top down Subspace Clustering

Top down subspace clustering methods analyse the full dimensional space to find patterns spotting clusters, where each database object multiple meaningful groupings might exist. The subspaces where clusters exist were identified based on the data distribution surrounding the patterns. Multi-resolution Correlation Cluster detection (MrCC) acts as a scalable method to detect correlation clusters in the range of around 5 to 30 axes (Cordeiro et al., 2010), while detecting alternative subspace clusters based on an already known subspace clustering. can detection of alternative subspace clusters, non-redundant cluster and has alternative cluster (Gunnemanns et al., 2009). Gunnemanns proposed ASCLU as alternative subspace clustering, his idea was based on a subspace cluster where  $C = (O; S)$  is a set of objects  $O \subseteq DB$  and a set of dimensions  $S \subseteq Dim$ . The objects  $O$  are similar within the relevant dimensions  $S$  while the dimensions  $Dim \setminus S$  are irrelevant for the cluster. K-means algorithm is able to generalize clustering for high dimensional data, as proposed in GKM (Generalized k-mean). GKM use advantages of k-mean arbitrarily, chooses  $k$  data points in  $X$  as the initial cluster centres, and each cluster centre  $i$  of  $C$  is associated with a vector  $i$  of  $W$  whose components equal one, then repeats steps to optimize the objective function  $E(W, C)$ .

### 2. 6. 3 Density Based Subspace Clustering Concept

Subspace clustering is a method to determine the clusters that form one different subspace. This method is better in handling multi-dimensional data (than other methods) shows the two dimensions of the clusters placed in a different subspace. On the dimension of the subspace cluster  $c_{a1}$  in the subspace  $c_c$ , and  $c_d$  in the subspace  $\{y\}$  can be found. Meanwhile,  $c_c$  is included in the cluster,  $c_{ab}$  and  $c_{ad}$  are identified as clusters Figure 2.22.

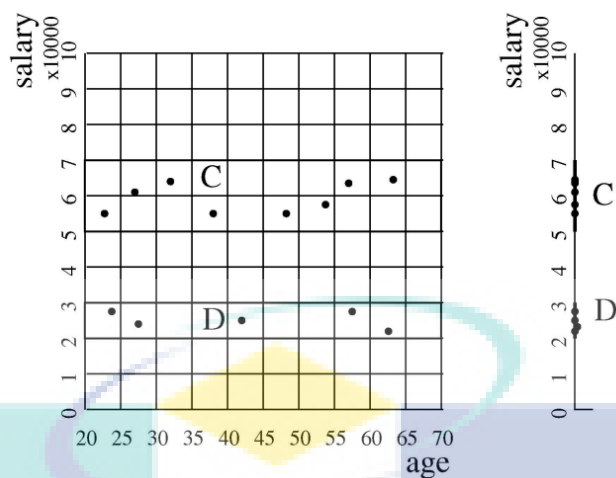


Figure 2.22. Subspace Clustering

Source : Agrawal et al., 2005

SC2D (Subspace Clustering with Dimensional Density) puts objects into the same cluster if they have similar dimensional density. Then clusters are separated from each other if there is more than one cluster in the same subspace. The algorithm starts from calculating the dimensional density of each object (Huang et al., 2010). DBSCAN is used to cluster dimensional density, and the clustering results figure out the objects which have the similar dimensional density. Due to the fact that objects in the same cluster means they are in the same subspace and refinement the result. It is obvious that two or more different clusters may exist in the same subspace. Thus, it is necessary to check this situation by calculating the intra-cluster compactness of clusters.

## 2.7 EDUCATIONAL DATA MINING

Today's rapid technological development has led to the exponential growth of human needs, many aspects of human life increase dependent with technological change, as in the field of medical, financial, even up to education. This fact requires educational institution to perform continuous efforts to adapt information from industries. Both educators and employers have recognized the importance of educating

the professionals that design, develop, and deploy information systems (Callahan and Bob, 2007).

Therefore, we can use data mining to recognize the relationship between educational institutions and industries, data mining application in education is also called the Educational Data Mining. Educational Data Mining (EDM) is the process of converting data from educational systems to useful information that can be used by educational software developers, students, teachers, parents, and other educational researchers (Baker and Yacef, 2009). Many studies discuss EDM, also required skill in the industry for college graduates, such as data mining for students' needs based on the improved RFM model (Bin et al., 2008), educating experienced IT professionals by addressing industry's needs (Callahan and Bob, 2007), jobs for young university graduates (Cardoso, 2007), measure the relationship between training and job performance (Devaraj and Babu, 2004).

Many authors consider research of EDM, such as needs of the mining engineering sector, education market demand, jobs, workplace performance, international job competitiveness, search job in academia, and students performance (Heuchan, 2003.; Hsia et al., 2008; Jones et al., 2009; McEntire et al., 2006; Moser et al., 2008.; Richardson, 2010; Wang et al., 2007; Yusof and Rukaini, 2005).

Some research predicting student performance through approach to classifying students in order to predict their final grade based on features extracted from logged do to in an education web-based system (Bidgoli et.al.2003), then assessing the recommended as well as personalized plans of study in terms of their affect on students' performance using data mining techniques (Siddiqui and Shehab, 2013). Another research discover meaningful, interpretable, and relevant patterns by injecting research contexts and domain knowledge applying an Educational Data Mining (EDM) technology to former students curricula and their degree of success (GPA) and their education (Knauf et.al., 2012).

For a simple case, student classroom seat choices, and focus on the issue of deployment stability, which denotes the student unwillingness to move from the usual

classroom seat, then examine different aspects of the deployment stability and introduce the necessary metric (Ivancevic et.al., 2012).

Educational Data Mining also use classification and clustering techniques, to predict the learning style of the peer learners based on the activities they have completed in the teaching learning activity of a particular course (Chellatamalin et.al.2011). Using spectral clustering able to further improve the student performance through prediction test score (Trivedi et.al., 2011), in other research data clustering combined with Neural Networks (NN) enables academicians to predict students' GPA according to their foreign language performance at a first stage, then classify the student in a well-defined cluster for further advising and follow up by forming a new system entry (Chady, 2011), as well as extract useful knowledge from graduate students (Tair, 2012).

Classification via clustering of student participation in forums can predict final marks based on student (Lopez et.al.2012), while cluster analysis can be used to help researchers develop profiles that are grounded in learner activity (Antonenko et.al., 2012), and range of clustering and sequential pattern mining and a theory driven approach can extract the mirror information (Prasad et.al., 2012).

Many clustering method used in Educational Data Mining, such as spectral clustering as a graph theoretic technique for metric modification to notion of similarity between data point (Trivedi et.al., 2011). In other research, a statistical model (a mixture of probability distributions) using LCA (Latent Class Analysis) can postulated for the population based on a set of sample data (Magidson and Vermunt, 2004).

Table 2.1 present research summarizes using of clustering for Educational Data Mining (EDM) (Xu et.al., 2013).

Table 2.1 Summarizes using of clustering for Educational Data Mining (EDM)

Clustering method	Authors / year	Topics
SOM (Self Organizing Map)	Durfee et.al. (2007)	Evaluating students computer based learning
EM (Expectation–Maximization)	Anaya and Boticario (2009)	Collaborative learning
ISODATA (Iterative Self Organizing Data Analysis Technique)	Wang et.al. (2004)	Learning Portfolio Analysis
Step wise HMM (Hidden Markov Models)	Shih et.al. (2010)	Discovery of Student Learning Tactics
Hierarchical cluster, K-,means	Hubshcer et.al. (2007)	Domain specific interactive data mining
K-means, EM ((Expectation–Maximization)	Mauil et.al. (2010)	Online curriculum planning behavior of teachers
PCA (Principal Component Analysis) over SOM (Self Organizing Map) K-means	Lee (2007)	Data mining in integrated learning environments
K-means	Dogan and Camurcu (2008)	Clustering of multidimensional Educational data
K-means	Perera et.al. (2009)	Pattern mining of online collaborative learning data

## 2.8 PERFORMANCE EVALUATION

The huge amount and using a vary data type in business and scientific fields caused explosive data collection, data mining can use to analyse useful knowledge from it. Nevertheless huge size of data and amount of computation involved in data mining, high-performance computing comes as a challenge and an essential component to ensure successful data mining application.

This challenge due to several factors i.e. (Grossman, 1996):

a. Many records.

Data became fundamental and put in its regimes, the most important challenge are accuracy and resource using (time, memory)

b. Many attributes.

Since usage of high dimension and multidimensional data became sparse and difficult to identify the boundaries

c. Many locations.

Mobile data and web application caused data became mobile too and need to measure its trading accuracy.

d. Many predictive models

e. Ease of use.

Main issue is how to simplify the task of building model from learning set.

Due to above challenge, performance evaluation of data mining became very important. The prediction of correct number of clusters unsupervised learning process is a hurdle, nevertheless can be cleared by using cluster validity indices to assess the quality of the clusters. Validation is the importance way of a cluster clarifies for its quality criterion, since connected operators are able to extract the number of clusters present. The clusters can be evaluated based on two criteria, compactness and separation (Visvanathan, 2009). Some method has proposed, Yang presents Vapnik–Chervonenkis-Bound (VB) index, estimated based on the structural risk minimization (SRM) principle, which optimizes the bound simultaneously over both the distortion function of empirical risk and the VC-dimension of model complex (Yang and Xindong, 2006).

Cluster validation refers to the quantitative evaluation of the quality of a clustering solution. In general, there are three approaches to investigate cluster quality; there are internal criteria, external criteria and relative criteria for the investigation of cluster quality.



Performance evaluation consider to validation and interpretation of the results (Halkidi et al., 2001). In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analysis in order to draw the right conclusion.



Figure 2.23. Purity as an external evaluation criterion for cluster quality.

Source : Manning et al., 2009

Cluster purity is a simple and transparent evaluation measure (Manning, 2009). To compute purity normally  $purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$ , where  $\Omega = \{\omega_1, \omega_2, \dots, \omega_K\}$  is the set of cluster and  $\mathbb{C} = \{c_1, c_2, \dots, c_j\}$  the set of classes. Each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of this assignment is measured by counting the number of correctly assigned documents and dividing by N.

From Figure 2.23 we can see, class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and  $\diamond$ , 3 (cluster 3), so the purity is  $(1/17) \times (5 + 4 + 3) \approx 0.71$ .

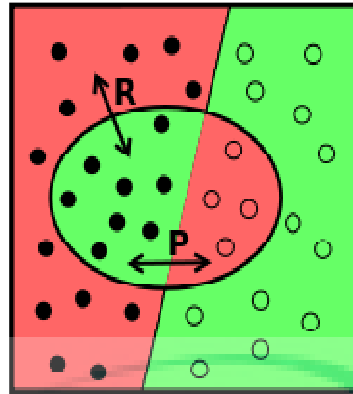


Figure 2.24. Precision (P) and Recall (R)

Another performance evaluation is to measure Precision (P) and Recall (R). From Figure 2.24, we can see the relevant items are to the left of the straight line while the retrieved items are within the oval. The red regions represent errors. On the left these are the relevant items not retrieved (false negatives), while on the right they are the retrieved items that are not relevant (false positives). Precision is the fraction of retrieved documents that are relevant to the search, can calculate as  $precision = \frac{|{\text{relevant data}} \cap {\text{retrieved data}}|}{|{\text{retrieved data}}|}$ .

	actual (observation)	class
predicted class (expectation)	tp (true positive) Correct result	fp (false positive) Unexpected result
	fn (false negative) Missing result	tn (true negative) Correct absence of result

Figure 2.25. Illustration of true false of expectation

While Recall is the fraction of the documents that are relevant to the query that are successfully retrieved, can calculate as  $recall = \frac{|{\text{relevant data}} \cap {\text{retrieved data}}|}{|{\text{relevant data}}|}$ .

Figure 2.25 illustrate the terms *positive* and *negative* refer to the classifier's prediction (sometimes known as the *expectation*), and the terms *true* and *false* refer to whether that prediction corresponds to the external judgment (sometimes known as the *observation*). Figure 2.25 show  $Precision = \frac{tp}{tp+fp}$ , and  $Recall = \frac{tp}{tp+fn}$ , and accuracy calculate as  $Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$ .

Another performance evaluation for clustering is F1 score, can be interpreted as a weighted average of the precision and recall, where an F<sub>1</sub> score reaches its best value at 1 and worst score at 0. From calculation of Precision and Recall, we can find the harmonic mean of Precision and Recall, named as F1 score, with formula 1 =

$$2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

## 2.9 SUMMARY

On the larger amount of data become more difficult to process it into useful information. Data mining as part of knowledge discovery has to be used to solve this problem. This technique increased and implemented in various area, such as health, medical, marketing, sales, medical, financial, e-commerce, multimedia, security, and lately developed for educational purposes too. Classification and clustering are two most important areas of data mining. The clustering process is perceived as an unsupervised process to identify groups of unlabelled data. Clustering high-dimensional data such as subspace clustering is one challenge approaches to solve the problem in data mining.

## CHAPTER 3

### METHODOLOGY

This chapter discusses the research methodology which makes up the entire research programme, and analyses the problem statement generated. Based on research problem, the research methods were tested and justified, and the data was analyzed as a whole. The methodology evolved during the initial research, which began with the investigation of the literature and new learning in data mining research, followed with the proposed methods. The literature enlightened the researcher's understanding of practical and effective methods for researching multidimensional data mining in education. The proposed method of clustering to allow proper adaptation problem as the focus of this research.

#### 3.1 RESEARCH DESIGN

Literature review of this research is the learning of the various techniques in data mining that were used to conduct research design, proposed algorithm, tests, experiments, surveys, and critical studies. The research literature of the data mining research and its application contains numerous discussions of curse of dimensionality.

This research makes on key studies from these literatures to establish the issue. Critical review and analyses were drawn from the literature as a starting point to highlight the design and development of subspace clustering based on density

connection. Educational data mining apply in a required skill competence. The critique of these case studies also provided a literature basis for specifying criteria for developing the research instrument. These data gathering and analysis instruments were employed for testing and training session in experimental phases. Emergent themes and relationship among data and proposed models were analysed and documented.

This research focused primarily on understanding values of subspace clustering based on density connection, and suggesting a proposed model. The theoretical development of values embedded in research process also required some of evaluation and validity. Figure 3.1 characterizes the research design of this thesis.

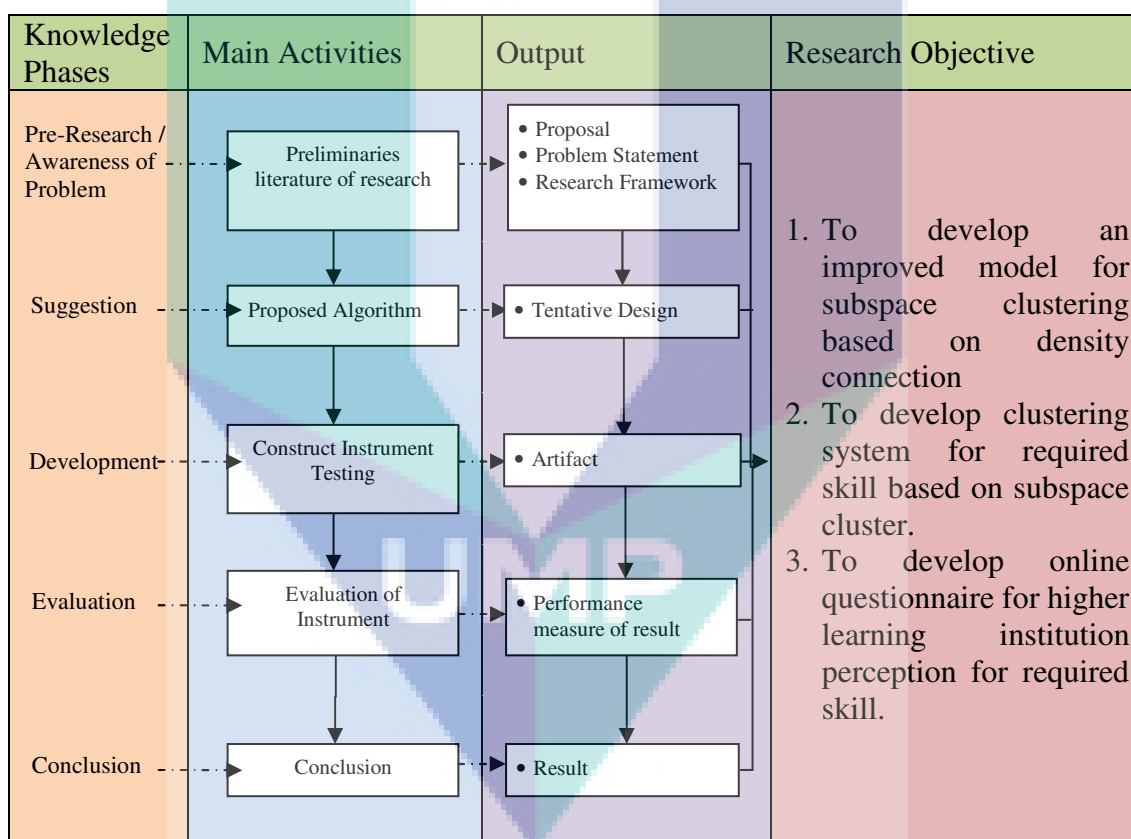


Figure 3.1. Research Design

### 3. 1. 1 Pre-Research / Awareness of Problem

Rapid development of computer and information technology in the last decade, create a huge amount of data in science, social life, and engineering. These data were exactly generated in massive scale, being either stored in gigantic storage devices, or growing through the system in the form of data streams (Han, research challenge in DM). Moreover, such data has made widely available, e.g., via the Internet, even PIPA (Protect IP Act) and SOPA (Stop Online Piracy Act) implement.

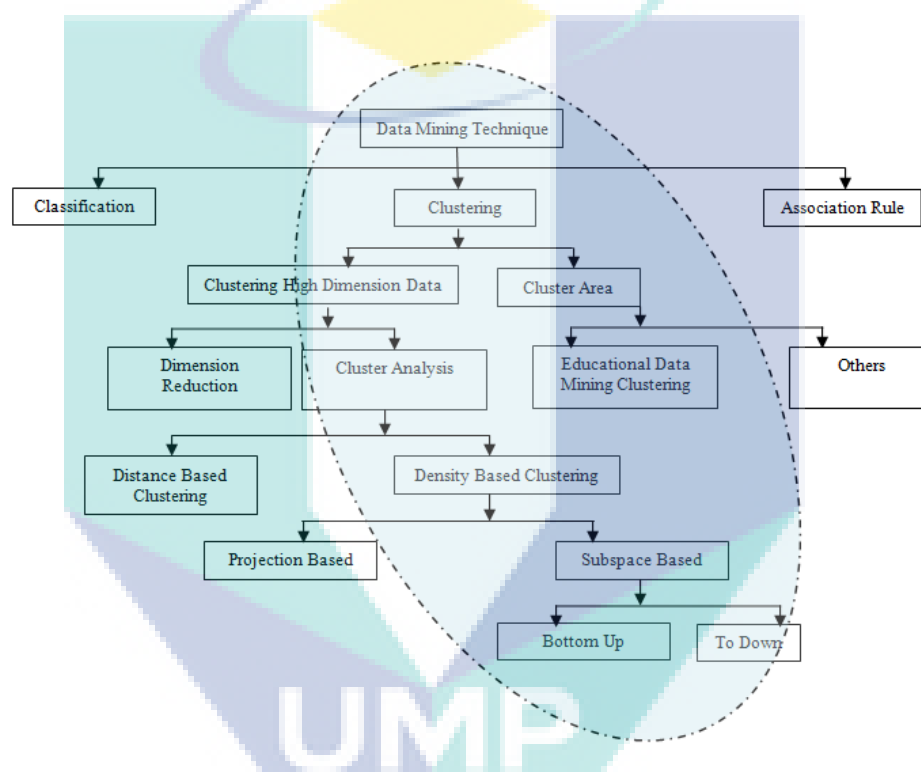


Figure 3.2. Path of literature review

Literature review as shown in Figure 3.2 begin with data mining; there are some techniques, this research focus on clustering, shown in ellipse area. Cluster dimension is the use as a basis to solve problem statement. Subspace clustering based on density based is a main research focus. The output of this phase is a proposal, problem statement, and research methodology.

Based on the awareness of the current problem, the researcher then proposed a research, asserting that clustering algorithms measure the similarity between data points by considering all features/attributes of a data set in high dimensional data sets and tend to break down both in terms of accuracy, as well as efficiency. Meanwhile, the dimensionality increases, the farthest neighbour of a data point is expected to be almost as close as its nearest neighbour for a wide range of data distributions and distance functions.

### 3. 1. 2 Suggestion

The suggestion phase mention behind the proposal and connected with proposal and tentative design. In this research we suggest an improved configuration of either existing or new clustering algorithms. As known clustering is the subject of active data mining research in several fields such as statistics, pattern recognition, and machine learning.

Clustering is a division of data into groups of similar objects, each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Many algorithms are developed, primarily in machine learning, that either have theoretical significance. In case multidimensional data, traditional clustering algorithms have no clear cluster. We proposed a new algorithm subspace clustering implement for multidimensional data.

### 3. 1. 3 Development

The design was implemented in this phase. The clustering techniques for implementation of high dimensional data will vary depending on the artefact to be constructed.

A clustering algorithm may require construction of a formal proof. An expert system embodying novel assumptions about human cognition in an area of interest will require software development, probably using a high-level package or tool. The novelty

of subspace clustering algorithm is primarily in the design, not the construction of the artefact.

#### 3. 1. 4 Evaluation

Once subspace clustering constructed, the artefact is evaluated according to criteria that are always implicit and frequently made explicit in the Proposal. Deviations from expectations, both quantitative and qualitative are carefully noted and *must be tentatively explained*. That is, the evaluation phase contains an analytic of several subspace clustering algorithm are made about the behaviour of the artefact.

This phase exposes an advantages and interpretation of the positivist stance. At an equivalent point in positivist research, analysis either confirms or contradicts between several algorithms. Essentially, save for some consideration of future work as may be indicated by experimental results, the research effort is over. For the design science researcher, by contrast, things are just getting interesting. Instead, the evaluation phase results and additional information gained in the construction and running of the artefact are brought together and fed back to another round of suggestion. This suggests a new design, frequently preceded by new library research in directions suggested by deviations from theoretical performance.

#### 3. 1. 5 Summary

This phase is the finale of a specific research effort. Typically, it is the result of satisfying, that is, though there are still deviations in the behaviour of the artefact from the initial problem, the results are adjudged “good enough”.

Not only are the results of the effort consolidated and written up at this phase. The knowledge gained in the effort is frequently categorized as either firm, facts that have been learned and can be repeatable applied or behaviour.



## 3.2 RESEARCH FRAMEWORK

Figure 3.3 characterizes the path of research design this research proposed with the research framework below:

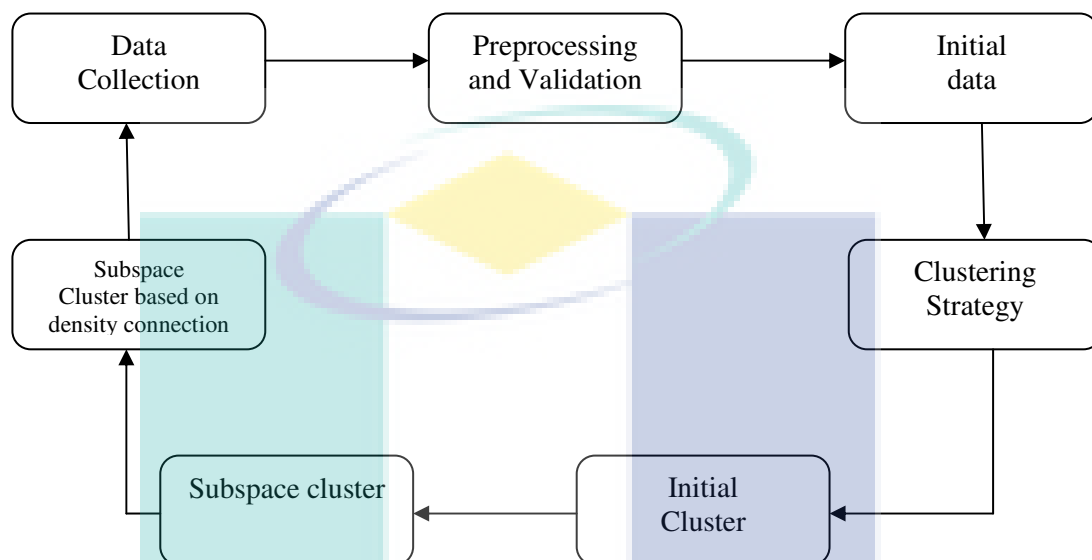


Figure 3.3. Research Framework

### 3.2.1 Data Collection

This research use benchmark real world data from the UCI archive, shown in Table 3.1 (Asuncion, 2007), these data sets have been used in some publications (Layton and Gales, 2004.; Matthew, 1998.; Assent et al., 2008.; Moise and Sander, 2008.; Sequeira and Zaki, 2004.; Michalski et al., 1986.; Clark and Niblett, 1987.; Tan and Eshelman, 1988.; Cestnik et al. 1987).

Table 3.1 Benchmark Real World Data Set

Dataset	Attributes	No of data	Types	Sources
Glass	10	214	CSV	<a href="http://repository.seasr.org/Datasets/UCI/csv/">http://repository.seasr.org/Datasets/UCI/csv/</a>
Liver-dis	7	345	CSV	<a href="http://dme.rwth-aachen.de/sites/default/files/public_files/data.zip">http://dme.rwth-aachen.de/sites/default/files/public_files/data.zip</a>
Job satisfaction	11	990	CSV	<a href="http://brainmass.com/statistics/all-topics/62117">http://brainmass.com/statistics/all-topics/62117</a>

Due to research problem, this research use main real dataset from education survey, as shown in Table 3.2.

Table 3.2 Number of sample in each respondent group

No	Respondent	Total
1	Student Industrial Training (SIT)	100
2	Industry	71

The samples of this study are student industrial training from three study program of FSKKP University Malaysia Pahang. The sampling method that was used in this study is stratified random sampling technique. The goal in stratified sampling is to form groups or strata of units within a stratum, the units are very similar on the characteristic of interest.

Based on Krecjie and Daryle (1970), table for determining needed size  $s$  of a randomly chosen sample from a given finite population of  $n$  cases, showed that industry population sample is 110 and the appropriate sample for it is 86 respondents (Krecjie and Daryle, 1970). The actual sample of this study is 71; this means that the sample is most adequate according to Krecjie and Morgan table. While, the population of SIT sample is 150, then according to Krecjie and Morgan's table, the adequate sample is 108. The actual sample of this study is 100, and then it means sample study is adequate too.

The use of a sequential explanatory mixed research design means that quantitative data were collected and analyzed before the researcher followed up with a qualitative approach (Creswell, 2009.; Creswell and Vicki, 2007).

A self-report questionnaire was used to collect data in the first phase. Information letters in English were mailed to the contact persons. To ensure a higher response rate, the information letters, consent form (offline and online), and

questionnaires were shared to each participant candidate during October 2010 until December 2011.

SIT Candidates who were willing to participate were asked to circle the number that most accurately represented the extent to which they experienced while industrial training. Industrial respondent also shared a similarly questionnaire, represented their perception about knowledge skill and soft skill needed for computer science graduated. After finishing their responses, for offline questionnaire, participants were asked to send questionnaire by email. While for online questionnaire, participants were asked to fill up questionnaire through website ([www.damira-ws.com](http://www.damira-ws.com)).

Several procedures were used to protect the right of participants in this research. Prospective participants received an information letter detailing the purpose of the research and the rights of the participants. All candidates were informed that participation was voluntary, and that they could withdraw from the study at any time or for any reasons without penalty. For ethical considerations, all information was disclosed only for the purpose of research. The survey process was anonymous to protect the privacy of the participant.

### 3. 2. 2 Pre-processing Data and Validation

The questionnaire of SIT is divided into section sections (see Appendix A). The first section acquires perception of student of frequency of course implemented with industrial training, sections two how are importance knowledge competence while industrial training, and last section contain of perception of importance of soft skills competence while industrial training. While questionnaire of industries divided into three sections (see Appendix B). The first section acquires employee profile, sections two how importance knowledge competence and last section contain of perception of importance of soft skills while working. By using SPSS, this research have validate each variable of student industrial data set, consist of 100 data shown result Cronbach Alpha greater than 0.800 (see Table 3.2), while validate each variable of industry data set, consist of 72 data shown result Cronbach Alpha greater than 0.800 (see Table 3.3).

Table 3.3 Student Industrial Training Dataset

Code	Question	Cronbach's Alpha	No of data	No of Items
	<i>Frequency of course implemented with industrial training environment</i>			
b1	University Course	.883	100	14
b2	Faculty Course	.927	100	18
b3	Program Course	.929	100	13
b4	Elective Course	.896	100	4
	<i>Importance knowledge competence</i>			
c1	Algorithm Capability	.861	100	4
c2	Application Program	.940	100	6
c3	Computer Programming	.950	100	4
c4	Hardware and Device	.954	100	8
c5	Human Computer Interaction	.808	100	2
c6	Information System	.930	100	5
c7	Information Management (Database	.971	100	7
c8	IT Resources Planning	.980	100	5
c9	Intelligent System	.966	100	2
c10	Network and Communication	.972	100	8
c11	System Development through Integration	.986	100	8
	<i>Importance of Soft Skills Competence</i>			
d1	Resources Management	.940	100	7
d2	Communication and Interpersonal	.924	100	9
d3	Leadership	.963	100	15
d4	Information Management	.971	100	18
d5	System Thinking	.974	100	6
d6	Technical and Functional Competence	.941	100	4

### 3. 2. 3 Initial Data

There are three issues with regard to sampling equivalence, which refers to whether competence samples can be compared. Those issues explain the operational definition of variables in this study.

Table 3.4 explain how that issue will use in this research, as follows:

- a. *Frequency of course implemented with industrial training environment* is defined as how SIT's perceptions for the use or benefit course while industrial training. There are four categories of course: University Course, Faculty Course, Program Course, and Elective Course. These course implemented in FSKKP Universiti Malaysia Pahang.
- b. *Importance knowledge competence* is defined as SIT's perceptions for the importance of knowledge while industrial training. There are 11 knowledge competences to be evaluate: Algorithm Capability, Application Program, Computer Programming, Hardware and Device, Human Computer Interaction, Information System, Information Management (Database), IT Resources Planning, Intelligent System, Network and Communication, and System Development through Integration (ACM and IEEE, 2005).
- c. *Importance soft skill competence* defined as SIT's perceptions for the importance of soft skill while industrial training. There are six soft skill competences to be evaluated: Resources Management, Communication and Interpersonal, Leadership, Information Management, System Thinking, and Technical and Functional Competence.

Table 3.4 Industrial Dataset

Code	Question	Cronbach's Alpha	No of data	No of Items
	<i>Employee Profile</i>			
b1	Education and Training	.016	72	2

Code	Question	Cronbach's Alpha	No of data	No of Items
b2	Licence and Certification	.828	72	2
b3	Computer Skill and Software Used	.831	72	3
	<i>Importance knowledge competence</i>			
c1	Algorithm Capability	.766	72	4
c2	Application Program	.901	72	6
c3	Computer Programming	.889	72	4
c4	Hardware and Device	.926	72	8
c5	Human Computer Interaction	.804	72	2
c6	Information System	.916	72	5
c7	Information Management (Database)	.954	72	7
c8	IT Resources Planning	.960	72	5
c9	Intelligent System	.699	72	2
c10	Network and Communication	.968	72	8
c11	System Development through Integration	.973	72	8
	<i>Importance of Soft Skills Competence</i>			
d1	Resources Management	.933	72	7
d2	Communication and Interpersonal	.956	72	9
d3	Leadership	.965	72	15
d4	Information Management	.977	72	18
d5	System Thinking	.975	72	6
d6	Technical and Functional Competence	.925	72	5

### 3. 2. 4 Clustering Strategy

Multidimensional data mining involves five step processes: deciding between census and sample data, identifying relationship within the data, modifying or

transforming data, developing a model that explains the data relationship, and testing the model's accuracy. To offer more detailed data to complete and best understand the research problem.

This research used three cluster strategies of multidimensional data mining analysis, including clustering analysis, subspace cluster, and subspace cluster based on density connection (Figure 3.4).

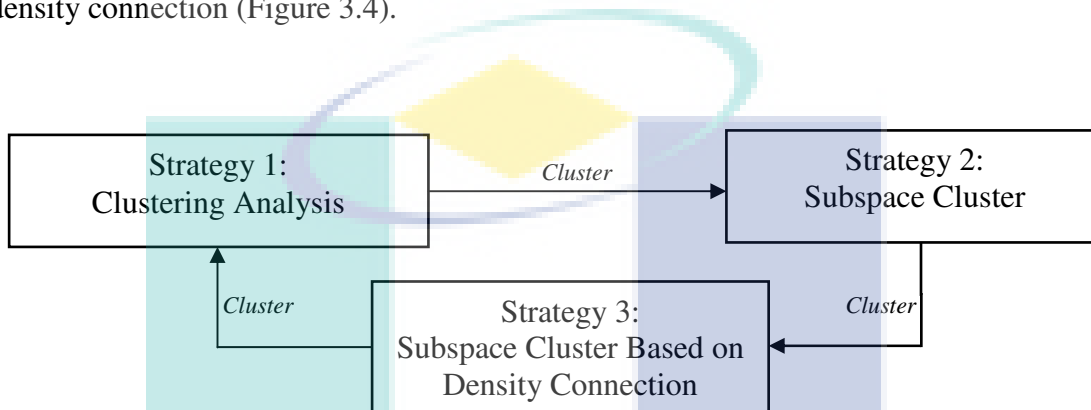


Figure 3.4. The strategy of multidimensional data mining analysis

### 3. 2. 5 Strategy 1 of Data Analysis – Clustering Analysis

Human beings are skilled at dividing objects into groups (known as clustering), assigning particular object to these groups (known as classification), and then doing prediction for certain objects. DBSCAN use in this research to understand member of objects for each data sets. These cluster prototypes can be used as the basis for a number of data analysis or data processing technique.

The purpose of the cluster initialization (Figure 3.5) is to determine an initial set of maximal subspace  $\alpha$ -clusters, based on the last  $W$  values of each streaming time series. The Cluster Initialization (CI) process comprises a series of steps. In the first step, each time instance (dimension) is inspected separately to determine simple  $\alpha$ -clusters (which are defined in one dimension only). Next, all clusters containing  $m=2$  streams in the maximum possible number of dimensions are generated.

In each subsequent step the algorithm tries to increase the number of streams per cluster ( $m=m+1$ ), until all possible maximal subspace  $\alpha$ -clusters are generated, according to the values of  $\alpha$ , minRows and minCols . Clusters that contain less than minCols dimensions are discarded permanently in each step of the algorithm, since they cannot contribute to the final answer.

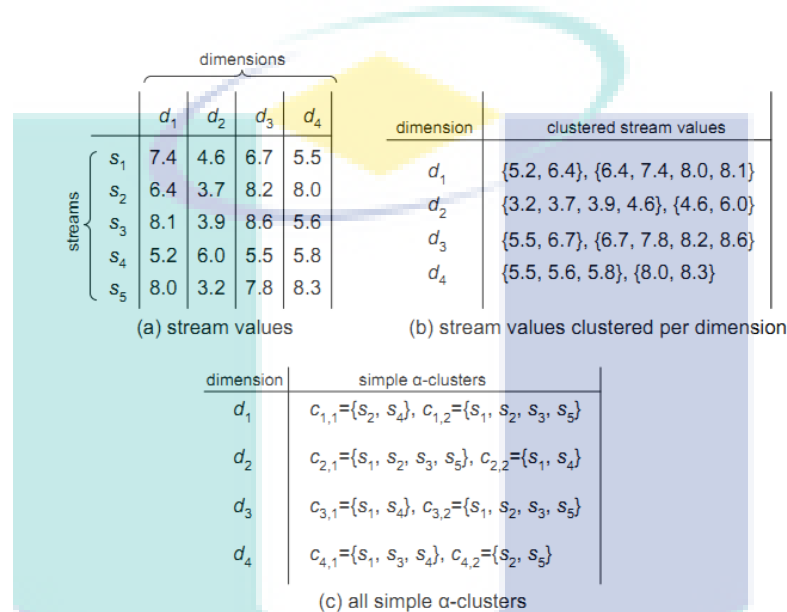


Figure 3.5. Cluster initialization

Clustering analysis groups data objects based only on information found in data sets that describe the objects and their relationship. Groups object's must be similar each other. But in fact, there are many situations in which an object could reasonably than placed in more than one cluster. In this case, subspace cluster should be used to solve the problem.

To overcome the curse of dimensionality in multidimensional data, in this research subspace cluster to identify a specific object and place in cluster related, to ensure that relation we use DBSCAN as a basis clustering technique. There are some subspace cluster technique, such as SUBCLU, FIRES and INSCY. These clustering will use to train data sets. We will train each data set until cluster result convergent.



As a novelty of this research we proposed an improved subspace clustering algorithm based on density based connection (named as DAMIRA), and use it as main clustering technique to job prediction of Higher Learning Institution.

### 3. 2. 6 Strategy 2 of Data Analysis – Subspace Cluster Analysis

Initial data to be cluster through implementation of subspace cluster, as operator learns from both nominal and numerical data.

### 3. 2. 7 Strategy 3 of Data Analysis – Subspace Cluster Based on Density Connection

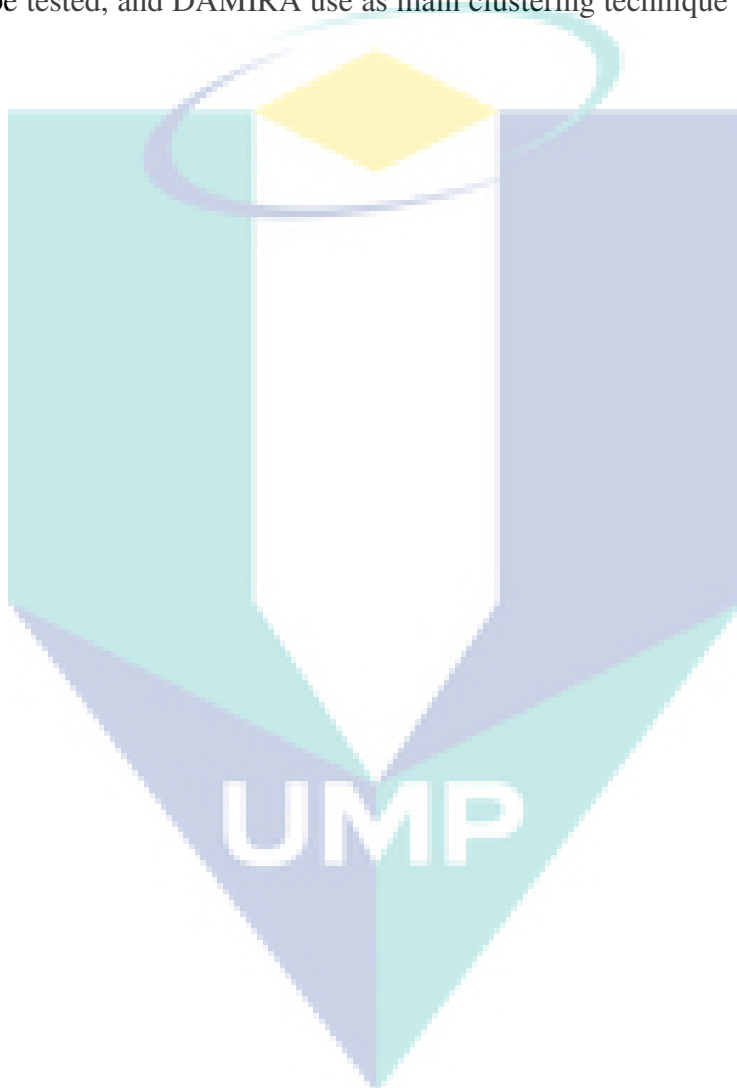
When rule has been found, this research will use density based connection to identify and to explore required skill. This result used as knowledge bases model to represent objects within information. Models capture relationships among many factors to allow assessment of potential associated with a particular set of conditions.

This research will develop an experiment for evaluating method. Numerical experiments will conduct in the PC (Pentium i5, CPU 3.66 GHz 2GB RAM) hardware environment based on and Windows 7. This research will compare on the six set testing. It will conduct a comparison of data mining clustering tools to select a suitable platform for experiment. Data collection for real data will use web-based data tool, via [www.damira-ws.com](http://www.damira-ws.com). MATLAB and/or PHP will use to simulate testing of high dimensional data clustering. WEKA and Rapid-Miner will use to training data to find out rule based analysis and prediction analysis.

## 3. 3 SUMMARY

The learning of the various techniques in data mining was used to conduct research design, and improved proposed algorithm. While an increasing using data mining on educational system is a motivation for this research. In this research we suggest an improved clustering algorithm, testing in use benchmark real world data from the UCI archive, and real dataset from education survey from SIT and industrial.

The questionnaire survey is divided into perception of student of how is importance knowledge competence while industrial training, and perception of importance of soft skills competence while industrial training. Clustering research by using clustering technique was applied to analyze the possible cluster between the knowledge competence skill and soft skill competence among student industrial training and industries. There are some subspace cluster technique, such as SUBCLU, FIRES and INSCY to be tested, and DAMIRA use as main clustering technique to based on density connection.



## CHAPTER 4

### SUBSPACE CLUSTER BASED ON DENSITY CONNECTION

This chapter described proposed technique to assess multidimensional data mining via subspace clustering also described preview of improved model for subspace clustering based on density connection, and how to cluster required skill based on subspace cluster.

#### 4.1 CLUSTERING STRATEGY

The main point of this thesis is the density-based clustering approach, in particular the concepts of density-connected clusters underlying the algorithms DBSCAN (Density-Based Spatial Clustering of Applications with Noise). We propose an improved technique to cope with the challenges clustering in educational data mining, named as DAMIRA (multidimensional DATA MINing subspace clusteRing Approach).

This data mining research started from a density based. As seen at images in Figure 4.1, easily identifies the cluster of points and also identify outliers that are formed (Ester et al., 1996). As a cluster these points have a closer density of points than the others. While point outside the group referred to as noise.

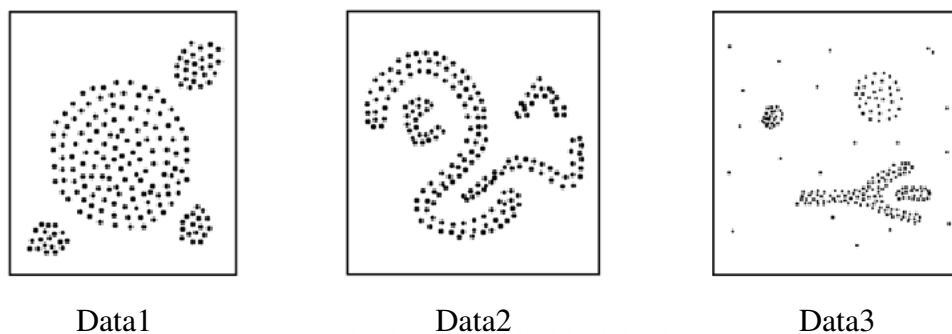


Figure 4.1. The cluster of points and also identify outliers

Source : Ester et al., 1996

The main idea of the density in each cluster is that any data has the minimum number of neighbouring data, where data density must more than a certain threshold. Neighbouring shape will be determined based on the function of the distance from point  $p$  and  $q$ , denoted as  $dist(p, q)$ .

Definition 1: (*Eps* – neighborhood of a point)

The *Eps* – neighborhood of a point  $p$ , denoted by  $N(p)$ , is defined by  $N_{Eps}(p) = \{q \in D | dist(p, q) \leq Eps\}$  (Ester et al., 1996).

There are a minimum number of points (*MinPts*) in the cluster within *Eps* – neighbourhood, then there will be two types of points in the cluster, core points and border points (Figure 4.2). The number of points on the border would be less than at the core, but should be set a minimum number of points included in the same cluster. This value, however, will not form the specific characteristics for each cluster, especially if there is noise. Therefore, it is required for each point  $p$  in cluster  $C$  and  $q$  in  $C$  so that  $p$  is in the *Eps* – neighborhood of  $q$  and  $N(q)$  contains at least *MinPts* number of points.

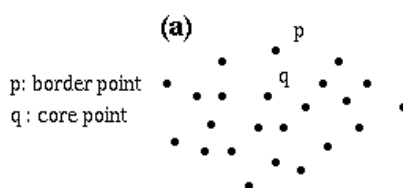


Figure 4.2. Border and core point

Definition 2: (*directly density – reachable*)

A point  $p$  is *directly density – reachable* (from a point  $q$  ( $Eps, MinPts$ )) if:

- a)  $p \in N_{Eps}(q)$  and
- b)  $|N_{Eps}(q)| \geq MinPts$  (core point condition).

Density-reachable is symmetric for pairs of core points, however it is not symmetric if one core point and one border point are involved (Ester et al., 1996).

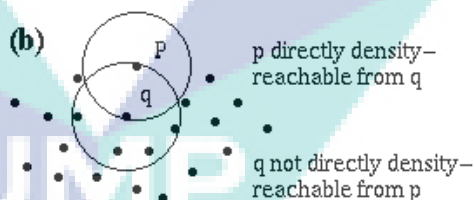


Figure 4.3. Density reachable

Definition 3: (*density-reachable*)

Point  $p$  is *density-reachable* from a point  $q$  ( $Eps$  and  $MinPts$ ) if there is a chain of points  $p_1, \dots, p_n$ ,  $p_1 = q$ ,  $p_n = p$  such that  $p_{i+1}$  is directly density-reachable from  $p_i$ .

Two border points of the same cluster  $C$  are possibly not density reachable from each other because the core point condition might not

hold for both of them. However, there must be a core point in  $C$  from which both border points of  $C$  are density-reachable.

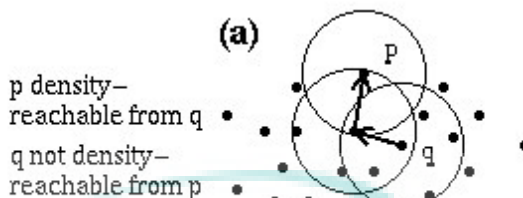


Figure 4.4. Another density reachable

Definition 4: (density-connected)

The notion of density-connectivity shown in (Figure 4.5) which covers this relation of border points. Point  $p$  is *density connected* to a point  $q$  ( $Eps$  and  $MinPts$ ) if there is a point  $o$  such that both,  $p$  and  $q$  are *density-reachable* from  $o$ .

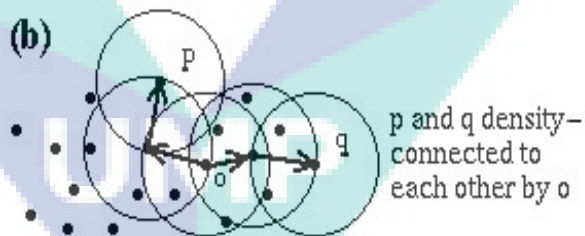


Figure 4.5. Density connected

A set of density connected points which is maximal (*density-reachability*) defined as a cluster.

Definition 5: (cluster)

Let  $D$  be a database of points. A cluster  $C$  ( $Eps$  and  $MinPts$ ) is a non-empty subset of  $D$  satisfying the following conditions:

1.  $\forall p, q$ : if  $p \in C$  and  $q$  is density-reachable from  $p$  ( $Eps$  and  $MinPts$ ), then  $q \in C$ . (Maximality)
2.  $\forall p, q \in C$  :  $p$  is density-connected to  $q$  ( $Eps$  and  $MinPts$ ). (Connectivity)

Refer to Definition 5, this research following term to define cluster:

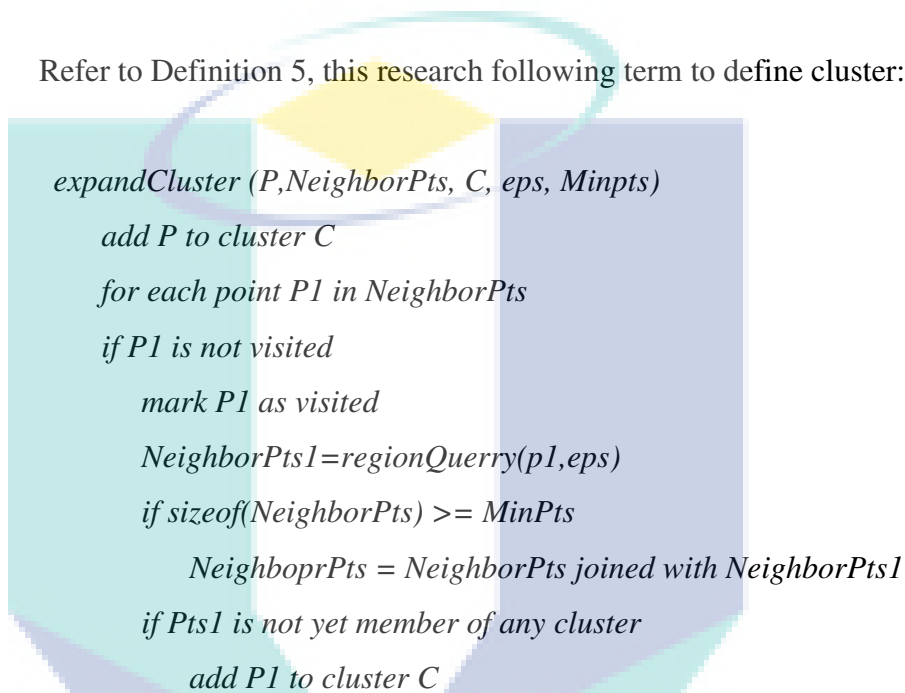


Figure 4.6. Define cluster

Definition 6: (noise)

Let  $C_1, \dots, C_k$  be the clusters of the database  $D$  (parameters  $Eps_i$  and  $MinPts_i, i = 1, \dots, k$ ). Then we define the *noise* as the set of points in the database  $D$  not belonging to any cluster  $C_i$ , i.e.  $noise = \{p \in D \mid \forall i: p \notin C_i\}$ .

Important lemmas to validate the correctness of clustering algorithm of DBSCAN stated that parameters  $Eps$  and  $MinPts$  can discover a cluster in a two-step. Begin with choose an arbitrary point from the database satisfying the core point, then retrieve all points that are density-reachable from cluster.

Lemma 1: Let  $p$  be a point in  $D$  and  $|NEps(p)| \geq MinPts$ . Then the set  $O = \{o \mid o \in D \text{ and } o \text{ is density-reachable from } p \text{ (} Eps \text{ and } MinPts)\}$  is a cluster.

It is not obvious that a cluster  $C$  is uniquely determined by any of its core points. However, each point in  $C$  is density-reachable from any of the core points of  $C$  and, therefore, a cluster  $C$  contains exactly the points which are density-reachable from an arbitrary core point of  $C$ .

Lemma 2: Let  $C$  be a cluster and let  $p$  be any point in  $C$  with  $|NEps(p)| \geq MinPts$ . Then  $C$  equals to the set  $O = \{o \mid o \text{ is density-reachable from } p\}$ .

## 4.2 SUBSPACE CLUSTERING

In line with technological developments, the application of biology, geography, and even education produces large amounts of data. The bigger the data to be processed require efficient methods and effective treatment. To this end, implementation of data mining becomes very necessary. One of the main functions of data mining is clustering. But in fact, traditional clustering algorithms often fail to identify the clusters correctly, too much noise, or perhaps the result to be biased. This failure is due to a data set of high-dimensional form, the data is inherently rare. To overcome this, the research cluster subspace becomes an important issue to be studied.

This research based on a bottom-up greedy algorithm to detect the density-connected clusters in all subspaces of high dimensional data. The algorithm begins with generating all 1 – *dimensional* clusters by applying DBSCAN to each 1 – *dimensional* subspace. For each detected cluster we have to check, whether this cluster is still existent in higher dimensional subspaces. No other clusters can exist in higher dimensional subspaces.



Definition 7: ( $\varepsilon$  – neighborhood)

Let  $\varepsilon \in \mathbb{R}$ ,  $S \subseteq \mathcal{A}$  and  $o \in DB$ . The  $\varepsilon$  – neighborhood of  $o$  in  $S$ , denoted by  $\mathcal{N}_\varepsilon^S(o)$ , is defined by:

$$\mathcal{N}_\varepsilon^S(o) = \{x \in DB \mid \text{dist}(\pi_S(o), \pi_S(x)) \leq \varepsilon\}.$$

Refer to Definition 7, this research use normalization.

for each  $P$  in datasets ( $D$ )

Mark  $P$  as visited

$\text{NeighborPts} = \text{neg}(\text{Dataset}, P, \text{eps})$

If ( $\text{count}(\text{neighborPts}) < \text{minPts}$ )

Mark  $P$  as noise

Figure 4.7. Normalization

Definition 8: (core object)

Let  $\varepsilon \in \mathbb{R}$ ,  $m \in \mathbb{N}$ , and  $S \subseteq \mathcal{A}$ . An object  $o \in DB$  is called core object in  $S$ , denoted by  $\text{Core}_{\varepsilon, m}^S(o)$ , if its  $\varepsilon$  – neighborhood in  $S$  contains at least  $m$  objects, formally:

$$\text{Core}_{\varepsilon, m}^S(o) \Leftrightarrow |\mathcal{N}_\varepsilon^S(o)| \geq m.$$

Usually clusters contain several core objects located inside a cluster and border objects located at the border of the cluster. In addition, objects within a cluster should be “connected”. Refer to Definition 8, this research use neighbour as member to define point and border point:

```
{
  $this->_data[$i];
  $this->expandCluster($this-
>_data[$i], $neighborPts, $cid, $eps, $minpts);
  $cid = count($this->perRegion());
}
```

Figure 4.8. Define point and border point

Definition 9: (*direct density – reachability*)

Let  $\varepsilon \in \mathbb{R}$ ,  $m \in \mathbb{N}$ , and  $S \subseteq \mathcal{A}$ . An object  $p \in DB$  is *directly density – reachable* from  $q \in DB$  in  $S$  if  $q$  is a core object in  $S$  and  $p$  is an element of  $N\varepsilon S(q)$ , formally:

$$DirReach_{\varepsilon,m}^S(q,p) \Leftrightarrow Core_{\varepsilon,m}^S(q) \wedge p \in \mathcal{N}_{\varepsilon}^S(q).$$

Refer to Definition 9, this research use distance as bases to define reachable each other point:

For each  $P1$  in NeighborPts  
 If ( $distance(p_1, p_2) \leq eps^2$ )  
 Mark  $P1$  as noise  
 Add  $P1$  to region

Figure 4.9. Density reachable

Definition 10: (*density – reachability*)

Let  $\varepsilon \in \mathbb{R}$ ,  $m \in \mathbb{N}$ , and  $S \subseteq \mathcal{A}$ . An object  $p \in DB$  is *density-reachable* from  $q \in DB$  in  $S$  if there is a chain of objects  $p_1, \dots, p_n$ ,  $p_1 = q$ ,  $p_n = p$  such that  $p_{i+1}$  is directly density-reachable from  $p_i$ , formally:

$$Reach_{\varepsilon,m}^S(q,p) \Leftrightarrow \exists p_1, \dots, p_n \in DB: p_1 = q \wedge p_n = p \wedge \forall i \in \{1 \dots n - 1\}: DirReach_{\varepsilon,m}^S(p_i, p_{i+1}).$$

Definition 11: (*density – connectivity*)

Let  $\varepsilon \in \mathbb{R}$ ,  $m \in \mathbb{N}$ , and  $S \subseteq \mathcal{A}$ . An object  $p \in DB$  is *density – connected* to an object  $q \in DB$  in  $S$  if there is an object  $o$  such that both  $p$  and  $q$  are *density – reachable* from  $o$ , formally:

$$Connect_{\varepsilon,m}^S(q,p) \exists o \in DB: Reach_{\varepsilon,m}^S(o,q) \wedge Reach_{\varepsilon,m}^S(o,p).$$

Refer to Definition 11, this research use distance function as bases to define connection of each other point:

```

Distance (p1,p2)
  Difx=p2.x-p1.x
  Dify=p2.y-p1.y
  Return Difx^+Dify^2

```

Figure 4.10. Define connection of each other point

Definition 12: (*density – connected set*)

Let  $\varepsilon \in \mathbb{R}$ ,  $m \in \mathbb{N}$ , and  $S \subseteq \mathcal{A}$ . A non-empty subset  $C \subseteq DB$  is called a *density – connected set* in  $S$  if all objects in  $C$  are *density – connected* in  $S$ , formally:

$$ConSet_{\varepsilon,m}^S(C) \Leftrightarrow \forall o, q \in C : Connect_{\varepsilon,m}^S(o, q).$$

A straightforward approach would be to run DBSCAN in all possible subspaces to detect all density-connected clusters. The problem is that the number of subspaces is  $2_d$ . A more effective strategy would be to use the clustering information of previous subspaces in the process of generating all clusters and drop all subspaces that cannot contain any density-connected clusters.

Unfortunately, density-connected clusters are not monotonic, i.e. if  $C \subseteq DB$  is a density-connected cluster in subspace  $S \subseteq A$ , it need not be a density-connected cluster in any  $T \subseteq S$ . The reason for this is that in  $T$  the density-connected cluster  $C$  need not be maximal w.r.t. *density – reachability* any more. There may be additional objects which are not in  $C$  but are density-reachable in  $T$  from an object in  $C$ .

However, density-connected sets are monotonic. In fact, if  $C \subseteq DB$  is a density-connected set in subspace  $S \subseteq A$  then  $C$  is also a density-connected set in any subspace  $T \subseteq S$ .

Lemma 3. (monotonicity)

Let  $\varepsilon \in \mathbb{R}$ ,  $m \in \mathbb{N}$ ,  $o, q \in DB$ ,  $C \subseteq DB$ , where  $C \neq \emptyset$  and  $S \subseteq A$ . Then the following monotonicity properties hold:

$\forall T \subseteq S$  :

- (1)  $Core_{\varepsilon, m}^S(o) \Rightarrow Core_{\varepsilon, m}^T(o)$
- (2)  $DirReach_{\varepsilon, m}^S(o, q) \Rightarrow DirReach_{\varepsilon, m}^T(o, q)$
- (3)  $Reach_{\varepsilon, m}^S(o, q) \Rightarrow Reach_{\varepsilon, m}^T(o, q)$
- (4)  $Connect_{\varepsilon, m}^S(o, q) \Rightarrow Connect_{\varepsilon, m}^T(o, q)$
- (5)  $ConSet_{\varepsilon, m}^S(o, q) \Rightarrow ConSet_{\varepsilon, m}^T(o, q)$

### 4.3 SUBSPACE CLUSTERING BASED ON DENSITY CONNECTION

This research use interface based on web to conduct training on this algorithm. The procedure prior to running the Step-1, which determines whether to use data sets that already exist, or can also add a new data set. If add the data set is selected then the user must select the type of data set types *csv* text file. Step-1 procedure is described in Figure 4.11.

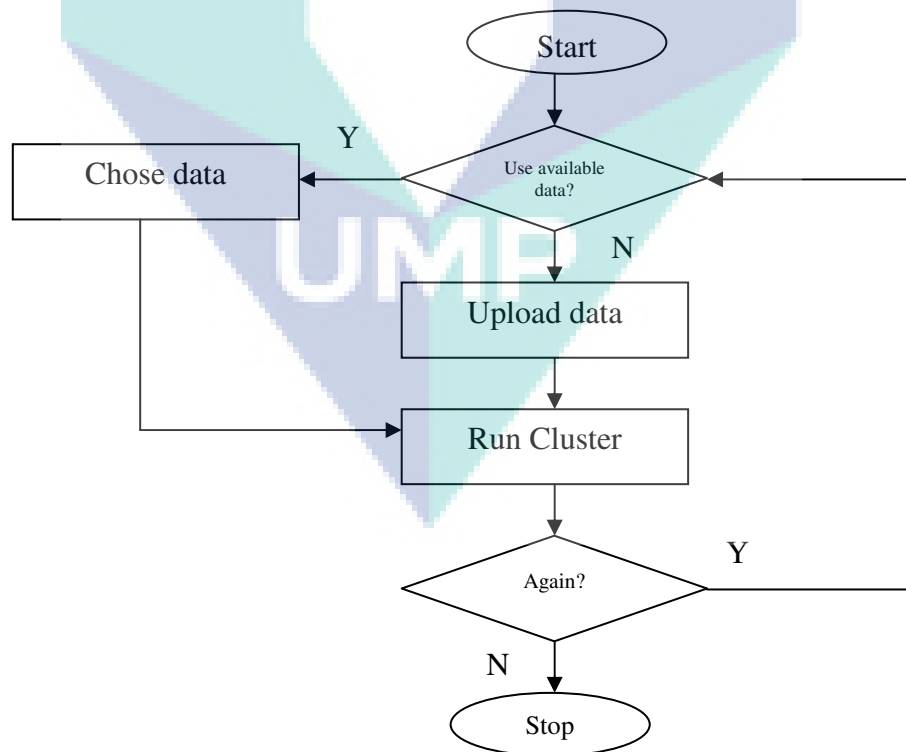


Figure 4.11. A procedure of data sets usages.

This research proposed subspace clustering based on density connection, named DATA Mining subspace clusteRing Approach (DAMIRA). Figure 4.12 show Pseudocode of DAMIRA.

```

Subclu (DB,  $\epsilon$ , Minpts)
D (D.format data in 1 dimension)
   $S_{(1)}=0$ ;
   $C_I=0$ ;
  for each data  $\in D$ 
     $C=DBSCAN(data, \epsilon, Minpts)$ 
     $S_I=S_I \cup C^S$ 
     $C_I=C_I \cup C$ 
  end for each
while (count( $C_{(k)}$ )<>0)
  Cand  $S_{k+1}$ =generate Cand  $S_{k+1}(S_k)$ 
   $S_{k+1}=0$ 
   $C_{k+1}=0$ 
  for each cand  $\in$  Cand  $S_{k+1}$ 
    best subspace=get best subspace(Cand)
    for each  $Cl \in$  best subspace
       $C=DBSCAN(Cl, \epsilon, Minpts)$ 
      if  $S \notin C$ 
         $S_{(k+1)}=S_{k+1} \cup C^S$ 
         $C_{(k+1)}=C_{k+1} \cup C$ 
      end if
    end for each
  end for each
   $k=k+1$ 
end while
candidate  $S_{k+1}(S_k)$ 
for each  $S_1 \in S_k$ 

```

```

for each  $S_2 \in S_k$ 
  cand  $S_{k+1} = S_1 \cup S_2$ 
end for each
end for each

```

Figure 4.12. Pseudocode of DATA Mining subspace clusterRing Approach (DAMIRA)

There are five steps of DAMIRA, the first step is change n-dimension to 1-dimension. After this, find out the initial cluster by using DBSCAN. This step uses the following statement:

```

For each  $Cand \in Candidate$ 
   $Bestsubspace = bestsubspace(cand)$ 
  For each  $cl \in bestsubspace$ 
     $C = DBSCAN(cl, eps, minpts)$ 
     $S(k+1) = S(k+1) \cup C^s$ 
     $C_{(k+1)} = C_{k+1} \cup C$ 

```

Figure 4.13. Change n-dimension to 1-dimension

The next step is find first cluster and first subspace. Figure 4.14 shows how it works.

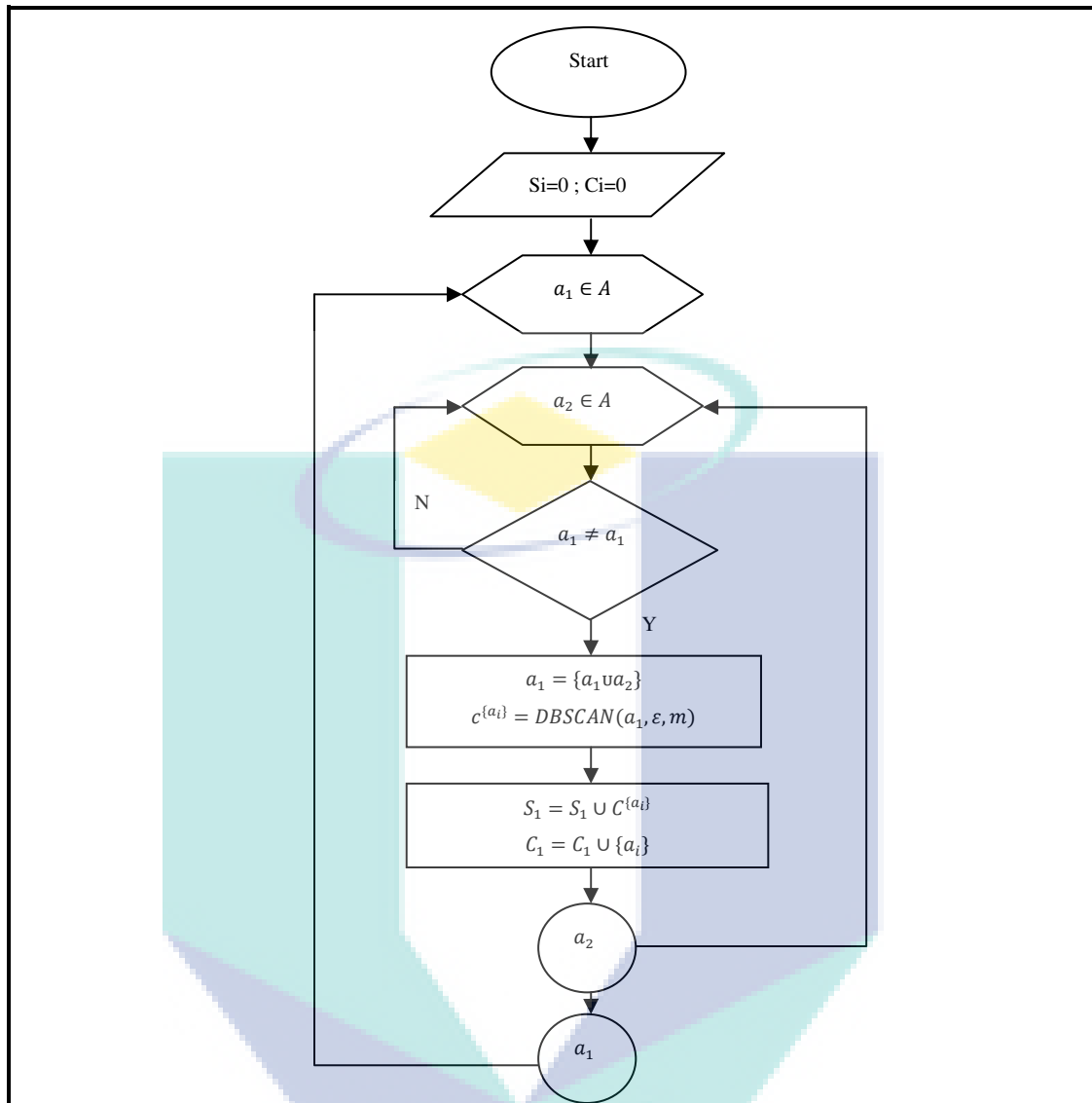


Figure 4.14. Flowchart to find first cluster and first subspace

For example, if initial data as shown in Figure 4.15, have five attributes (a), one subspace (S) and subspace T where  $T \subset S$ .

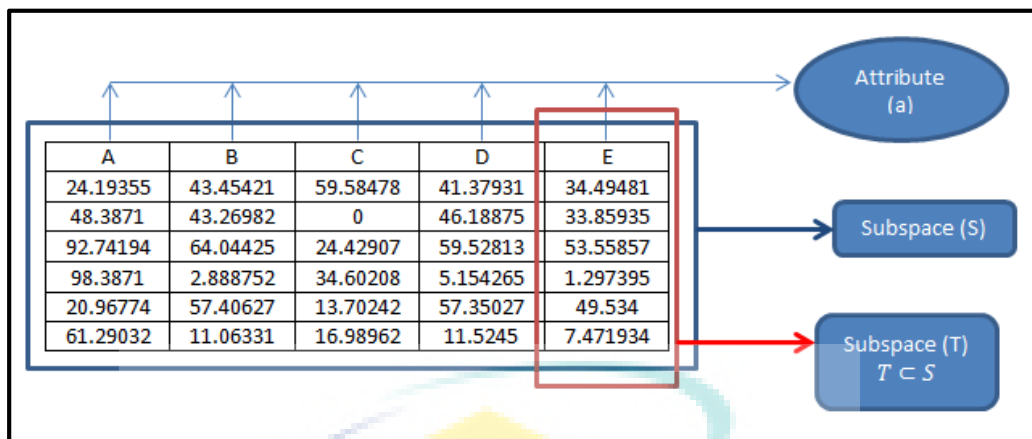


Figure 4.15. Initial data (database)

After change n-dimension to 1-dimension (Figure 4.15), will have result as shown in Figure 4.16. Each attribute will be paired with another attribute, as shown in A and B, A to C, and so on.

A	B
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

A	C
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

A	D
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

A	E
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

B	C
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

B	D
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

B	E
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

C	D
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

C	E
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

D	E
24.19355	43.45421
48.3871	43.26982
92.74194	64.04425
98.3871	2.888752
20.96774	57.40627
61.29032	11.06331

Figure 4.16. 1-dimension of cluster

Having successfully established a one-dimension on each attribute, the next step is to determine the candidate subspace over clusters, the flow as shown in Figure 4.17.



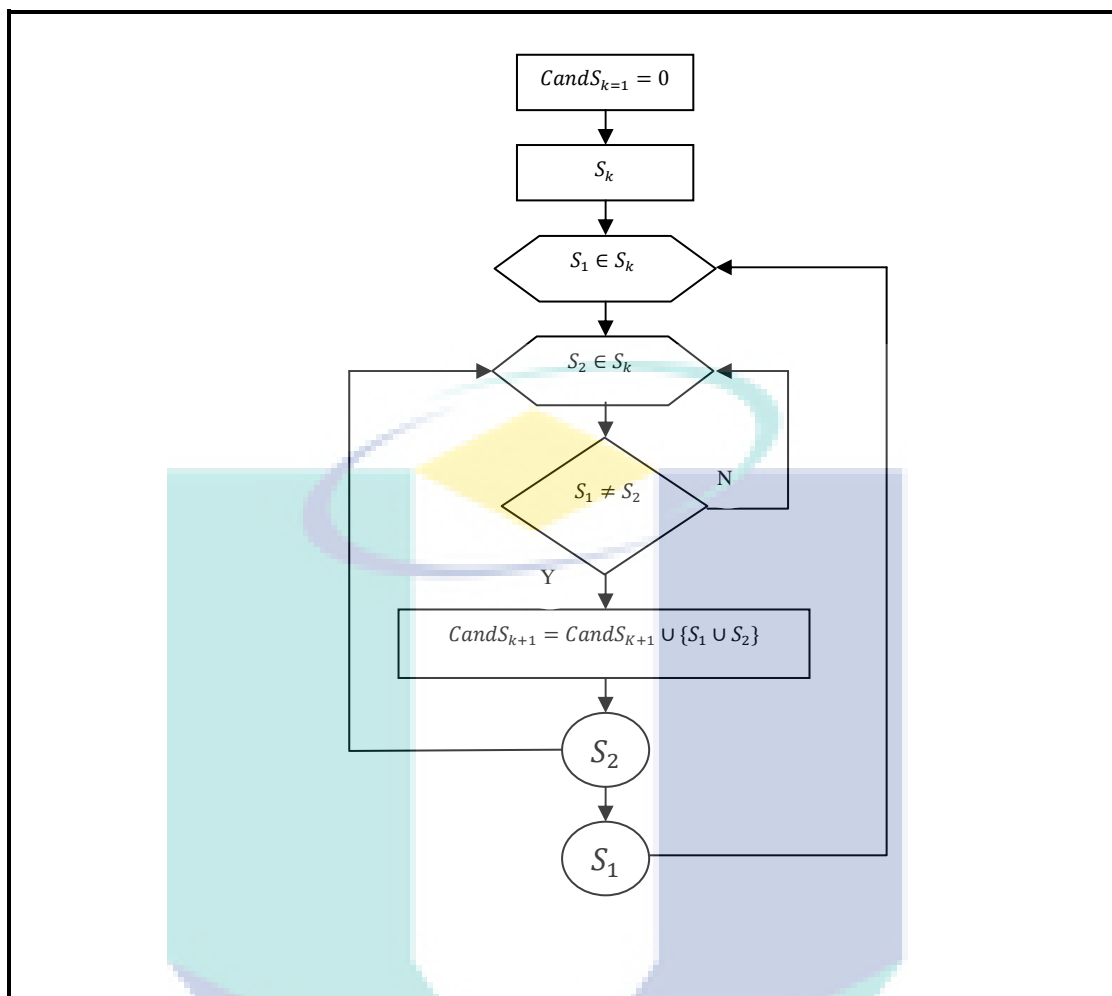


Figure 4.17. Flowchart to determine candidate subspace

From result on Figure 4.17, the next step is joint subspace  $S_1$  and  $S_2$  if the subspace 1 and 2 differ only 1 attribute. For example, if  $S_1 = \{A, C\}$  and  $S_2 = \{A, D\}$  (Figure 4.18) then  $S_1 \cup S_2 = \{A, C, A, D\}$  (Figure 4.19).

$S_1 = \{A, C\}$		$S_2 = \{A, D\}$		$S_1 \cup S_2 = \{A, C, A, D\}$			
A	C	A	D	A	C	A	D
24.19355	43.45421	24.19355	43.45421	24.19355	43.45421	24.19355	43.45421
48.3871	43.26982	48.3871	43.26982	48.3871	43.26982	48.3871	43.26982
92.74194	64.04425	92.74194	64.04425	92.74194	64.04425	92.74194	64.04425
98.3871	2.888752	98.3871	2.888752	98.3871	2.888752	98.3871	2.888752
20.96774	57.40627	20.96774	57.40627	20.96774	57.40627	20.96774	57.40627
61.29032	11.06331	61.29032	11.06331	61.29032	11.06331	61.29032	11.06331

Figure 4.18. 1-dimension of cluster      Figure 4.19. 1-dimension of cluster

Figure 4.20 show each cluster form by  $C_1 = C^{a_1} \cup \dots \cup C^{a_{10}}$ , and each subspace form by  $S_1 = \{a_1\} \cup \dots \cup \{a_{10}\}$ .

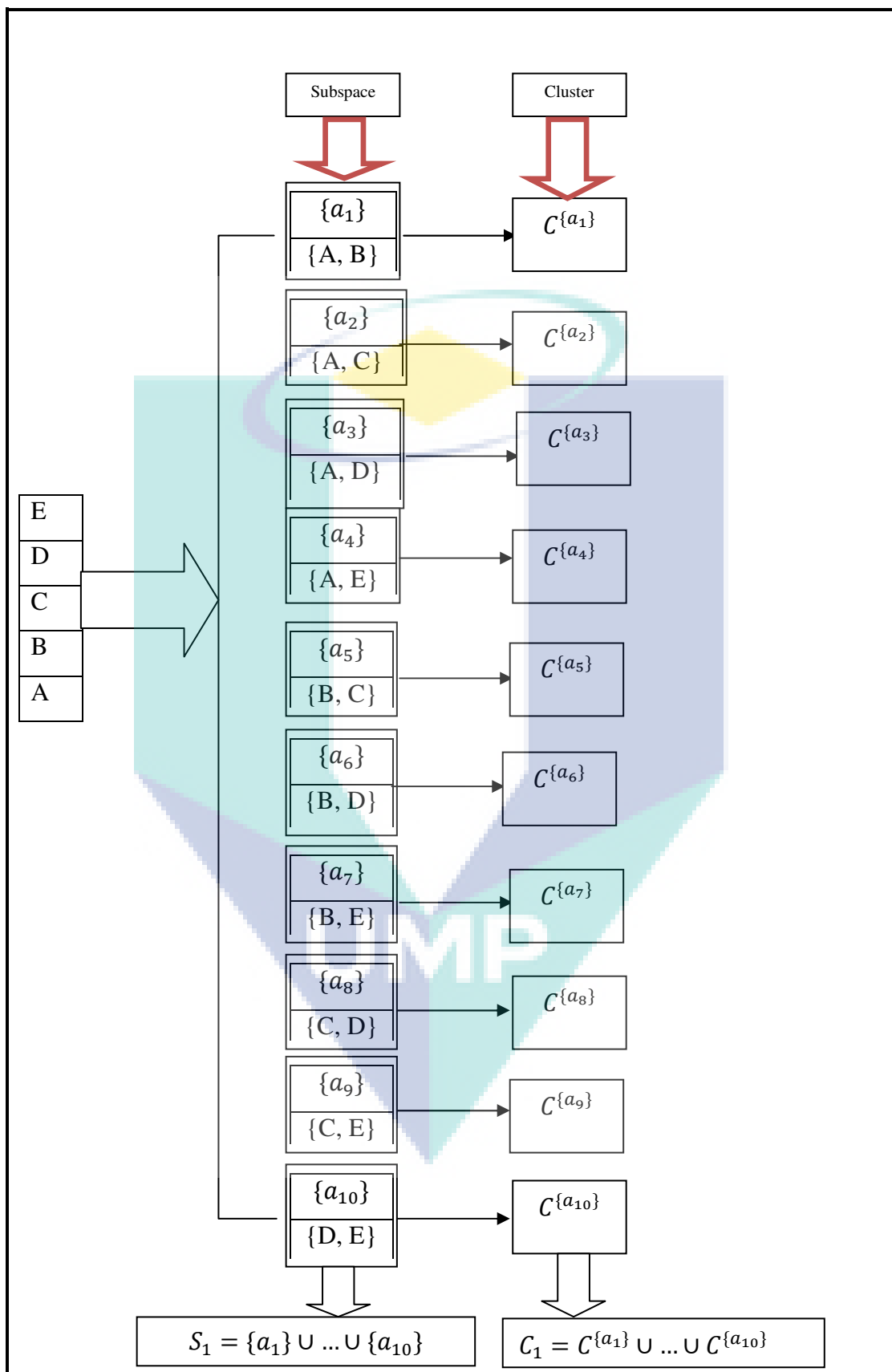


Figure 4.20. Detail of candidate subspace

After cluster and subspace determined as candidate, the next step is choose best subspace. The argument is a subspace has a lower cluster, the process shown in Figure 4.21.

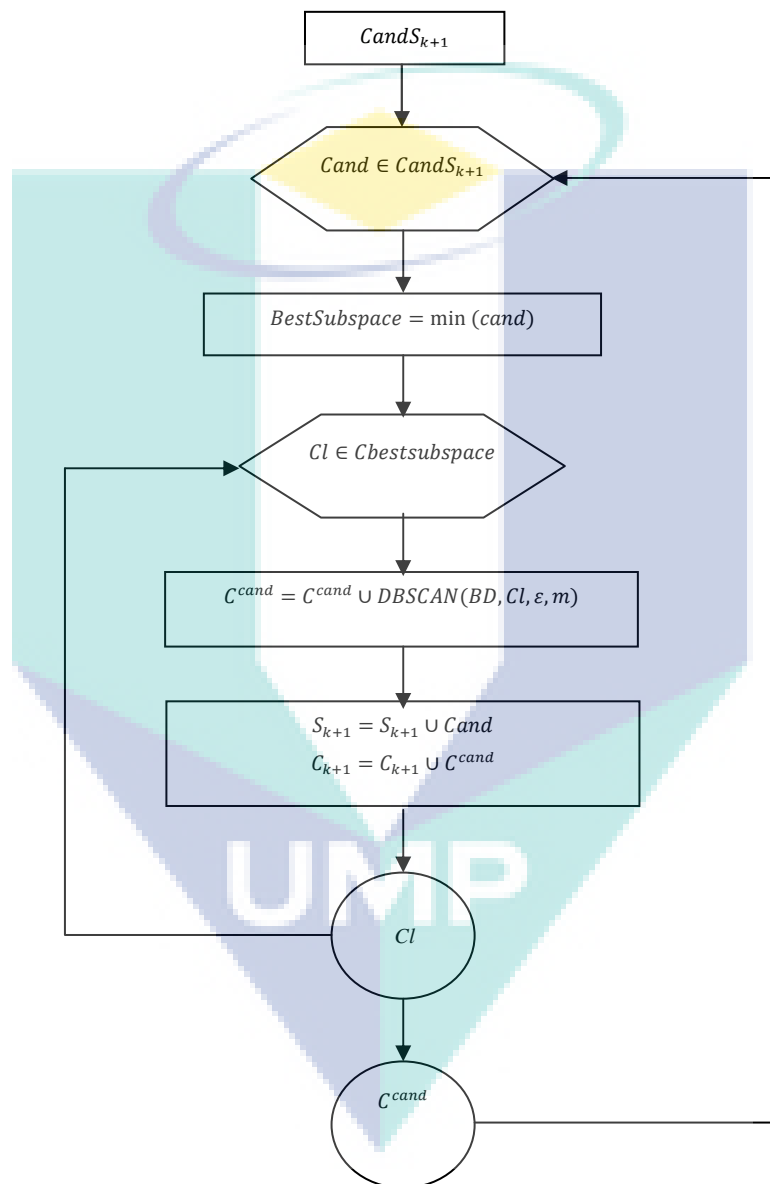


Figure 4.21. Flowchart to determine best subspace

From Figure 4.22 we can see from candidate subspace will choose best subspace best on lowest number of cluster, this statement declare in script in Figure 4.23.

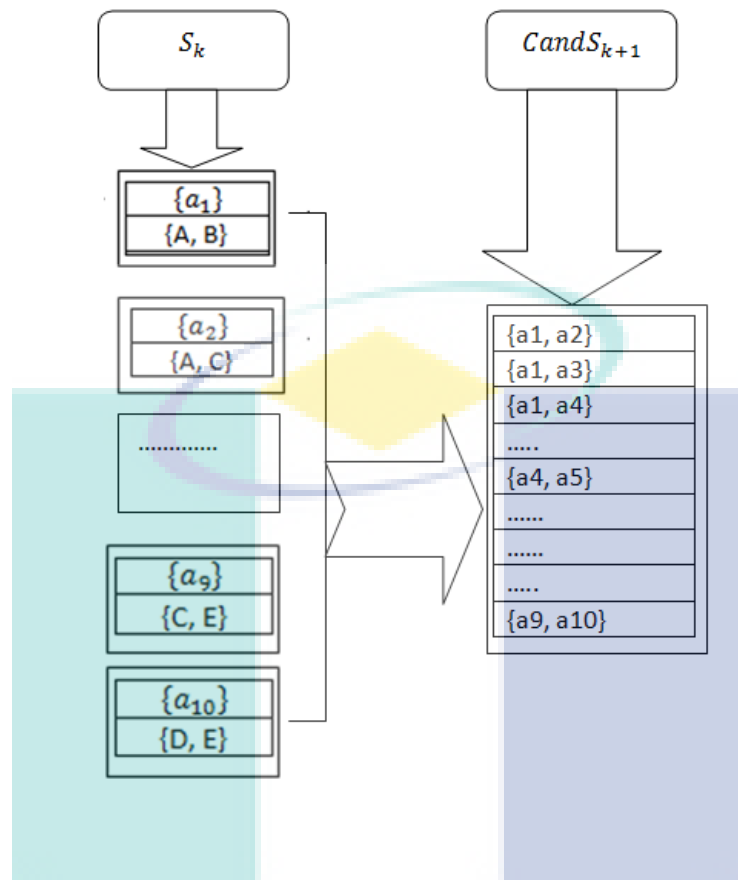


Figure 4.22. Determine best subspace

For each  $Cand \in Candidate$

$Bestsubspace = bestsubspace(cand)$

For each  $cl \in bestsubspace$

$C = DBSCAN(cl, eps, minpts)$

$S(k+1) = S(k+1) \cup C^s$

$C_{(k+1)} = C_{k+1} \cup C$

Figure 4.23. Script to determine best subspace

#### 4.4 SUMMARY

The main point of this thesis is propose an improved technique to overcome challenges clustering in educational data mining, named as DAMIRA, as the density-based clustering approach. Another objective are to develop clustering system for perception for skill required based on subspace cluster, use educational dataset higher learning institution real world datasets, as well as using of small datasets, and benchmark datasets, and develop online questionnaire for Higher Learning Institution (HLI) perception for skill required.

The main idea of is that any data has the minimum number of neighbouring data, use a bottom-up greedy algorithm to detect the density-connected clusters in all subspaces of high dimensional data. The step of DAMIRA are: the first step is change n-dimension to 1-dimension, second is find out the initial cluster by using DBSCAN, third step is found first cluster and first subspace, fourth step is to determine the candidate subspace over clusters, and lastly choose best subspace.



UMP

## CHAPTER 5

### RESEARCH FINDING AND DISCUSSION

This chapter described the research finding, presented the implementation of DAMIRA. The experimental research are analysed and comparison are done with the baseline technique, i.e., SUBCLU, FIRES, and INSCY based on three UCI benchmark datasets, and six higher learning institution datasets.

#### 5.1 INTRODUCTION

Clustering a set of data with subspace clustering presents a special challenge since two data points belong to the same cluster look as dissimilar as an arbitrary pair of data points. The main idea of subspace clustering based on the density in each cluster is that any data has the minimum number of neighbouring data, where data density must more than a certain threshold.

The proposed subspace clustering process involves five phases:

- a. Format data 1 dimension using DBSCAN

Table 5.1 show a data matrix of initial data.

Table 5.1 Example of Initial Data

data1	data2	data3
0	0	120
0	0	98
0	275	150
100	150	100
200	100	125
250	200	122

By using the following function:

```

public function formatID()
{
    $this->_headerData=array('x','y','mark','cid','rid','kom');
    $dataPoint =array();
    $dmAll = new DataOperation;
    #print_r($this->_data);
    foreach($this->_header as $k)
    {
        foreach($this->_header as $j)
        {
            if($k==$j) break;
            $dm = new DataOperation;
            $dm->attribute(array("x","y"));
            $i=0;
            foreach($this->_data->data as $e)
            {
                $data=array($e->{$j},$e->{$k},0,0,$i,($j.'-U-'. $k));
                $dm->addData(array_combine($this->_headerData, $data));
                $i++;
            }
        }
    }
}

```

```

    }
    $dmAll->addData(array($j.'_u_'. $k=>$dm));
  }
}
$this->_data1D= $dmAll;
}

```

Figure 5.1. Separate multidimensional into 1-dimension

It will produce a 1-dimensional matrix table as follows (Table 5.2):

Table 5.2 Multidimensional separate into 1-dimension

Dimension-1		Dimension-2		Dimension-3	
data1	data2	data1	data3	data2	data3
0	0	0	120	0	120
0	0	0	98	0	98
0	275	0	150	275	150
100	150	100	100	150	100
200	100	200	125	100	125
250	200	250	122	200	122

- b. Generate cluster based on density connection

Clustering using DBSCAN for each 1-Dimensional, and have resulted as follows (Table 5.3):

Table 5.3 Clustering result based on DBSCAN

0	1	0	0	75625	0	32500	0	50000	0	102500	0
0	1	0	0	75625	0	32500	0	50000	0	102500	0
75625	0	75625	0	0	3	25625	0	70625	0	68125	0
32500	0	32500	0	25625	0	0	4	12500	0	25000	0
50000	0	50000	0	70625	0	12500	0	0	5	12500	0
102500	0	102500	0	68125	0	25000	0	12500	0	0	6



c. Generate subspace cluster

Then the following functions will be the creation of subspace clustersTest candidate and generate dimensional cluster

```
$subspace[1]->addData(array('subspace'=>$d->perRegion()));
```

That will be generated subspace clustering of clusters as follows (Table 5.4):

Table 5.4 Result of Generate Subspace Cluster

S-1	s-1	C-0
		C-1
	s-2	C-0
		C-1
	s-3	C-0
		C-1

After that each cluster are grouped together formed a separate cluster, so the results are as follows (Table 5.5):

Table 5.5 Result of Group of Subspace Cluster

CandSk+1	S1 + S2
	S1 + S3
	S2 + S3

d. Test candidate and generate dimensional cluster

The next step is test and generate dimensional cluster with the following functions:

```

for($i=0;$i<count($data);$i++)
{
  $s1=$data[$i];
  #print_r($s1);
  #echo "<hr />";
  for($j=$i+1;$j<count($data);$j++)
  {
    $s2 = $data[$j];
    $candidate = array($s1->subspace,$s2->subspace);
    $candidates[] = $candidate;
    $x++;
    #print_r($candidate);
    break;
  }
}

```

Figure 5.2. Separate multidimensional into 1-dimension

- e. Search best subspace clustering.

Finally search best subspace clustering by following function:

```
$clustering[1]->addData(array('clustering'=>$d->getCluster()));
```

## 5.2 DATASET PROPERTIES

To verify the quality of the clustering obtained through our technique (DAMIRA) and to expedite the first phase, we run DBSCAN, FIRES, INSCY, SUBCLU, and DAMIRA. Setup parameter was done at subspace cluster bracketing, and average dimension and number of cluster were defined.

Table 5.6 shown the property of dataset, we implemented in 3 real dataset, and 6 higher education dataset. For each test research uses MinPoints = 6 and Epsilon = 0.9, based on previous experiment this criteria create best cluster.

Table 5.6 The Property of Dataset

Dataset	Attributes	No of data	M	E
Glass	10	214	6	0.9
Liver-dis	7	345	6	0.9
Job satisfaction	8	288	6	0.9
Ump_student_b1_b4	44	100	6	0.9
Ump_student_c1_c11	67	100	6	0.9
Ump_student_d1_d6	62	100	6	0.9
Ump_industry_b1_b3	8	71	6	0.9
Ump_industry_c1_c11	60	71	6	0.9
Ump_industry_d1_d6	61	71	6	0.9

Figure 5.3 shows the distribution of the data of glass datasets. In this data there are 345 data instance, 14 dimensions with no missing value. Data with range 46.67 and 53.33 has 68 instances.

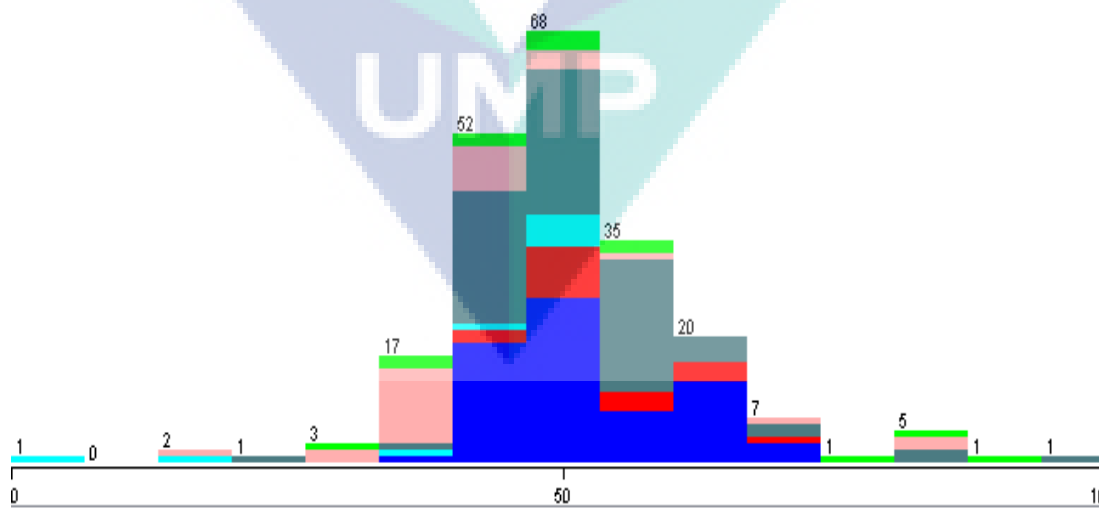


Figure 5.3. Data distribution of glass datasets

Figure 5.4 is the distribution of liver datasets. In this data there are 345 data instance, 6 dimensions with no missing value. Data with range 64.8 has 41 instances.

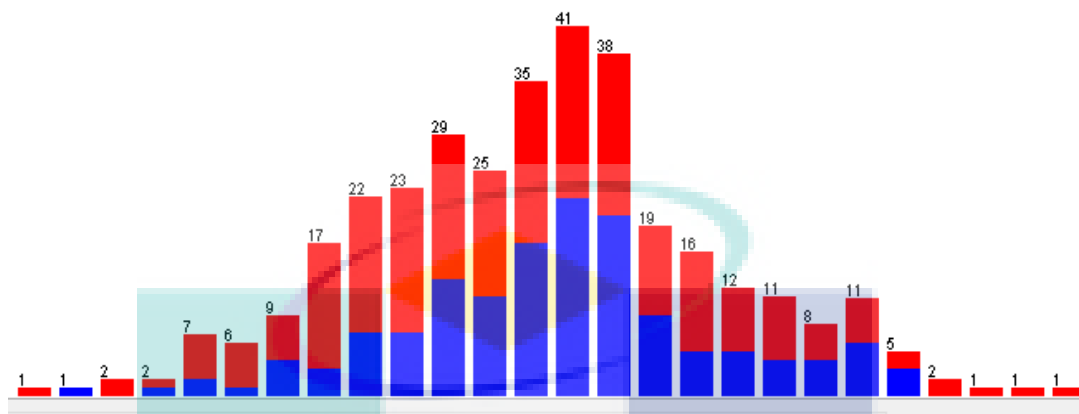


Figure 5.4. Data distribution of liver datasets

Figure 5.5 is the data distribution of job satisfaction datasets. In this data there are 345 data instance, 17 dimensions with no missing value. Data with range 64.7 and 70.5 has 76 instances.

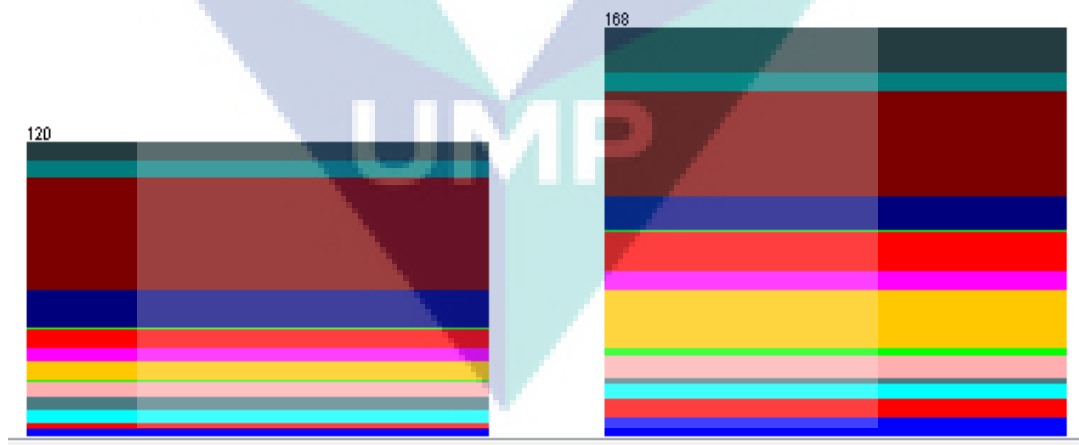


Figure 5.5. Data distribution of job satisfaction datasets

Figure 5.6 shows the distribution of the data of Ump\_student only for b1\_b4 datasets.

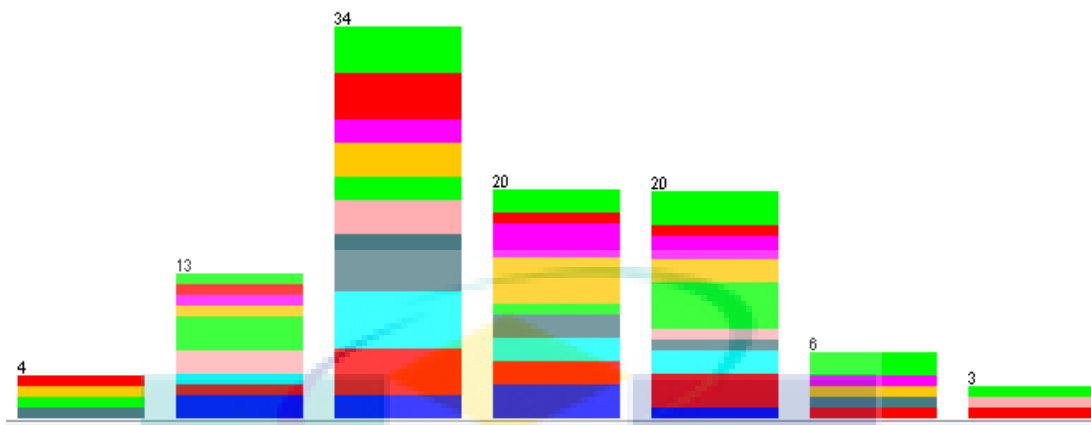


Figure 5.6. Data distribution of Ump\_student\_b1\_b4 datasets

Figure 5.7 shows the distribution of the data of Ump\_student only for c1\_c11 datasets.

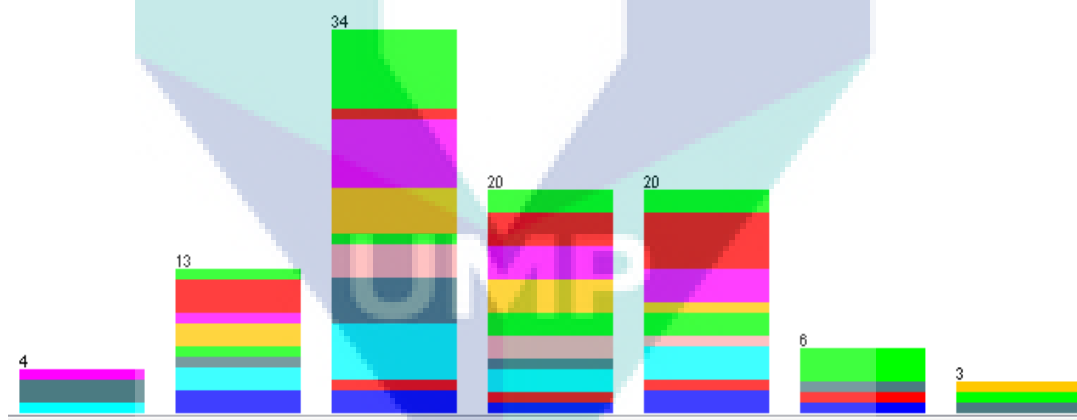


Figure 5.7. Data distribution of Ump\_student\_c1\_c11 datasets

Figure 5.8 shows the distribution of the data of Ump\_student only for d1\_d6 datasets.

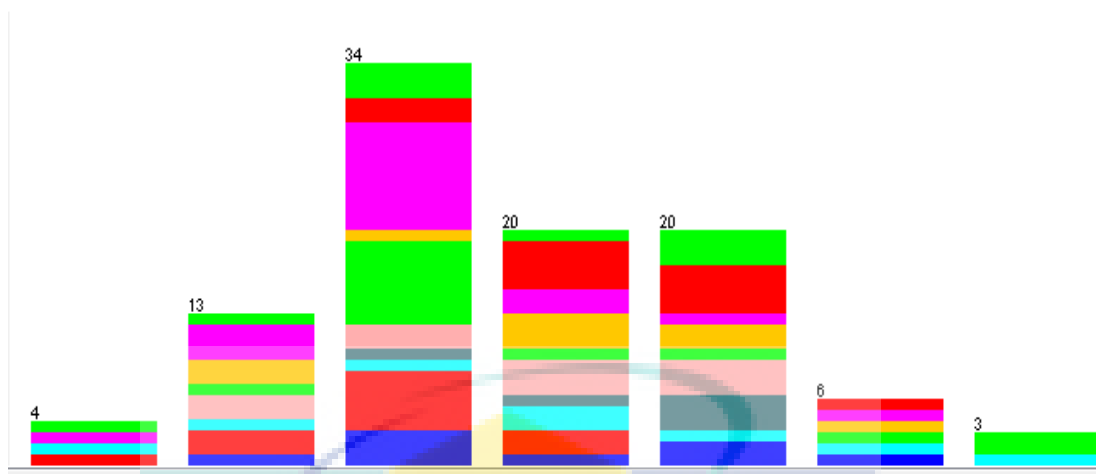


Figure 5.8. Data distribution of Ump\_student\_d1\_d6 datasets

Figure 5.9 shows the distribution of the data of Ump\_industry only for b1\_b4 datasets.

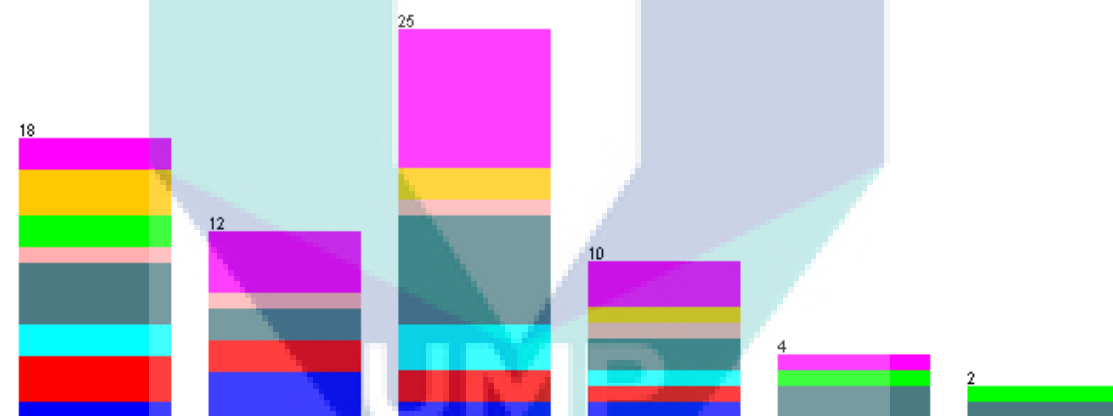


Figure 5.9. Data distribution of Ump\_industry\_b1\_b4 datasets

Figure 5.10 shows the distribution of the data of Ump\_industry only for c1\_c11 datasets.

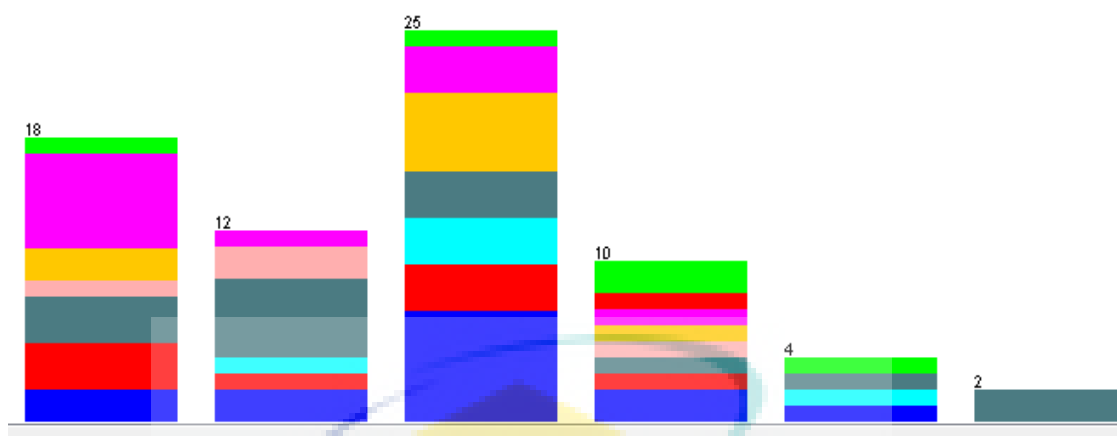


Figure 5.10. Data distribution of Ump\_industry\_c1\_c11 datasets

Figure 5.11 shows the distribution of the data of Ump\_industry only for d1\_d6 datasets.

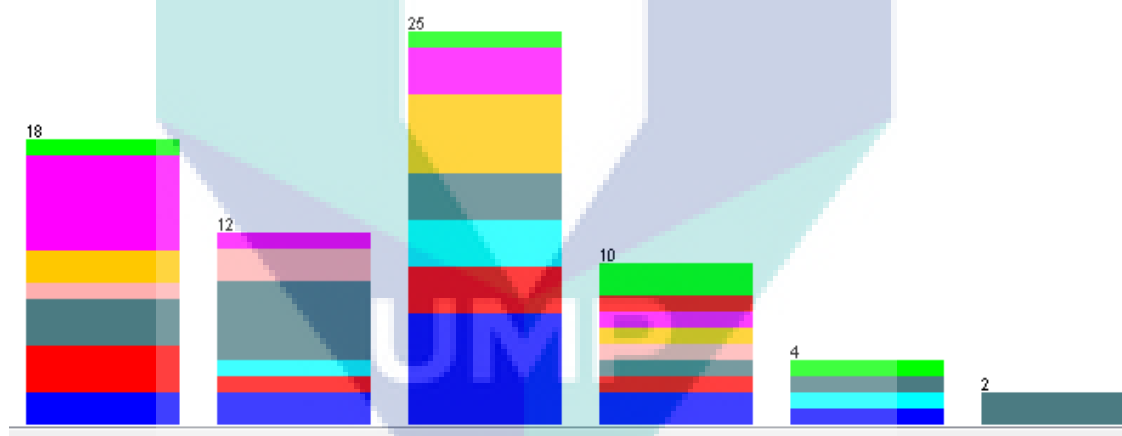


Figure 5.11. Data distribution of Ump\_industry\_d1\_d6 datasets

An important aspect of the proposed method is the number of clusters generated. Based on test results, DBSCAN clusters generated only 4 clusters obtained on job satisfaction dataset. The number of clusters in FIRES were identified very low, ranging from 5 clusters. Meanwhile, DAMIRA successfully established very large number of clusters for each dataset, as shown in Figure 5.12. With large number clusters means every data have its own subspace.

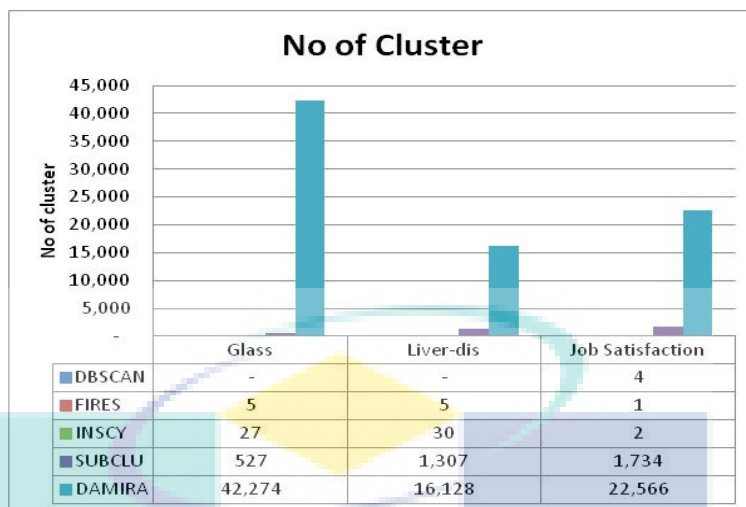
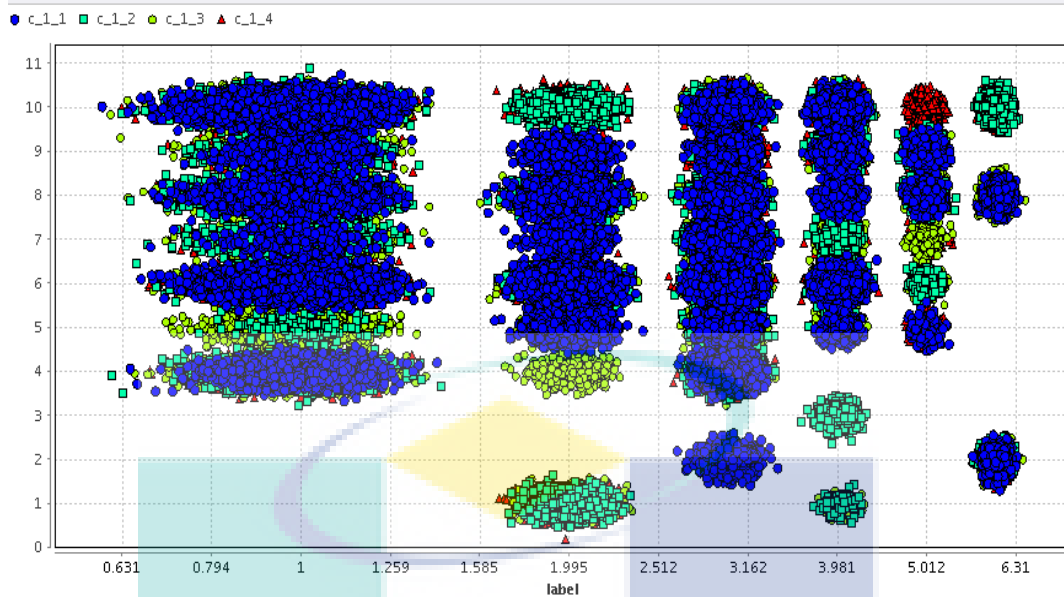


Figure 5.12. Number of cluster real datasets

### 5.3 EXPERIMENTAL RESULT

Based on SIT dataset, student perception for “Algorithm Capability” competence (Figure 5.13), cluster map of Ump\_student for c\_1 datasets shown that knowledge competence of “Prove Theoretical Results (c\_1\_1)” were needed in field of manufacturing, service and education. Knowledge competence of “Develop solutions to programming problems (c\_1\_2)” is needed in field of trade, while knowledge competence of “Determine if faster solutions possible (c\_1\_4)” is needed in another field.

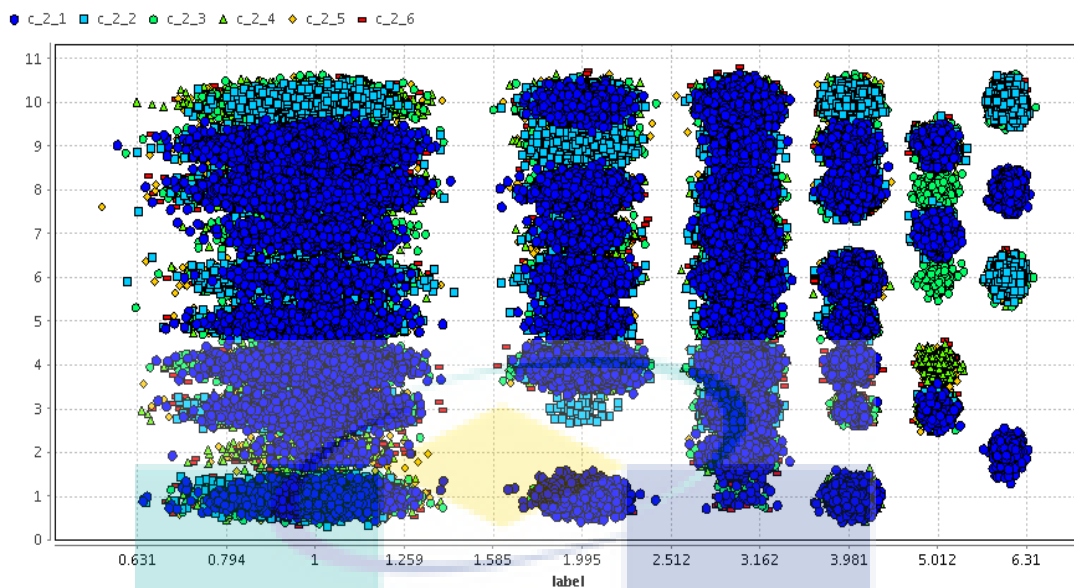




- a. Prove theoretical results (c\_1\_1)
- b. Develop solutions to programming problems (c\_1\_2)
- c. Develop proof-of-concept programs (c\_1\_3)
- d. Determine if faster solutions possible (c\_1\_4)

Figure 5.13. Algorithm Capability

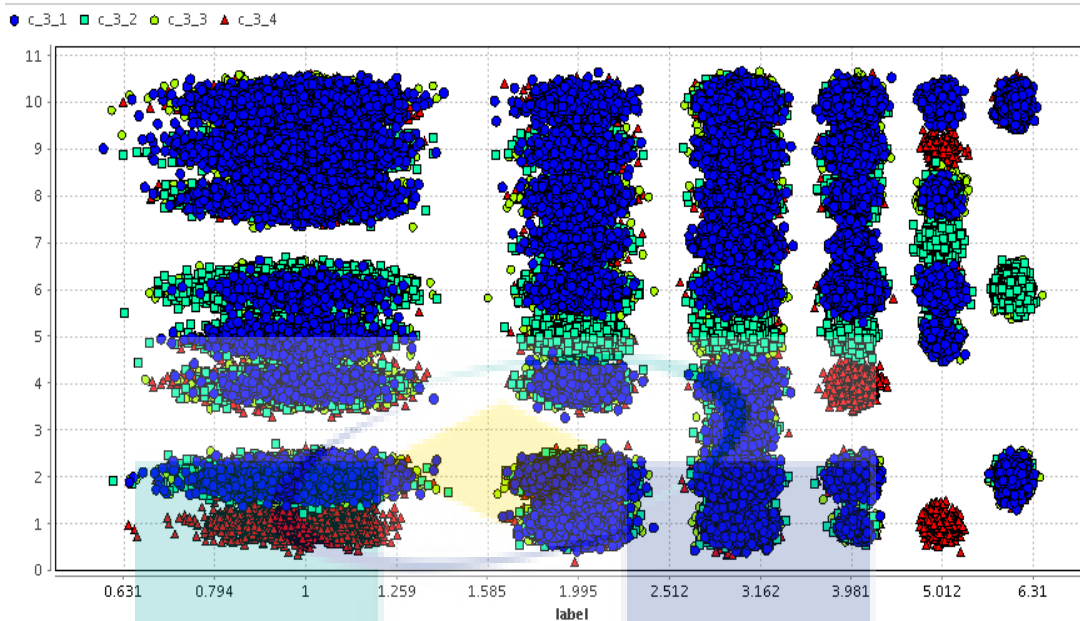
For “Application Programs” competence (Figure 5.14), cluster map of Ump\_student for c\_2 datasets shown that knowledge competence of “Use Word Processor Features (c\_2\_2)” were needed in field of manufacture, and education. Knowledge competence of “Design a word processor program (c\_2\_1)” is needed in field of trade and service.



- a. Design a word processor program (c\_2\_1)
- b. Use word processor features (c\_2\_2)
- c. Train and support word processor users (c\_2\_3)
- d. Design a spread sheet program (e.g., Excel) (c\_2\_4)
- e. Use spread sheet features (c\_2\_5)
- f. Train and support spread sheet users (c\_2\_6)

Figure 5.14. Application Programs

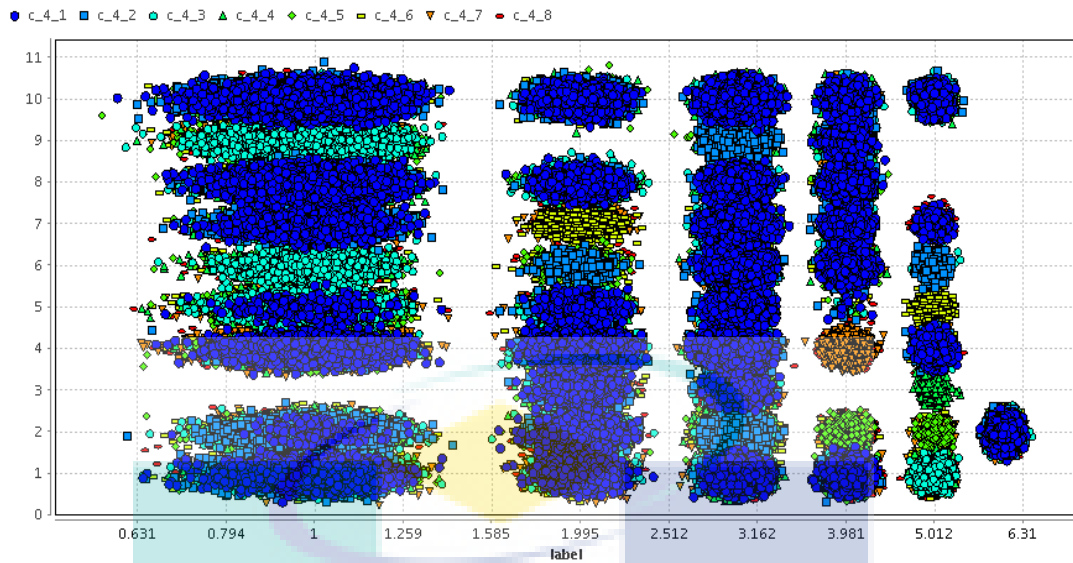
For “Computer Programming” competence (Figure 5.15), cluster map of Ump\_student for c\_3 datasets shown that knowledge competence of “Do small-scale programming (c\_3\_1)” were needed in all of field. Knowledge competence of “Develop new software systems (c\_3\_4)” is mostly needed in another field, but rather unneeded in manufacture and education.



- Do small-scale programming (c\_3\_1)
- Do large-scale programming (c\_3\_2)
- Do systems programming (c\_3\_3)
- Develop new software systems (c\_3\_4)

Figure 5.15. Computer Programming

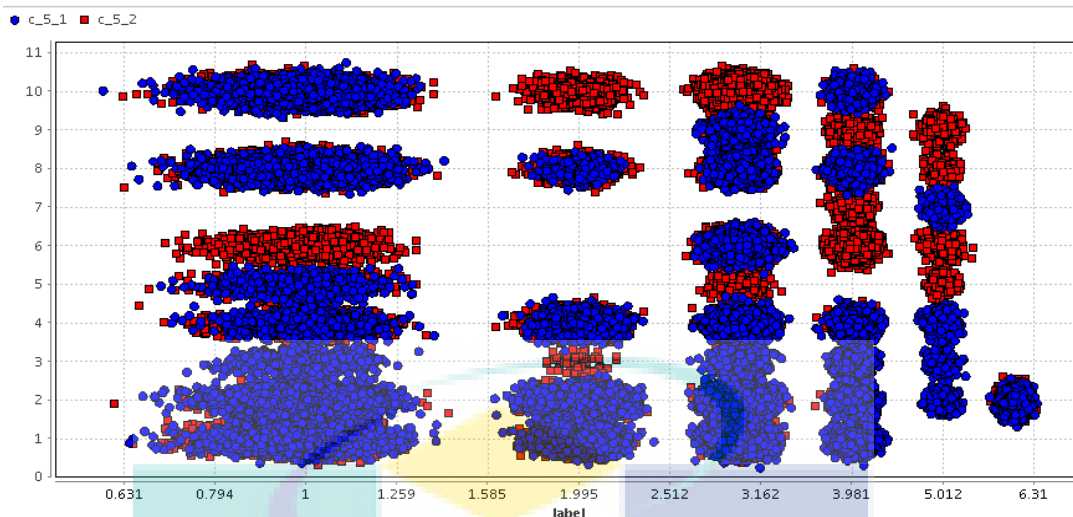
Figure 5.16 show “Hardware and Device” competence, cluster map of Ump\_student for c\_4 datasets shown that knowledge competence of “Create safety-critical systems (c\_4\_1)” were needed in all field. Knowledge competence of “Design complex sensor systems (c\_4\_6)” is needed in field of trade, but rather unneeded in service.



- a. Create safety-critical systems (c\_4\_1)
- b. Manage safety-critical projects (c\_4\_2)
- c. Design embedded systems (c\_4\_3)
- d. Implement embedded systems (c\_4\_4)
- e. Design computer peripherals (c\_4\_5)
- f. Design complex sensor systems (c\_4\_6)
- g. Design a chip (c\_4\_7)
- h. Program a chip (c\_4\_8)

Figure 5.16. Hardware and Device

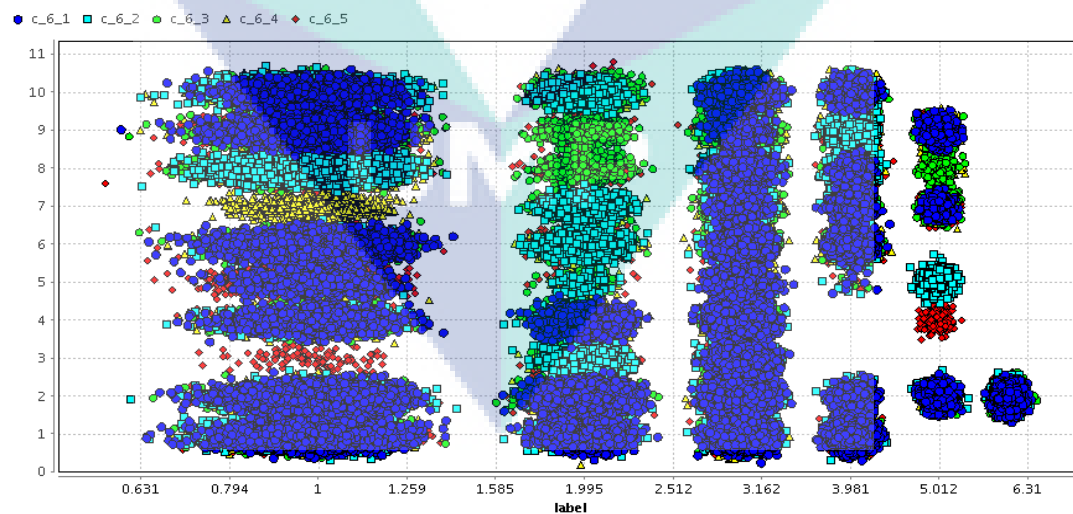
For “Human Computer Interaction” competence (Figure 5.17), cluster map of Ump\_student for c\_5 datasets shown that knowledge competence of “Design a computer (c\_5\_1)” were needed in field of manufacture. Knowledge competence of “Create a software user interface (c\_5\_2)” is needed in field of trade and service.



- a. Design a computer (c\_5\_1)
- b. Create a software user interface (c\_5\_2)

Figure 5.17. Human Computer Interaction

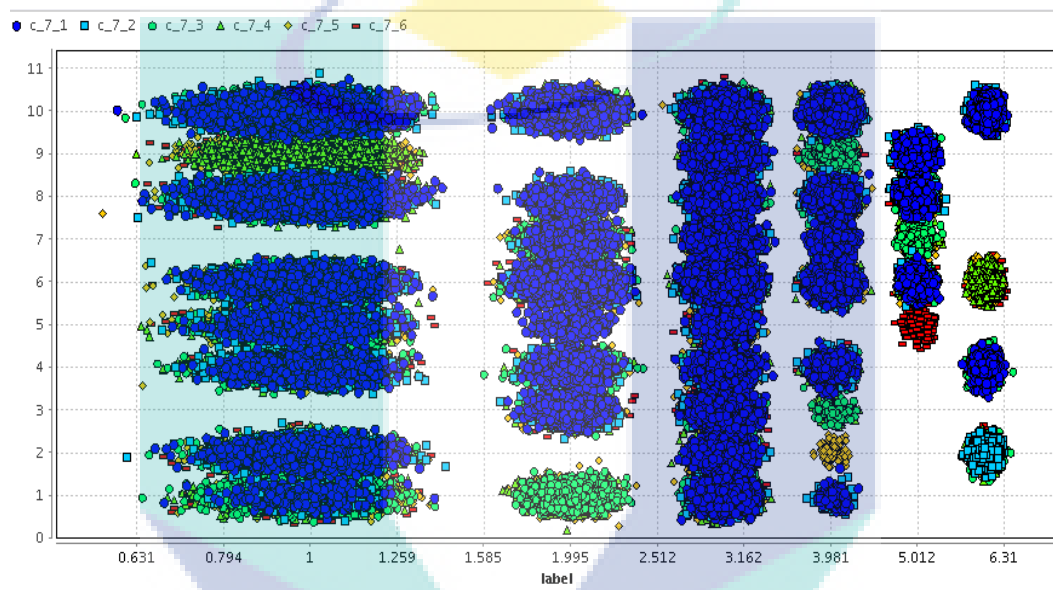
Figure 5.18 show “Information System” competence, cluster map of Ump\_student for c\_6 datasets shown that knowledge competence of “Produce graphics or game software (c\_6\_1)” were needed in field of manufacture, and service. Knowledge competence of “Design a human-friendly device (c\_6\_2)” is needed in field of trade and education.



- a. Produce graphics or game software (c\_6\_1)
- b. Design a human-friendly device (c\_6\_2)
- c. Define information system requirements (c\_6\_3)
- d. Design information systems (c\_6\_4)
- e. Implement information systems (c\_6\_5)

Figure 5.18. Information System

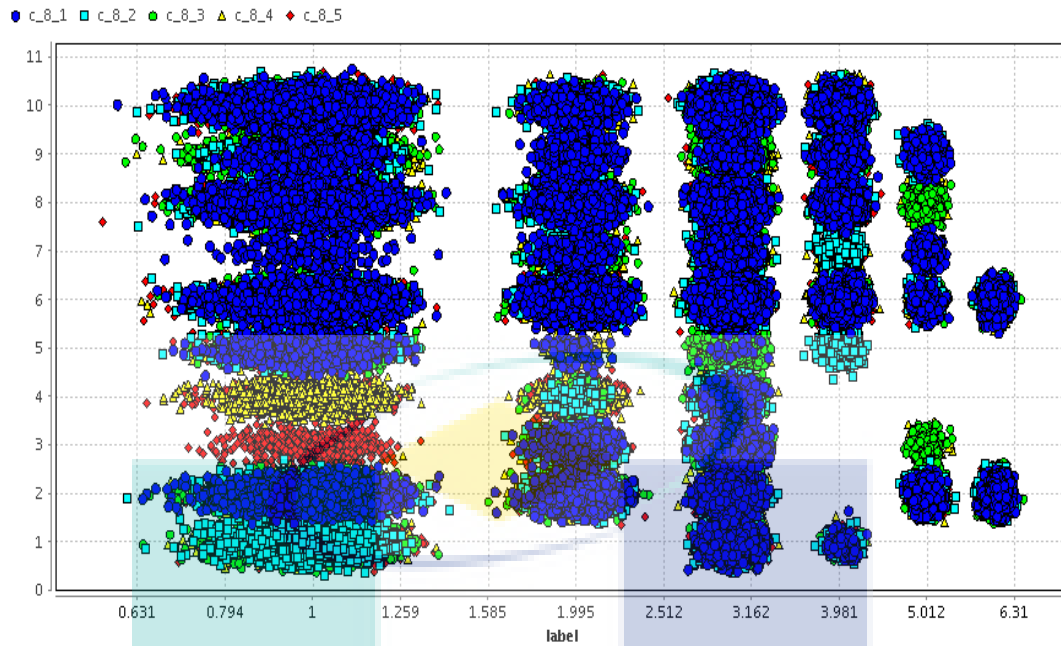
Figure 5.19 show “Information Management (Database)” competence, cluster map of Ump\_student for c\_7 datasets shown that knowledge competence of “Train users to use information systems (c\_7\_1)” were needed in all field of industry. Knowledge competence of “Maintain and modify information systems (c\_7\_2)” is rather needed in field of manufacture than others fields.



- Train users to use information systems (c\_7\_1)
- Maintain and modify information systems (c\_7\_2)
- Design a database management system (e.g., Oracle) (c\_7\_3)
- Model and design a database (c\_7\_4)
- Implement information retrieval software (c\_7\_5)
- Select database products (c\_7-6)

Figure 5.19. Information Management (Database)

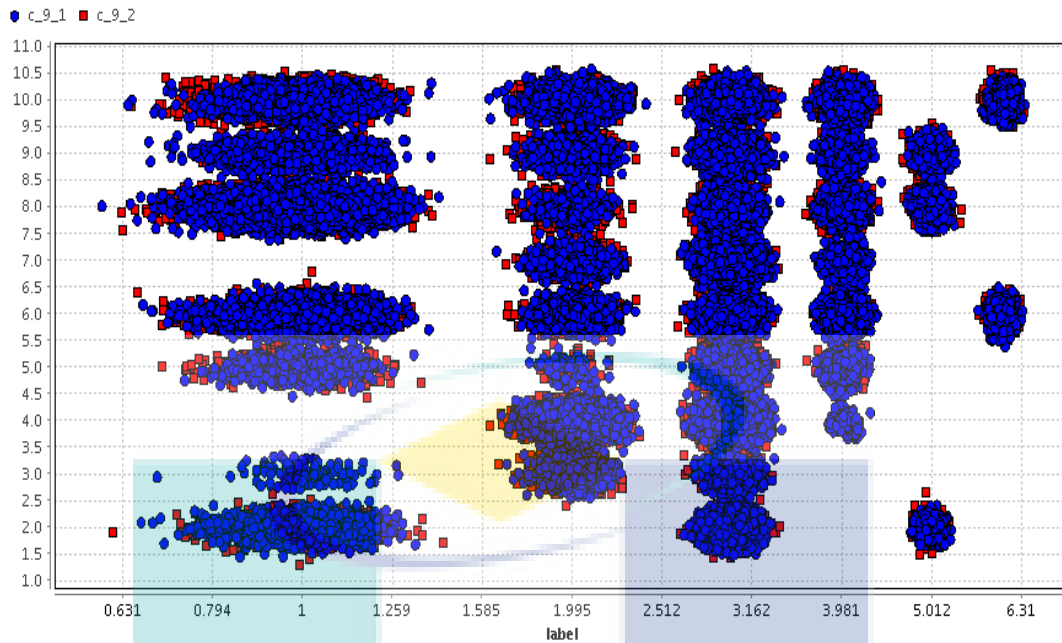
Figure 5.20 show “IT Resource Planning” competence, cluster map of Ump\_student for c\_8 datasets shown that knowledge competence of “Manage databases (c\_8\_1)” were needed in all of field. Knowledge competence of “Develop corporate information plan (c\_8\_3)” is needed in field of manufacture and service.



- a. Manage databases (c\_8\_1)
- b. Train and support database users (c\_8\_2)
- c. Develop corporate information plan (c\_8\_3)
- d. Develop computer resource plan (c\_8\_4)
- e. Schedule/budget resource upgrades (c\_8\_5)

Figure 5.20. IT Resource Planning

Figure 5.21 show “Intelligent System” competence cluster map of Ump\_student for c\_9 datasets shown that knowledge competence of “Install/upgrade computers (c\_9\_1)” and Install/upgrade computer software (c\_9\_2)” were needed in all fields.

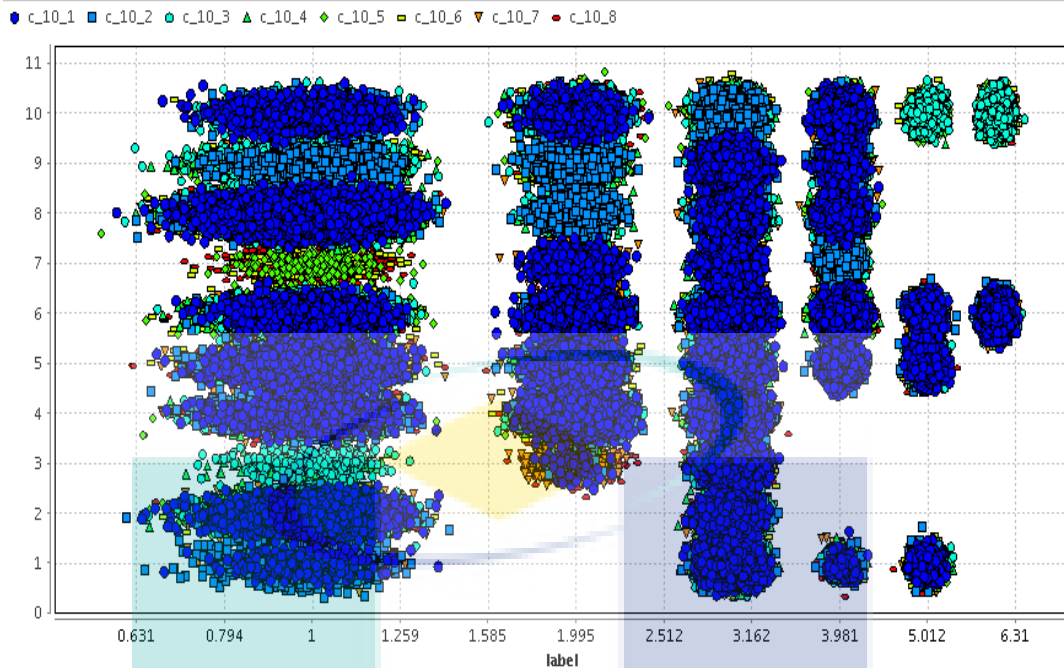


- a. Install/upgrade computers (c\_9\_1)
- b. Install/upgrade computer software (c\_9\_2)

Figure 5.21. Intelligent System

Figure 5.22 show “Networking and Communication” competence, cluster map of Ump\_student for c\_10 datasets shown that knowledge competence of “Design auto-reasoning systems (c\_10\_1)” were needed in field of manufacture, trade and education. Knowledge competence of “Design network configuration (c\_10\_3)” is mostly needed in all fields.

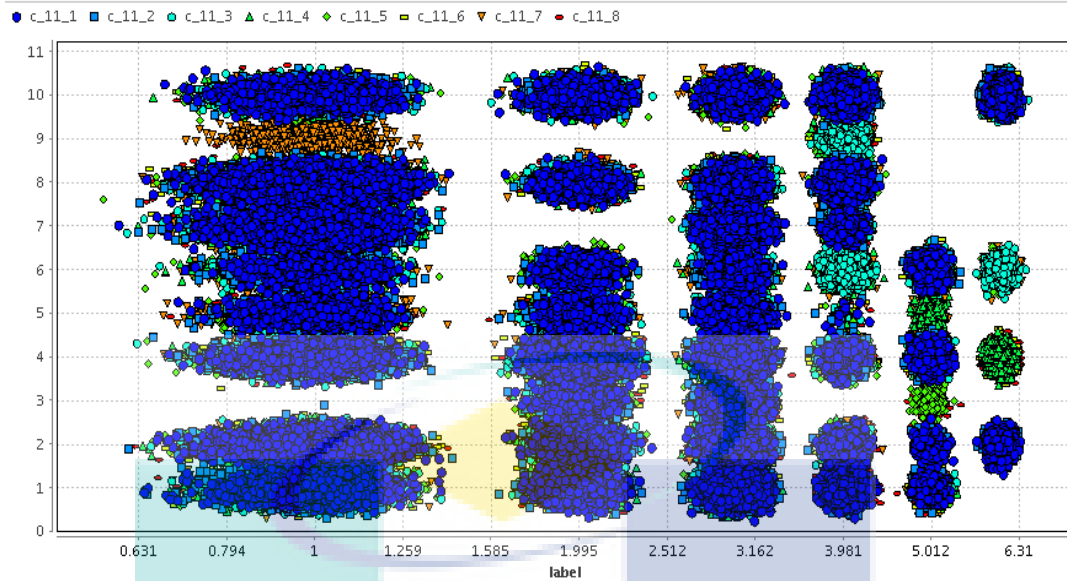




- a. Design auto-reasoning systems (c\_10\_1)
- b. Implement intelligent systems (c\_10\_2)
- c. Design network configuration (c\_10\_3)
- d. Select network components (c\_10\_4)
- e. Install computer network (c\_10\_5)
- f. Manage computer networks (c\_10\_6)
- g. Implement communication software (c\_10\_7)
- h. Manage communication resources (c\_10\_8)

Figure 5.22. Networking and Communication

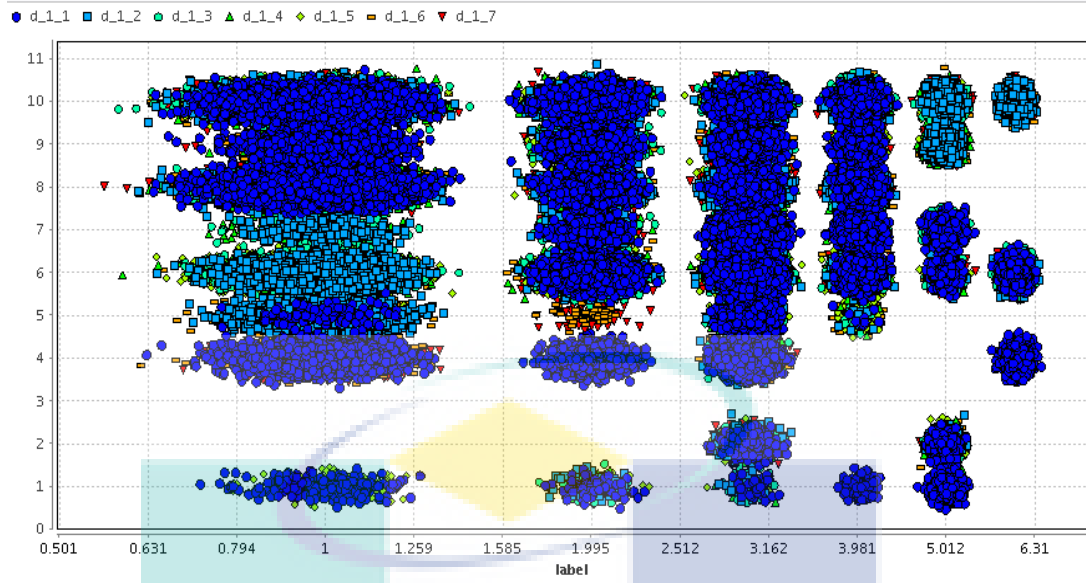
Figure 5.23 show “System Development through Integration” competence, cluster map of Ump\_student for c\_11 datasets shown that knowledge competence of “Implement mobile computing system (c\_11\_1)” were needed in all field. Knowledge competence of “Develop business solutions (c\_11\_7)” is needed in field of manufacture, while competence of “Manage an organization’s web presence (c\_11\_3) is mostly needed in education field.



- Implement mobile computing system (c\_11\_1)
- Manage mobile computing resources (c\_11\_2)
- Manage an organization's web presence (c\_11\_3)
- Configure & integrate e-commerce software (c\_11\_4)
- Develop multimedia solutions (c\_11\_5)
- Configure & integrate e-learning systems (c\_11\_6)
- Develop business solutions (c\_11\_7)
- Evaluate new forms of search engine (c\_11\_8)

Figure 5.23. System Development through Integration

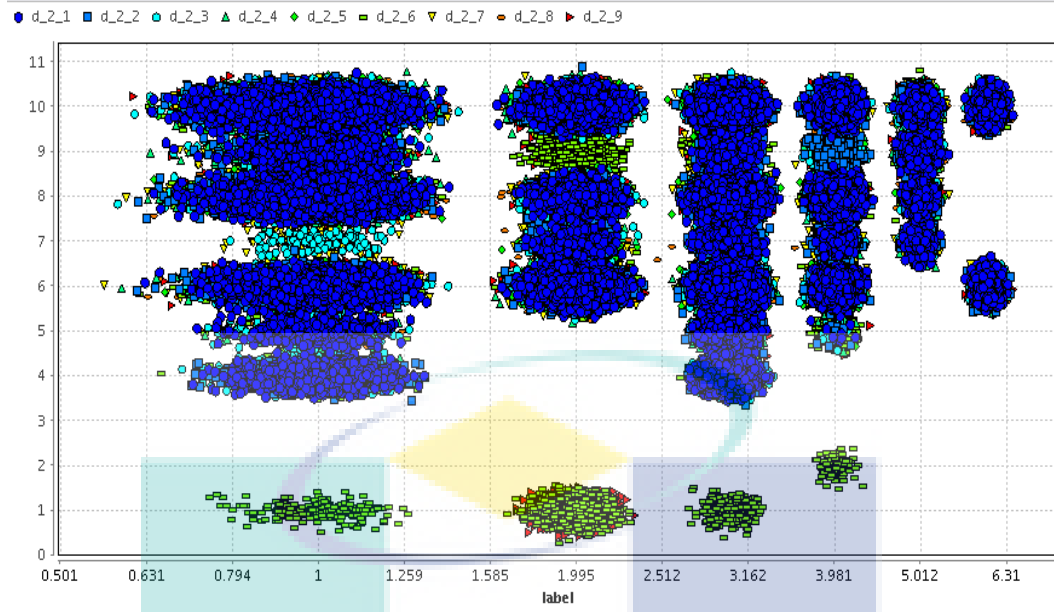
Figure 5.24 show "Resource Management" competence, cluster map of Ump\_student for d\_1 datasets shown that soft skill competence of "Budget management (d\_1\_1)" were needed in field of manufacture, trade, service, and education. Soft skill competence of "Coping with deadlines (d\_1\_2)" and "Establishing objectives (d\_1\_3)" is needed in field of education.



- a. Budget management (d\_1\_1)
- b. Coping with deadlines (d\_1\_2)
- c. Establishing objectives (d\_1\_3)
- d. Scheduling (d\_1\_4)
- e. Forecasting (d\_1\_5)
- f. Personal organization (d\_1\_6)
- g. Time management (d\_1\_7)

Figure 5.24. Resource Management

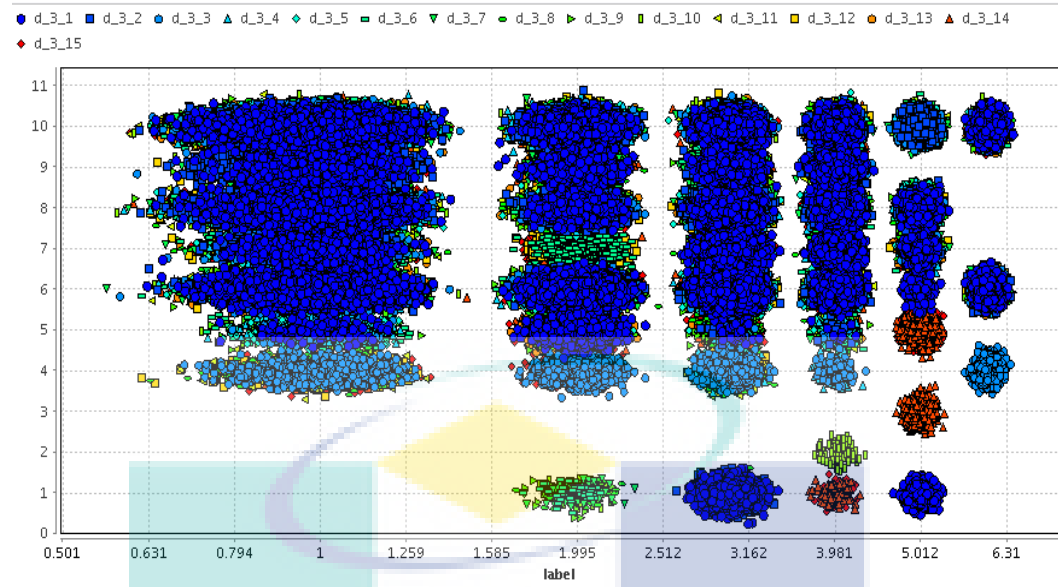
Figure 5.25 show “Communication and Interpersonal” competence, cluster map of Ump\_student for d\_2 datasets shown that soft skill competences of “Negotiation (d\_2\_1)” were needed in all. Soft skill competence of “Listening (d\_2\_7)” is mostly needed in field of trade.



- a. Negotiation (d\_2\_1)
- b. Customer relations (d\_2\_2)
- c. Recognizing value of diversity (d\_2\_3)
- d. Seeking and receiving feedback (d\_2\_4)
- e. Teamwork/collaboration (d\_2\_5)
- f. Selecting people/interviewing (d\_2\_6)
- g. Listening (d\_2\_7)
- h. Establishing work relationships (d\_2\_8)
- i. Speaking/presentations (d\_2\_9)

Figure 5.25. Communication and Interpersonal

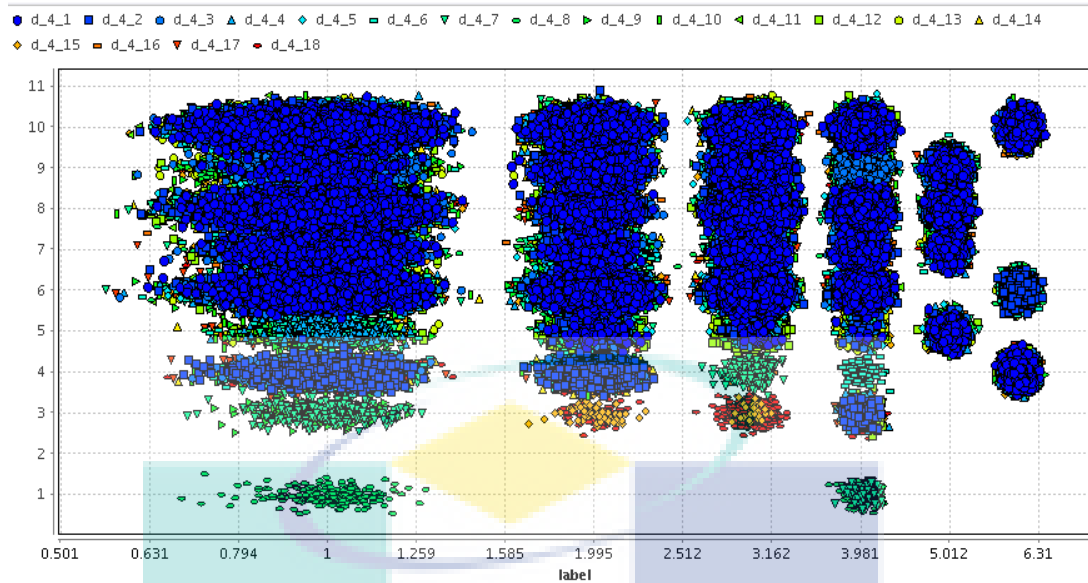
Figure 5.26 show “Leadership” competence, cluster map of Ump\_student for d\_3 datasets shown that soft skill competence of “Anticipating problems and taking action without waiting to be told (d\_3\_1)” were needed in all field.



- a. Anticipating problems and taking action without waiting to be told (d\_3\_1)
- b. Dealing with pressure (d\_3\_2)
- c. Delegating (d\_3\_3)
- d. Motivating others (d\_3\_4)
- e. Responsibly challenging the status quo (d\_3\_5)
- f. Championing change/new ideas/innovation (d\_3\_6)
- g. Providing feedback/initiating difficult conversations (d\_3\_7)
- h. Following through/ accountability (d\_3\_8)
- i. Holding others accountable (d\_3\_9)
- j. Initiating change/improvement (d\_3\_10)
- k. Developing self/self-directed learning (d\_3\_11)
- l. Persistence (d\_3\_12)
- m. Influencing and persuading (d\_3\_13)
- n. Supervising/coordinating the work of others (d\_3\_14)
- o. Developing people/mentoring/coaching (d\_3\_15)

Figure 5.26. Leadership

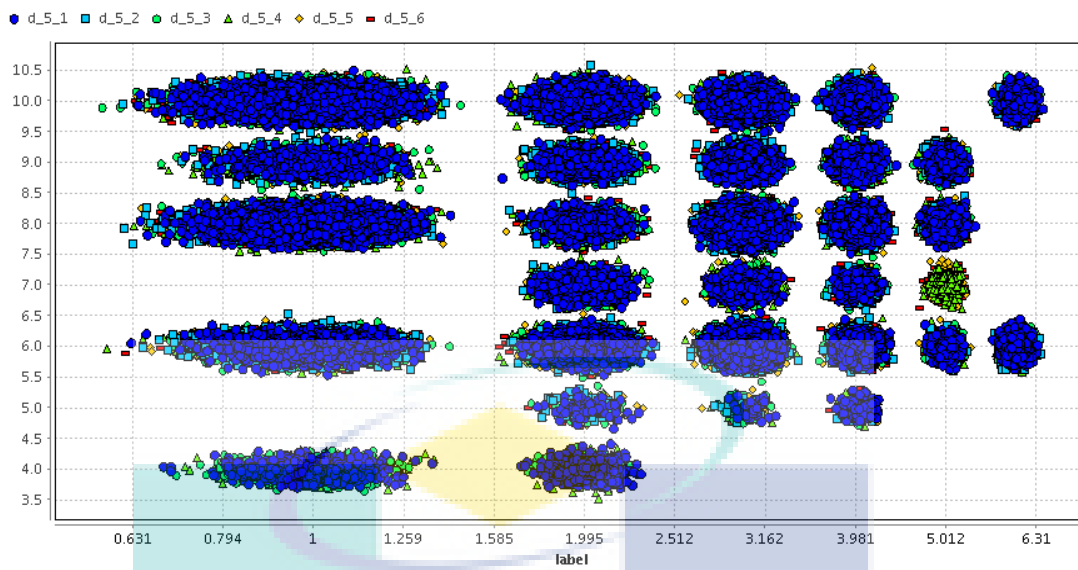
Figure 5.27 show “Information Management” competence, cluster map of Ump\_student for d\_4 datasets shown that soft skill competence of “Analyzing/classifying (d\_4\_1)” were needed in all field. Soft skill competence of “Abstracting (d\_4\_15)” and “Modelling (d\_4\_18)” and is unneeded in field of trade and service.



- a. Analyzing/classifying (d\_4\_1)
- b. Interpreting/translating (d\_4\_2)
- c. Observing (d\_4\_3)
- d. Integrating (d\_4\_4)
- e. Reporting (d\_4\_5)
- f. Conceptualizing (d\_4\_6)
- g. Calculating (d\_4\_7)
- h. Designing (d\_4\_8)
- i. Editing/revising (d\_4\_9)
- j. Investigating (d\_4\_10)
- k. Decision-making (d\_4\_11)
- l. Synthesizing (d\_4\_12)
- m. Writing (d\_4\_13)
- n. Reading (d\_4\_14)
- o. Abstracting (d\_4\_15)
- p. Dealing with ambiguity/uncertainty (d\_4\_16)
- q. Constructing (d\_4\_17)
- r. Modelling (d\_4\_18)

Figure 5.27. Information Management

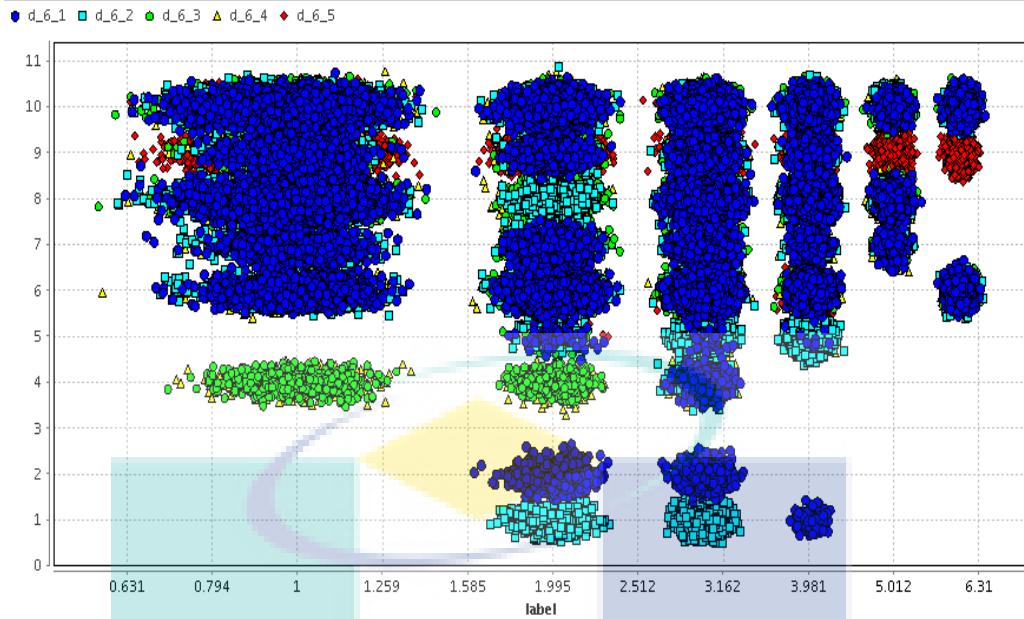
Figure 5.28 show “Systems Thinking” competence, cluster map of Ump\_student for d\_5 datasets shown that soft skill competence of “Thinking strategically (d\_2\_2)” were needed in all field.



- a. Thinking strategically (d\_5\_1)
- b. Thinking systematically (d\_5\_2)
- c. Establishing performance standards (d\_5\_3)
- d. Evaluating performance (d\_5\_4)
- e. Measuring performance (d\_5\_5)
- f. Correcting performance (d\_5\_6)

Figure 5.28. Systems Thinking

Figure 5.29 show “Technical/Functional Competence” competence, cluster map of Ump\_student for d\_6 datasets shown that soft skill competence of “Troubleshooting/maintaining technology (d\_6\_1)” were needed in all field.



- a. Troubleshooting/maintaining technology (d\_6\_1)
- b. Using instruments/equipment (d\_6\_2)
- c. Problem solving (d\_6\_3)
- d. Selecting applications (d\_6\_4)
- e. Another Technical/Functional skills (d\_6\_5)

Figure 5.29. Technical/Functional Competence

Similarly, the number of clusters of higher learning institution dataset, formed another cluster subspace methods, such as FIRES and INSCY, tendency to fail to form clusters in each subspace, as shown in Figure 5.30, while DAMIRA, produced many clusters, beyond than the other methods.



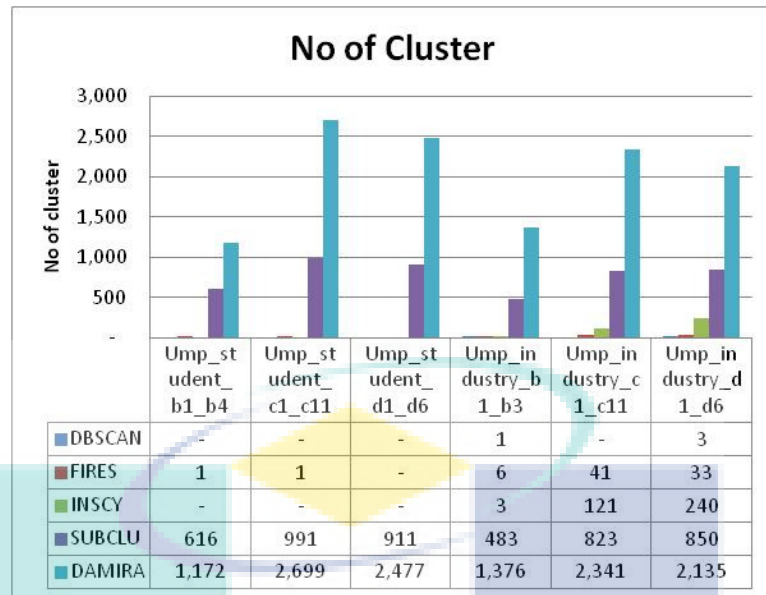


Figure 5.30. Number of cluster Higher Learning Institution datasets

Another concern of clustering is how many data is missing or un-clustered, more un-clustered data could bias information result. As shown in Figure 5.31, SUBCLU and DAMIRA has no un-clustered real datasets, thus the perception of the results of the cluster will produce more accurate information.

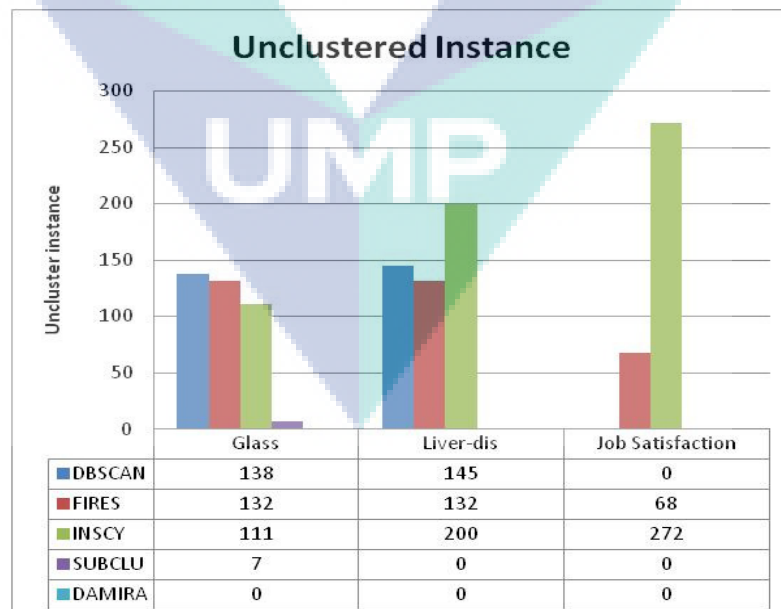


Figure 5.31. Un-cluster data of real datasets

Similarly, the process of data clustering higher learning institution, by FIRES and INSCY not all data can be in the cluster. As shown in Figure 5.32, the INSCY method most likely to produce un-clustered instance, while the SUBCLU and DAMIRA managed all data in the cluster.

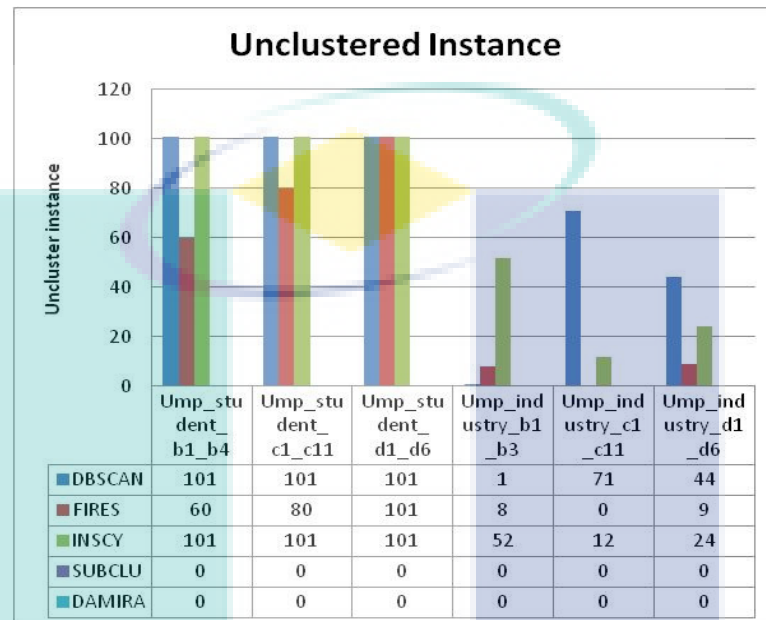


Figure 5.32. Un-cluster data of higher learning institution datasets

## 5.4 PERFORMANCE EVALUATION

Performance evaluation of data mining became very important, the prediction of correct number of clusters unsupervised learning process is a hurdle, nevertheless can be cleared by using efficiency, accuracy, cluster coverage and F1-Entropy indices to assess the quality of the clusters.

### 5.4.1 Efficiency

From the experimental results, we can see that the clustering time for glass dataset and liver dataset, DAMIRA method is more than 20 times longer than the

FIRES, INSCY and SUBCLU, meanwhile for job satisfaction dataset DAMIRA need shortest time than SUBCLU and INSCY methods as shown in Figure 5.33.

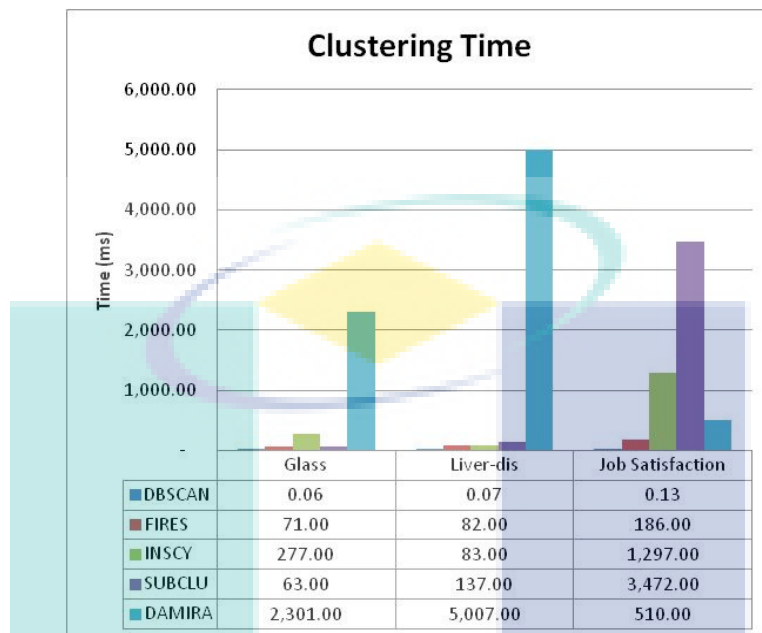


Figure 5.33. Time processing of clustering of real datasets

For larger and more complex the data, the performance DAMIRA looks more efficient than SUBCLU as shown in Figure 5.34. However, still lower than the FIRES and INSCY, especially compared with DBSCAN method, which can be done very quickly, averaging less than 1 second, but tend to be a lot of data as un-clustered.

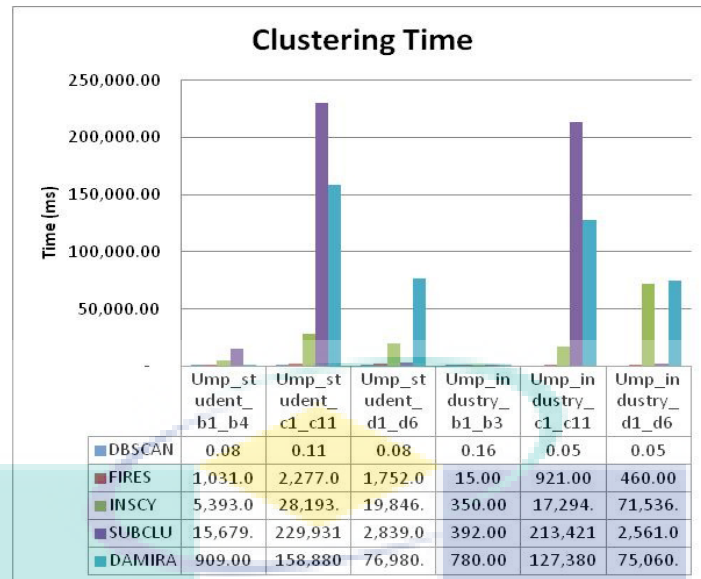


Figure 5.34. Time processing of clustering of higher learning institution datasets

#### 5. 4. 2 Accurate

In addition to evaluating the work efficiency of subspace clustering method, this research discusses related parameters of clustering results. The experimental results show that the accuracy of INSCY method is more accurate than SUBCLU and FIRES.

In the clustering experiments for the glass dataset, methods FIRES have better accuracy than INSCY, SUBCLU and DAMIRA. Meanwhile DAMIRA method it is more accurate than methods INSCY and SUBCLU, as shown in Figure 5.35. But for experimental liver dataset clustering accuracy DAMIRA lower than FIRES method, INSCY and SUBCLU.

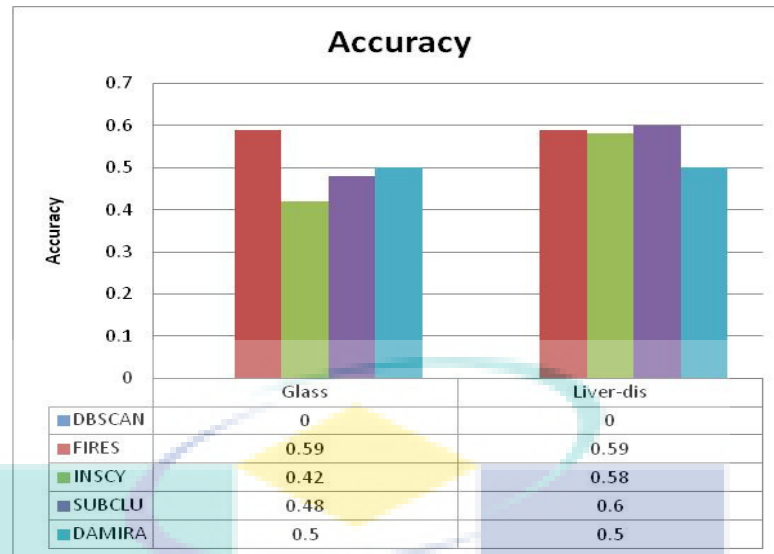


Figure 5.35. Accuracy of real datasets

#### 5.4.3 Coverage

Coverage is used to evaluate the scope of the size of clustering. In real datasets, DAMIRA successfully clustered all of the data. INSCY method has a lower coverage than FIRES method, as shown in Figure 5.36. INSCY and DAMIRA, even for job satisfaction clustering only 6%.

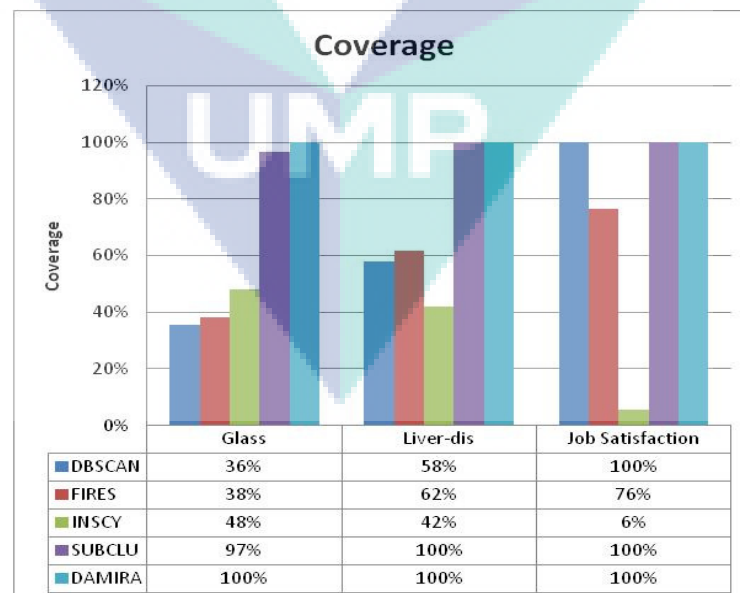


Figure 5.36. Coverage of real datasets

Similarly, higher learning institution clustering dataset, DAMIRA managed to cover all existing data, as shown in Figure 5.37. INSCY method even fails to cluster the data Ump\_student\_b1\_b4.

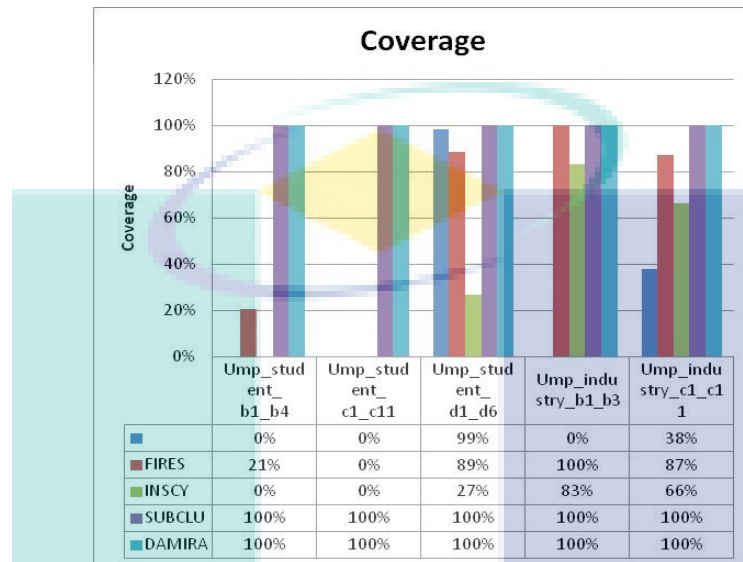


Figure 5.37. Coverage of higher learning institution datasets

#### 5. 4. 4 F1-Measure

F1-Measure generally used to evaluate classifier, but also can be used to evaluate or projected subspace clustering, by measuring the average value of harmony from the cluster, whether all the clusters detected and precision (if all the clusters detected with accuracy). In statistics, the  $F_1$  score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision  $p$  and the recall  $r$  of the test to compute the score:  $p$  is the number of correct results divided by the number of all returned results and  $r$  is the number of correct results divided by the number of results that should have been returned. The  $F_1$  score can be interpreted as a weighted average of the precision and recall, where an  $F_1$  score reaches its best value at 1 and worst score at 0. For F1 Measure SUBCLU method is better than FIRES, INSCY and DAMIRA, as shown in Figure 5.38.

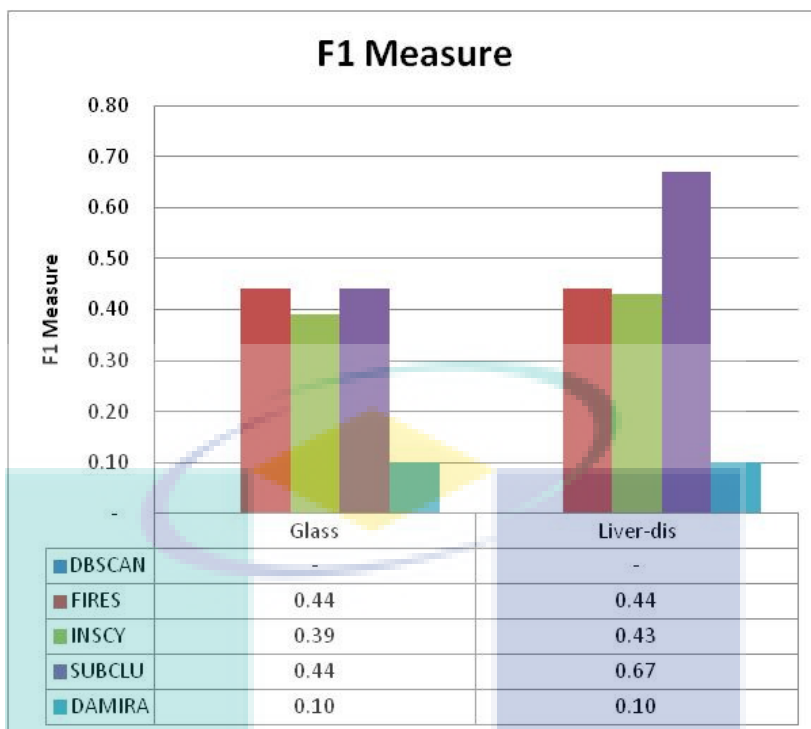


Figure 5.38. F1 measure of real datasets

## 5.5 SUMMARY

Experiment implement in 3 real dataset, and 6 higher education dataset. For each test research uses MinPoints = 6 and Epsilon = 0.9, based on experiment the clustering time of DAMIRA for glass dataset and liver dataset, method is more than 20 times longer than the FIRES, INSCY and SUBCLU, while for job satisfaction dataset DAMIRA need shortest time than others method. For more complex data, the performance DAMIRA looks more efficient than SUBCLU, also tend to produced many clusters, more than the other methods. For higher learning institution dataset, FIRES and INSCY tendency to fail to form clusters, For un-clustered data, SUBCLU and DAMIRA has no un-clustered real datasets, thus the perception of the results of the cluster will produce more accurate information, using FIRES and INSCY not all data can be in the cluster. FIRES have better accuracy than INSCY, SUBCLU and DAMIRA, meanwhile DAMIRA method it is more accurate than methods INSCY and SUBCLU.

## CHAPTER 6

### ONLINE QUESTIONNAIRE FOR EDUCATIONAL DATA MINING

This part described the implementation platform of online questionnaire for educational data. This data will employ for the proposed subspace clustering. The experimental based on three six higher learning institution datasets.

#### 6.1 PLATFORM OF ONLINE QUESTIONNAIRE

One of the instruments collecting real dataset in the research is a questionnaire, closely related to the research problem. This research used closed questions answers, this model were chosen because the answers are standard and can be compared with other people's answers; it is easier to coding and analysed. The answer also can be obtained directly from the question of existing coding, thus saving energy and time.

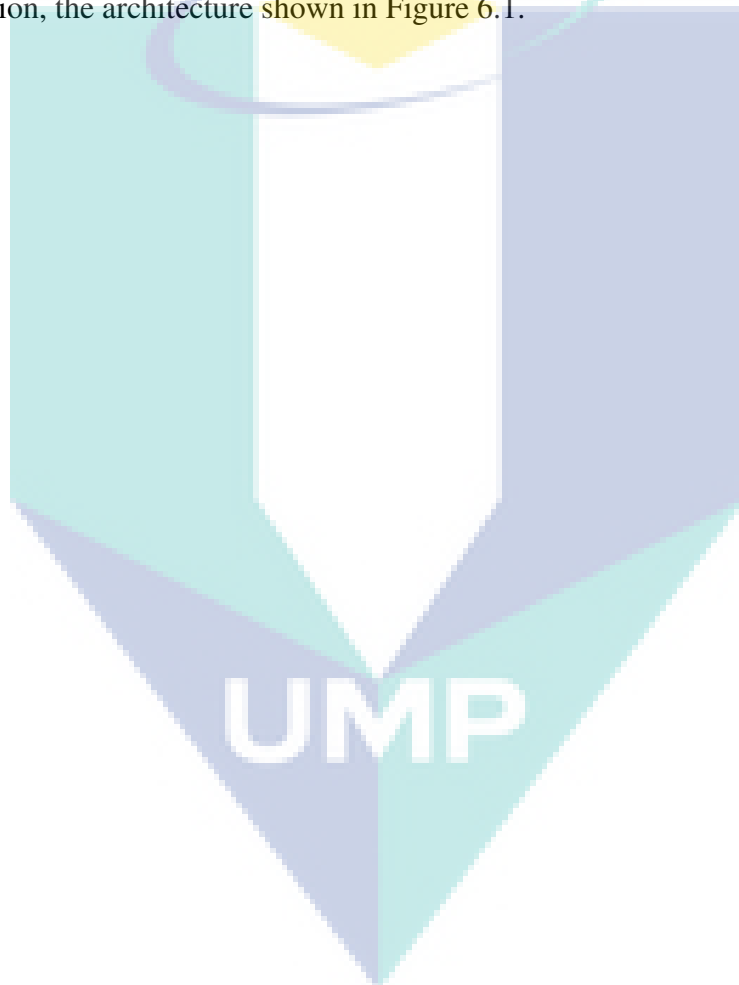
Before the questionnaires distributed to respondents, the trial has been conducted to a small number of respondents, to determine the validity and reliability of measuring instruments in question. Measurement scale used is the size of the Ordinal, which contains the meaning given number of levels. Scale used to measure attitudes, opinions, and perceptions of an alumni and industries.

With a Likert scale, the variables to be measured are translated into indicator variables. Then the indicator is used as a starting point to construct the instrument items



that can be a statement or question, both are favourable (positive) and unfavourable (negative). This research conduct the validity of the instrument accuracy, such as content validity, construct validity, and measure of reliability made for consistency problems.

This research use online questionnaire system, making it easier for the recapitulation of the data and supports the concept and go green. This research requires datasets from the real world. Therefore, was developed a website as a medium access data collection, the architecture shown in Figure 6.1.



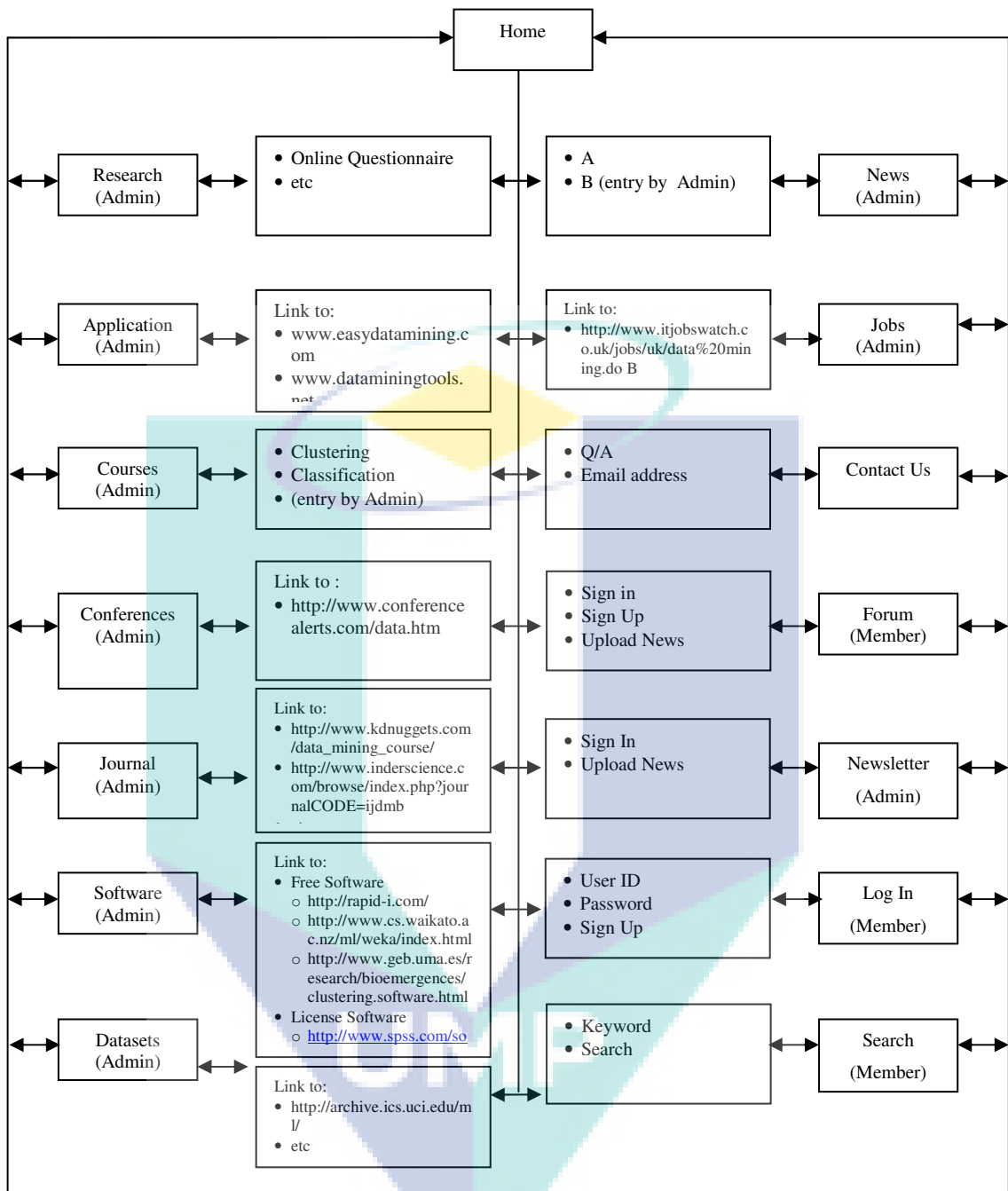


Figure 6.1. Web architecture

As known, internet is one of useful of communication technologies can use as media for online questionnaire, to capture the whole society, reception and dissemination of information for mutual interdependence. Associated with this research, we put online questionnaire in its domain.

In addition to the media access research data collection, the website is also aimed at sharing information related to research in the field of data mining. When the information is accessible online will display as shown in Figure 6.2.



Figure 6.2. Web homepage

This web page provides a “Research Central” features that can be used to access the data entry page. If selected central research will display data access page as shown in Figure 6.3.



Figure 6.3. Homepage of data access

Meanwhile, the custom data access to multiple user classification, arranged as a flowchart in Figure 6.4

UMP

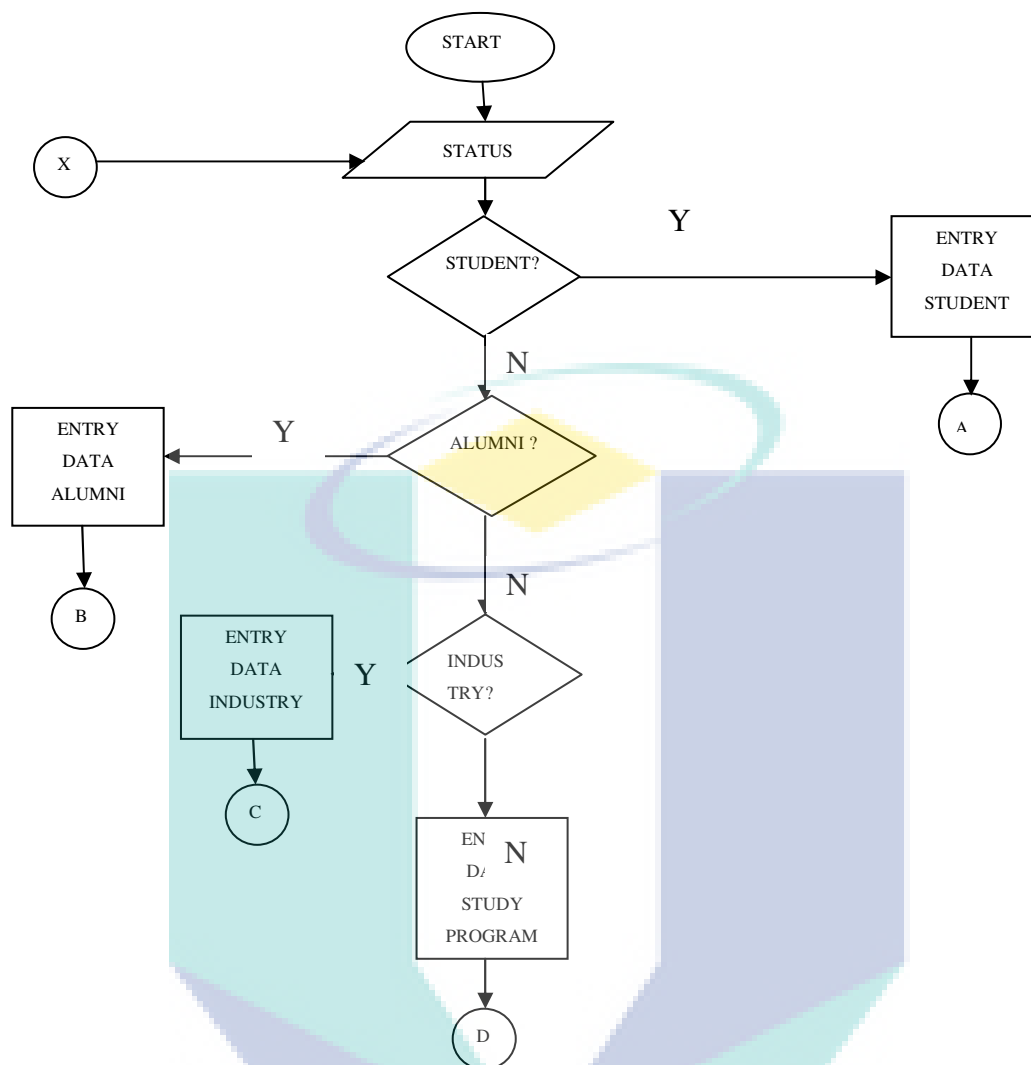


Figure 6.4. Architecture of online questionnaire

There are four categories of users to access the data, as shown in Figure 6.5. Choices Student, Alumni and Industry are to fill in the data access.



Figure 6.5. Key in for add new HLI

While the University is access options for building an online questionnaire required. For example, as shown in Figure 6.6, if choose the University it will show a page of HLI details.

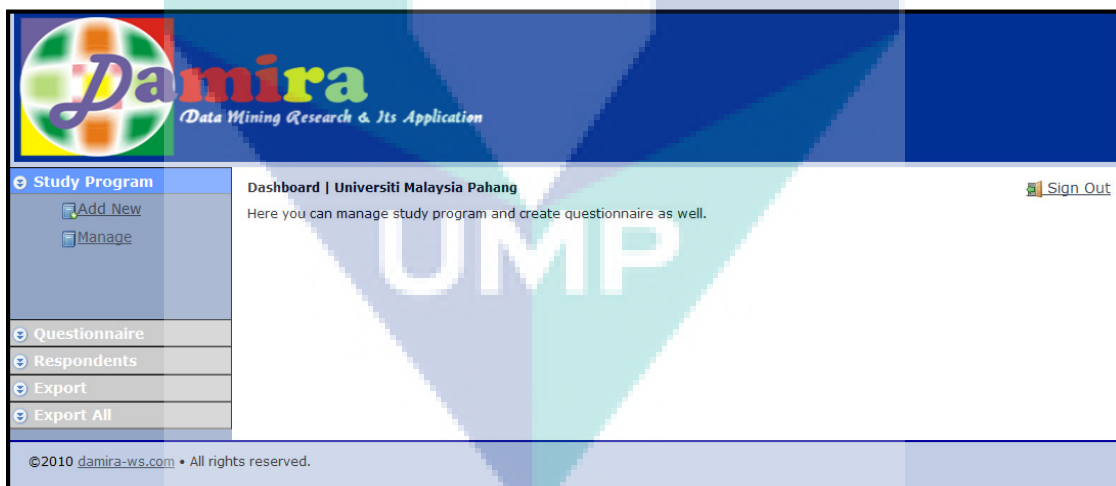


Figure 6.6. Dashboard of HLI detail

While at HLI detail page, a user can manage study program that will be included and editing online questionnaire, as shown in the Figure 6.7.

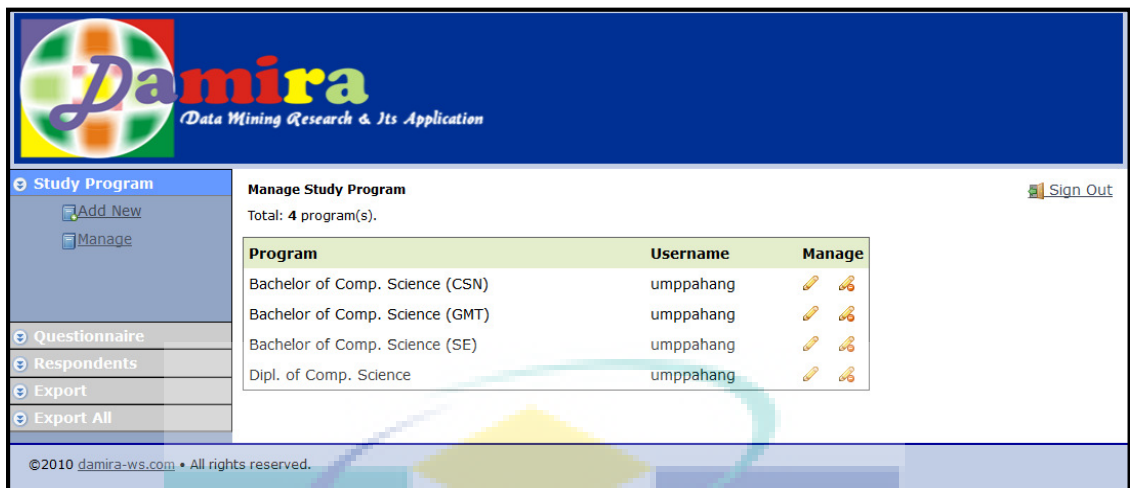


Figure 6.7. Editing online questionnaire

Compiled questionnaire can be edited as needed, as shown in Figure 6.8. Questions can be grouped in accordance with the wishes, and the amount can range from 150-200 questions.



Figure 6.8. Manage questionnaire

If necessary, the data that is already filled respondents can be display by HLI, for industrial data as shown in Figure 6.9



**Damira**  
Data Mining Research & Its Application

**Industrial Respondents Database** [Sign Out](#)

There are 13 respondent(s)

ID	Name	Industry Type	Street	City	State	Country	Zip Code	E-mail	Phone
1	TECHNIP	Other	2nd Floor, Wisma Technip	241, Jalan Tun Razak	Kuala Lumpur	Malaysia	50400	thiyagarajang@technip.com	0321167323
2	ONG MING TECK	Service	No: 103, JALAN UPPER FOOCHOW NO.1,	KUCHING	SARAWAK	MALAYSIA	93300	teck_5867@hotmail.com	0168542065
3	Unomedical Sdn. Bhd	Manufacture	Bekar Arang Industrial Estate	Sg. Petani	Kedah	Malaysia	08000	nua@convatec.com	044556126
4	Azanil Putra Bin Yussof/Telekom Applied Business S	Other	Lingkarang Usahawan 1 Timur	Cyberjaya	Selangor	Malaysia	63000	azanil@tab.com.my	
5	EcoGas Pte. Ltd.	Manufacture	Jln Ampang No. Kuala Lumpur	Kuala Lumpur	Persekutuan	Malaysia	21000		
	Roxy RW		Main Road	Shah					

Figure 6.9. Industrial Respondent Database



**Damira**  
Data Mining Research & Its Application

**Student Respondent Database** [Sign](#)

There are 64 respondent(s)

ID	First name	Last name	Gender	Birth date	Citizenship	Province	Matrices No.	Year intake	Year graduated(predicted)	C
1	Nurul Rahimah	Rahman	Female	1988-05-07	Malaysian	Kuala Lumpur	CA07041	2007	2010	3.
2	MOHD SYAFIQ	BACHOK	Male	1987-10-21	MALAYSIA	JOHOR	CB08013	2008	2011	3.
3	Nur Filzah Hani	Mohamad	Female	1987-02-24	Malaysian	N.Sembilan	CB06057	2006	2010	2.
4	CARLOS	MAH	Male	1987-03-14	MALAYSIAN	BUDDHA	CB07028	2006	2011	3.
5	CHAROMIE	A/L TAT WI	Male	1984-11-14	MALAYSIAN	KELANTAN	CA07059	2007	2010	3.
6	munirah	ab rahman	Female	1987-03-03	malaysia	melaka	cb08010	2008	2011	3.

Figure 6.10. Result of student respondent



Once completed by the respondents, the data can be downloaded and saved in Excel format, as shown in Figure 6.11.

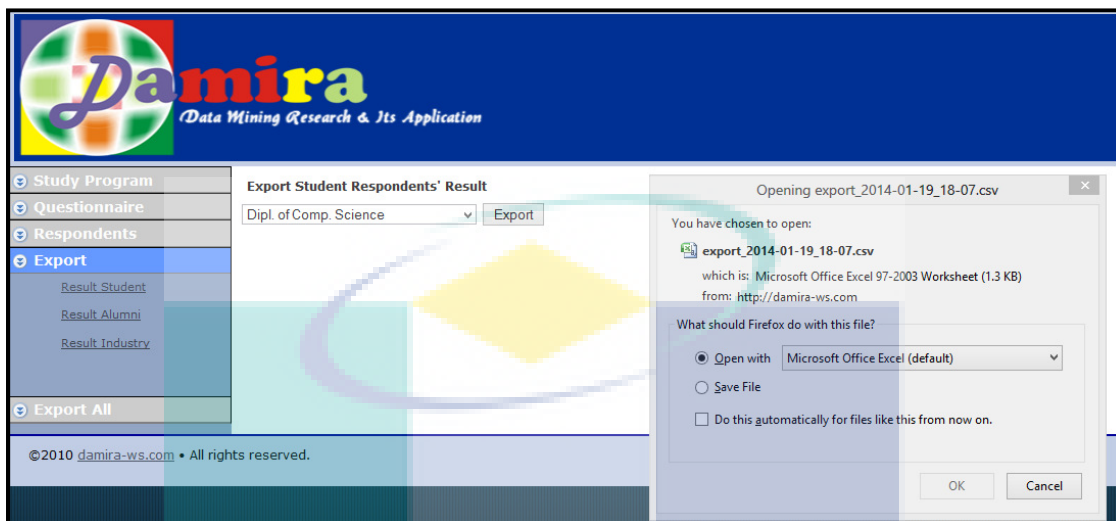


Figure 6.11. Export Student Respondent Result

When accessing a student, college names must be specified origin, and the program of study, as shown in Figure 6.12.

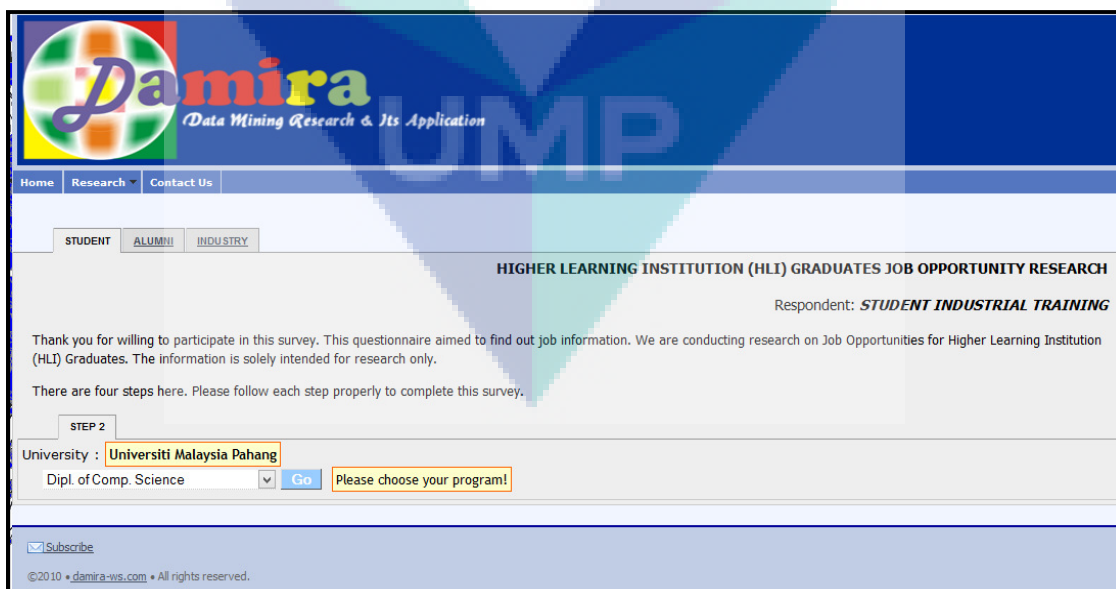


Figure 6.12. University and study program choose

Given the complexity of the possibility of the type and number of questions were provided, it would require a flowchart design as shown in Figure 6.13. This flowchart as a basis for the preparation of student detail page form, which is shown in the Figure 6.14.

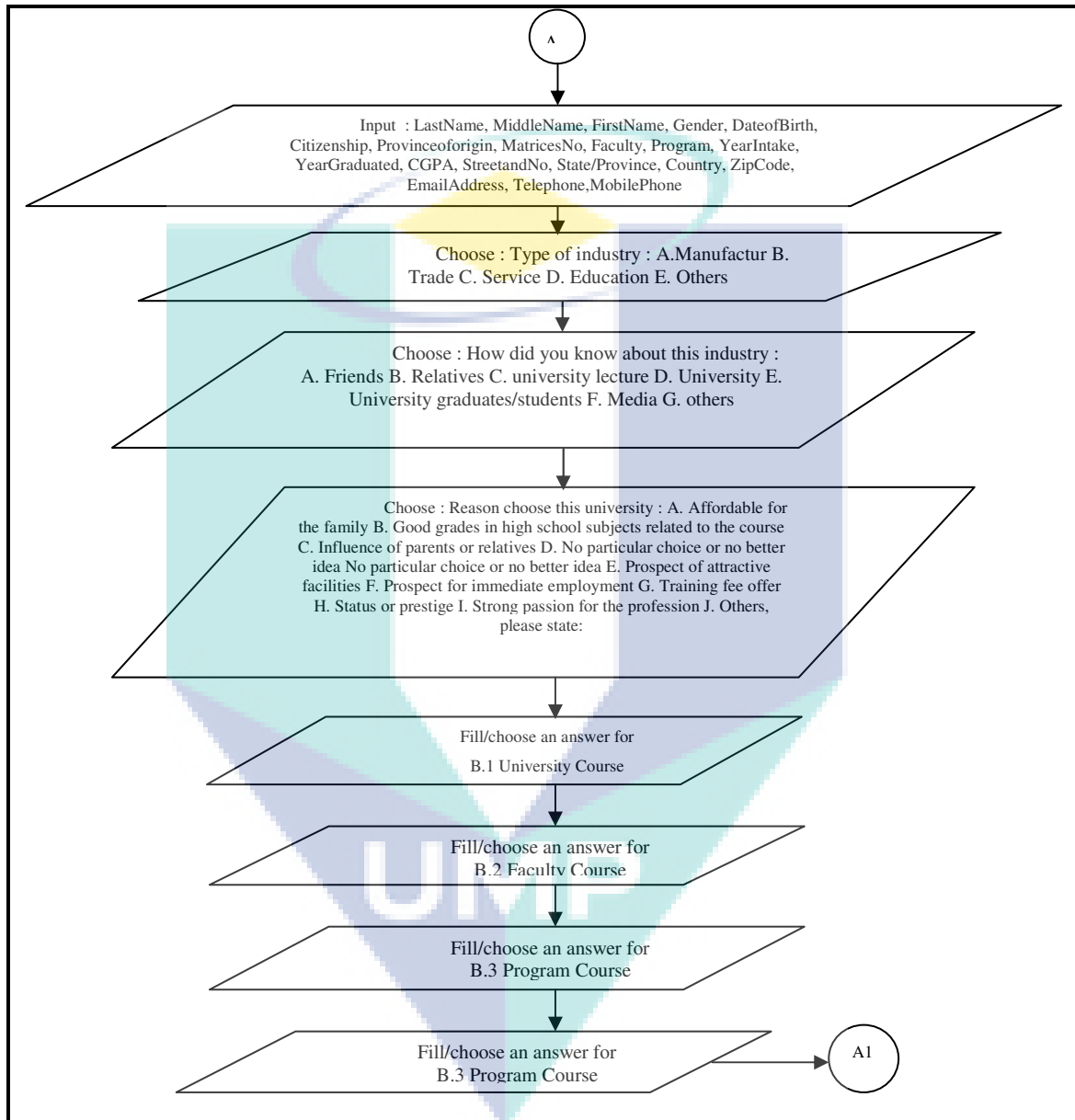


Figure 6.13. Flowchart query of questionnaire

STUDENT ALUMNI INDUSTRY

**HIGHER LEARNING INSTITUTION (HLI) GRADUATES JOB OPPORTUNITY RESEARCH**

Respondent: **STUDENT INDUSTRIAL TRAINING**

Thank you for willing to participate in this survey. This questionnaire aimed to find out job information. We are conducting research on Job Opportunities for Higher Learning Institution (HLI) Graduates. The information is solely intended for research only.

There are four steps here. Please follow each step properly to complete this survey.

STEP 3

University: **Universiti Malaysia Pahang**

Faculty : **Dipl. of Comp. Science**

**STUDENT PROFILE**

\*) is important

First name\*

Last name\*

Gender\*

Birth date\* 01 - 01 - 1970

Citizenship\*

Province of Origin

Matrices No. \*

Year intake \*

Year graduated (predicted) \*

CGPA \*

Street

Figure 6.14. Student details form

Given the complexity of the possibility of the type and number of questions were provided, it would require a flowchart design. Flowchart as shown in Fig., The basis for the preparation of student detail page form, which is shown in the Figure 6.15.

UMP

Figure 6.15. Online Questionnaire section

Question list as display in Figure 6.16, Figure 6.17 and Figure 6.18.

No.	Course	Choose the level of frequency of implemented course in your work environment							
		NEVER	2	3	4	5	6	7	FREQUENTLY
1	Communication Malaysian Studies	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2	Islamic Institutions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3	Mandarin for Beginners	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4	Soft Skills II	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5	English for General	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6	Briged Siswa	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 6.16. Question structure for frequency of course implemented

Home Research Contact Us

STUDENT ALUMNI INDUSTRY

HIGHER LEARNING INSTITUTION (HLI) GRADUATES JOB OPPORTUNITY RESEARCH

Respondent: *STUDENT INDUSTRIAL TRAINING*

Thank you for willing to participate in this survey. This questionnaire aimed to find out job information. We are conducting research on Job Opportunities for Higher Learning Institution (HLI) Graduates. The information is solely intended for research only.

There are four steps here. Please follow each step properly to complete this survey.

STEP 4

Name : [ ]

University: **Universiti Malaysia Pahang**

Faculty : **Dipl. of Comp. Science**

Click to open/close each section and choose according to your opinion

Section A: COMPANY PROFILE

Section B: FREQUENCY OF COURSE IMPLEMENTED WITH INDUSTRIAL TRAINING ENVIRONMENT

Section C: IMPORTANCE KNOWLEDGE COMPETENCE

In Section C, there are 11 parts of questions about importance of knowledge competence during industrial training. You need to choose by clicking on the level of importance of knowledge competence.

C.1 Algorithm Capability

Choose the importance of knowledge competence carrying out the task of your work environment

No.	Performance Capability	1	2	3	4	5	6	7	8	9	10
1	Prove theoretical results	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2	Develop solutions to programming problems	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 6.17. Question structure for important knowledge competence

Home Research Contact Us

STUDENT ALUMNI INDUSTRY

HIGHER LEARNING INSTITUTION (HLI) GRADUATES JOB OPPORTUNITY RESEARCH

Respondent: *STUDENT INDUSTRIAL TRAINING*

Thank you for willing to participate in this survey. This questionnaire aimed to find out job information. We are conducting research on Job Opportunities for Higher Learning Institution (HLI) Graduates. The information is solely intended for research only.

There are four steps here. Please follow each step properly to complete this survey.

STEP 4

Name : [ ]

University: **Universiti Malaysia Pahang**

Faculty : **Dipl. of Comp. Science**

Click to open/close each section and choose according to your opinion

Section A: COMPANY PROFILE

Section B: FREQUENCY OF COURSE IMPLEMENTED WITH INDUSTRIAL TRAINING ENVIRONMENT

Section C: IMPORTANCE KNOWLEDGE COMPETENCE

Section D: IMPORTANCE OF SOFT SKILLS COMPETENCE

In Section D, there are 6 parts of questions about soft skill competence during industrial training. Your competencies reflect a combination of talents, knowledge, skills, and behaviours that you use to get things done.

D.1 Resource Management

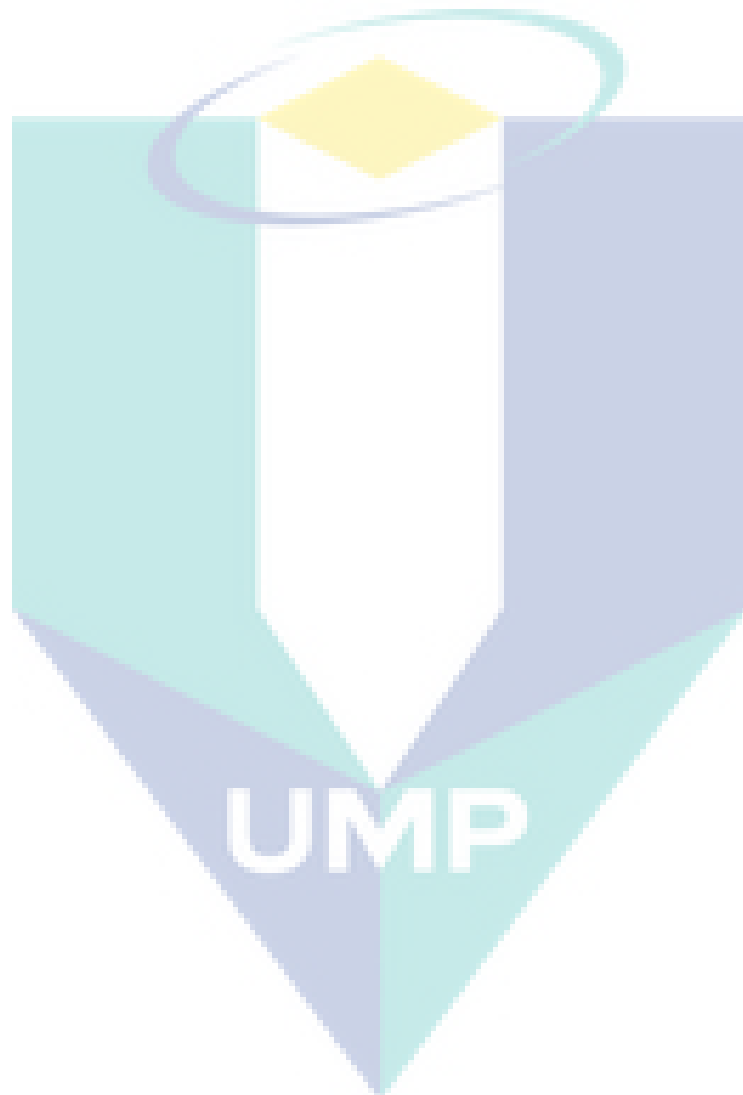
Choose the importance of soft skill competence carrying out the task of your work environment

No.	Competencies	1	2	3	4	5	6	7	8	9	10
1	Budget management	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 6.18. Question structure for importance of soft skill competence

## 6.2 SUMMARY

Experiment implement in 6 higher education dataset. There are four categories of users to access the data; there are student, alumni, industry and university. This online questionnaire can design by user needs.



## CHAPTER 7

### CONCLUSION AND FUTURE WORK

This part described summarizes and discusses the major contributions of the thesis. It concludes with pointing out some future research directions.

#### 7.1 CONCLUSION

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 methodology for classification, clustering, and outlier detection. The used of clustering algorithms to measure the similarity of high-dimensional data or attributes often achieve undesired results, causing the attributes to become unrelated or either too close together. Closed data can form overlap groups either from dense clusters. Data may be found in different clusters and also in a different subspace. While a cluster is seriously challenged in high dimensional spaces, where subspace is considered when subset of attribute or data points belongs to different clusters in different subspaces.

Subspace clustering is projected as a search technique for grouping data or attributes in different clusters. Grouping was done by determining the level of data density and was also done to identify outliers or irrelevant data that will create each to cluster exist in a separate subset.

The present research estimate density dimensions and the results used as input data to determine the initial cluster based on density connection, using DBSCAN algorithm. Each dimension 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.

This study improved model for subspace clustering based on density connection, to cope with the challenges clustering in educational data mining, named as DAMIRA. The main idea in DAMIRA 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.

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.

An improved clustering technique was applied to analyze the possible cluster between the knowledge competence skill and soft skill competence among student industrial training and industries. There are some subspace cluster technique, such as SUBCLU, FIRES and INSCY to be tested, and DAMIRA use as main clustering technique based on density connection.

The step of DAMIRA are: the first step is change n-dimension to 1-dimension, second is find out the initial cluster by using DBSCAN, third step is found first cluster and first subspace, fourth step is to determine the candidate subspace over clusters, and lastly choose best subspace.

Experimental implemented in 3 real dataset, and 6 higher education dataset. To verify the quality of the clustering obtained through our technique (DAMIRA) and to expedite the first phase, we run DBSCAN, FIRES, INSCY, and SUBCLU. DAMIRA successfully established very large the number of clusters for each dataset, while FIRES and INSCY, tendency to fail to form clusters in each subspace.



The clustering time for glass dataset and liver dataset, DAMIRA method is more than 20 times longer than the FIRES, INSCY and SUBCLU, meanwhile for job satisfaction dataset DAMIRA need shortest time than SUBCLU and INSCY methods.

For more complex data, the performance DAMIRA looks more efficient than SUBCLU, also tend to produced many clusters, beyond than the other methods. For higher learning institution dataset, FIRES and INSCY tendency to fail to form clusters, For un-clustered data, SUBCLU and DAMIRA has no un-clustered real datasets, thus the perception of the results of the cluster will produce more accurate information, using FIRES and INSCY not all data can be in the cluster. FIRES have better accuracy than INSCY, SUBCLU and DAMIRA, meanwhile DAMIRA method it is more accurate than methods INSCY and SUBCLU.

Coverage is used to evaluate the scope of the size of clustering, in real datasets, 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.

## 7.2 FUTURE WORKS

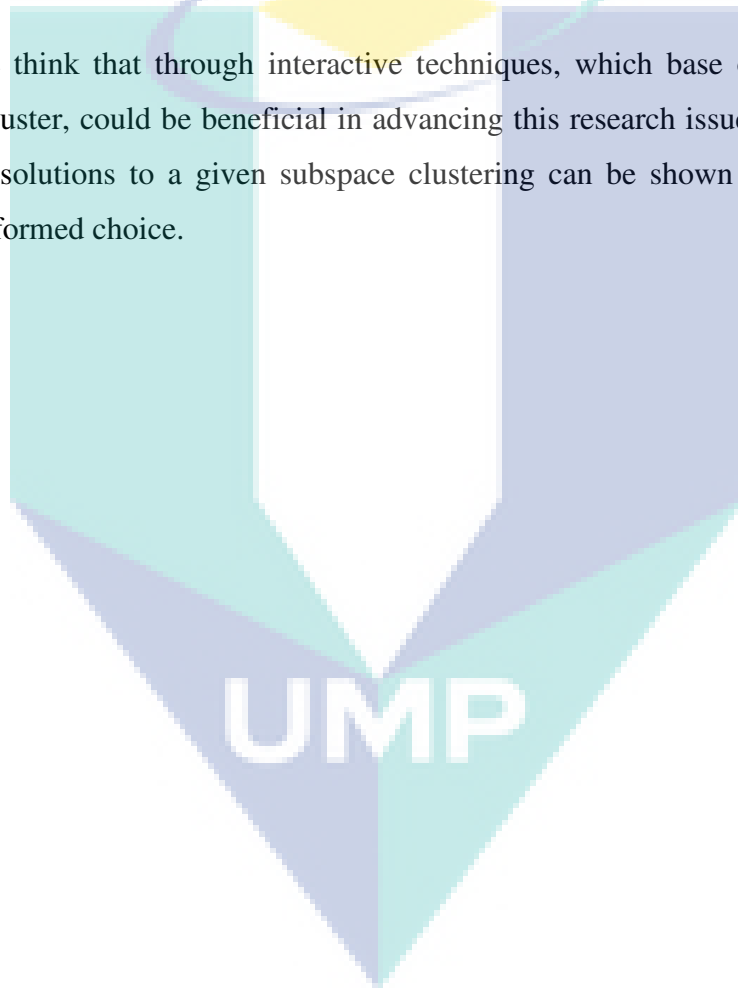
There are still more research avenues in subspace clustering to be developed in the future. As discussed above, we have studied an improved subspace clustering based on density connection, and compare with another method. Depending on the data, benchmark real world data from the UCI archive and real dataset from education survey from SIT and industrial.

Other issues that are important for research in this area are the runtime behavior of more complex data. Clearly an overall density based connection for non nominal data is difficult to solve in practice, and more efficient algorithmic solutions are necessary. Another issue that still needs to be resolved is the instantiation of image subspace clustering, and would require deepest research. This is crucial for being able to compute improved subspace clustering for interactive exploration within reasonable response times.

It should also be interesting to study the effectiveness of our method combined with pre-processing data, so it can be interesting to adapt our method for supervised or semi-supervised learning.

However, existing approaches use a density based connection approach comparing across the high dimensionality of the dataset. Our method finds similar data in subspaces of the dataset, each representing a particular subspace.

We think that through interactive techniques, which base on visualization of subspace cluster, could be beneficial in advancing this research issue. This implies that alternative solutions to a given subspace clustering can be shown to enable users to make an informed choice.



## REFERENCE

- Abonyi, J. and Balázs, F. 2007. *Cluster Analysis for Data Mining and System Identification*, Berlin (Ed.). Germany: Birkhauser Verlag AG.
- ACM and IEEE. 2005. Computing Curricula – *The Overview Report 2005*, ISBN: 1-59593-359-X, ACM Order Number: 999066, IEEE Computer Society Order Number: R0236, pp. 23-25
- Aggarwal, C.C., Cecilia P., Joel L.W., Philip S.Y and Jong S.P. 1999. Fast Algorithms for Projected Clustering. *Proceeding of SIGMOD*, pp.61-72
- Agrawal, R., Gehrke, J., Johannes, G. And Raghavan, P. 2005. *Automatic Subspace Clustering of High Dimensional Data*, Data Mining and Knowledge Discovery (11) 1: pp.5-33
- Alpaydin, E. 2010. *Introduction To Machine Learning Adaptive Computation And Machine Learning* - 2nd edition, The MIT Press Cambridge, Massachusetts London, England
- Aliguliyev, Ramiz. M. 2009. *Performance evaluation of density-based clustering methods*, Journal of Information Sciences-Elsevier, No. 179, pp.3583-3602
- Antonenko, P.D., Toy, S., Niederhauser, D.S., 2012, *Using Cluster Analysis for Data Mining Educational Tehcnology Reserahch*, Education Tech Research Dev., Vol. 60, pp.383-398
- Assent, I., Krieger, R., Muller, E. and Seidl, T. 2008. INSCY: Indexing subspace clusters with in-process-removal of redundancy. *Proceeding of Eight International Conference on Data Mining (ICDM)*, pp.719-724
- Asuncion, D.N. 2007. *UCI Machine Learning Repository*.
- Aviad, B and Gelbard, R. 2011. *Classification by clustering decision tree-like classifier based on adjusted clusters*. Expert Systems with Applications 38: pp.8220–8228
- Baker, R.S.J.D. 2005. Data Mining for Education, Pre-print draft, download from <http://www.columbia.edu/~rsb2162/>, (12 December 2012).
- Baker, R.S.J.D. and Yacef, K. 2009. The State of Educational Data Mining in. A Review and Future Visions Data Mining for Education, *Journal of Educational Data Mining, Volume 1, Issue 1*, pp. 3-17
- Baldi, P. and Søren, B. 2001. Bioinformatics, *The Machine Learning Approach*, The MIT Press, Cambridge, Massachusetts, London, England

- Banfield, J. D. and Raftery, A.E. 1993. Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49, pp. 803–821
- Baraldi, A and Alpaydin, E. 2002. Constructive Feedforward ART clustering networks–Part I and II. *IEEE Transactions on Neural Networks*, 13(3), pp.645–677
- Barrientos, F. and Gregorio, S. 2012. *Interpretable knowledge extraction from emergency call data based on fuzzy unsupervised decision tree*. Knowledge-Based Systems 25: pp.77–87
- Beckmann N., Kriegel H.P., Schneider R, and Seeger B. 1990. *The R\*-tree: An Efficient and Robust Access Method for Points and Rectangles*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Atlantic City, NJ, 1990, pp. 322-331.
- Bellman, R.E. 1957. *The Theory of Dynamic programming*, The Rand Corporation, Santa Monica, USA
- Berka, P., Jan, R. and Djamel A.Z. 2009. *Data Mining And Medical Knowledge Management*, Medical Information Science Reference, New York,
- Berry, W.M. and Malu, C. 2008. *Survey of Text Clustering*, Springer-Verlag London Limited, London, UK
- Bertini J.R., Liang, Z. and Robson, M. 2011. Alneu de Andrade Lopes, A *nonparametric classification method based on K-associated graphs*, *Information Sciences 181*: pp.5435–5456
- Bi, J., Kristin, B., Mark, E., Curt, B. and Minghu, S. 2003. *Dimensionality Reduction via Sparse Support Vector Machine*”, *Journal of Machine Learning Research* 3 pp.1229-1243
- Bicici, E. and Deniz, Y. 2007. Local Scaled Density Based Clustering. *ICANNGA*. pp. 739-748
- Bidgoli, B.M., Kashy, D.A., Kortemeyer, G., Punch, W.F., 2003, *Predicting Student Performance: An Application Of Data Mining Methods With an Educational Web-Based System*, 33rd ASEE/IEEE Frontiers in Education Conference, pp. 13-18.
- Bin, D., Shao P. and Zhao, D. 2008. *Data Mining for Needy Students Identify Based on Improved RFM Model: A Case Study of University*. International Conference on Information Management, Innovation Management and Industrial Engineering (ICIII), Volume. 1. pp.244-247.

- Boulicaut, J.F., Floriana, E.F.G. and Dino, P (Eds.). 2004. Machine Learning, ECML 2004, *LNAI 3201, Springer-Verlag Berlin Heidelberg*. pp. 239–249
- Brinkhoff T., Kriegel H.P., Schneider R., and Seeger B. 1994, *Efficient Multi-Step Processing of Spatial Joins*, Proc. ACM SIGMOD Int. Conf. on Management of Data, Minneapolis, MN, 1994, pp. 197-208.
- Brusilovsky, P. and Christoph, P. 2003. Adaptive and Intelligent Web-based Educational Systems. *International Journal of Artificial Intelligence in Education 13*. pp.156–169
- Callahan, D. and Bob, P. 2007. *Educating Experienced IT Professionals by Addressing Industry's Needs*, IEEE Software Magazine September/October 2002, pp.57-62
- Cardoso, A.R. 2007. Jobs for young university graduates. *Economics Letters, Volume 94, Issue 2, February*, pp 271-277.
- Cestnik, G., Kononenko, I and Bratko, I. 1987. Assistant-86: A Knowledge elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, Sigma Press. pp.31-45,
- Chady, E.M., Khair, M., Zakhem, W., 2011, Improving Student's Performance Using Data Clustering and Neural Networks in Foreign-Language Based Higher Education, *The Research Bulletin of Jordan ACM, Vol.II No.III*, pp.27-34.
- Chakrabarti, K. and Sharad, M. 2000. Local Dimensionality Reduction : A New Approach To Indexing High Dimensional Space, *Proceeding Of The 26th VLDB Conference, Cairo, Egypt*. pp.89-100.
- Chandra, B and Manish, G. 2011. *Robust approach for estimating probabilities in Naïve-Bayes Classifier for gene expression data*. Expert Systems with Applications 38: pp.1293–1298
- Chellatamilan, T., Ravichandran, M., Suresh, R. M., Kulanthaivel, G. 2011, *Effect of Mining educational Data to improve Adaptation of learning in e-Learning System*, Second International Conference on Sustainable Energy and Intelligent System, pp.922-927
- Chen, V.J., Razip, A.M., Ko, S., Qian, C.Z., Elbert, D.S., 2012, *SemanticPrism: a Multi-Aspect View of Large High-Dimensional Data*, IEEE Symposium on Visual Analytics Science and Technology, pp.259-260
- Chu, Y.H., Jen, W.H., Kun, T.C., De, N.Y and Ming, S.C. 2010. Density Conscious Subspace Clustering for High-Dimensional Data. *IEEE Transactions On Knowledge And Data Engineering*, Vol. 22, No. 1, January 2010

- Chua, F. and Chong, T. 2009. “*Dimensionality Reduction and Clustering of Text Document*”, [www.mysmu.edu/phdis2009/freddy.chua.2009/papers/probabilistic.pdf](http://www.mysmu.edu/phdis2009/freddy.chua.2009/papers/probabilistic.pdf)
- Clark, P. and Niblett, T. 1987. *Induction in Noisy Domains*. Machine Learning (from *the Proceedings of the 2<sup>nd</sup> European Working Session on Learning*), Bled, Yugoslavia: Sigma Press. pp.11-30
- Cocozza, Sergio, 2007, *Methodological aspects of the assessment of gene–nutrient interactions at the population level*, Nutrition, Metabolism and Cardiovascular Diseases, Volume 17, Issue 2, February 2007, pp.82–88.
- Cordeiro, R.L.F., Caetano, T.Jr., Agma, J.M.T., Julio, L., Kang, U., Christos, F. 2010. *Finding Clusters in Subspaces of Very Large Multi-dimensional Datasets*. 26th International Conference on Data Engineering (ICDE) pp.625-636.
- Cortes, C. and Vladimir, V. 1995. Support-Vector Networks, *Machine Learning*, pp.273-297.
- Creswell. J.W and Vicki L.P.C. 2007. *Designing, and Conducting Mixed Methods Research*, London: SAGE Publication Inc., London, UK
- Creswell. J.W. 2009. *Research Design: Qualitative, Quantitative and Mixed Methods Approaches*. SAGE Publication Inc.. 3<sup>rd</sup> ed. London, UK
- Cristianini, N. and John, S.T. 2000. *An Introduction to Support Vector Machines: And Other Kernel-Based Learning*, Cambridge University Press, Cambridge, UK,
- Cunningham, P. 2007. *Dimension Reduction*, Technical Report UCD-CSI.
- Datong, L., Peng, Y. and Peng, X. 2011. *Online Adaptive Status Prediction Strategy for Data Driven*. China. Fault Prognostics of Complex Systems
- Dave, R.N. 1996. *Validation Fuzzy Partitions Obtained Through C-Shells Clustering*, *Pattern Recognition Letters* 17, pp.613 – 623.
- Dayan, P. 2008. Unsupervised Learning, *The MIT Encyclopedia of the Cognitive Sciences*. pp.1-7
- Dekker, G.W., Pehcenizkiy, M., Vleeshouwers, J.M., 2009, Predicting Students Drop Out: A Case Study, *2nd International Conference On Educational Data Mining*, Cordoba, Spain, pp.

- Devaraj, S. and Babu, S.R. 2004. *How To Measure The Relationship Between Training and Job Performance*, Communications of the ACM, Volume 47, Issue 5, pp.62-67.
- Ding, C. and Tao, L. 2007. *Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering*, International Conference on Machine Learning, Corvallis, pp.1-8
- Donoho, D.L. 2000. *High Dimensional Data Analysis: The Curses and Blessing of Dimensionality*, <http://mlo.cs.man.ac.uk/resources/Curses.pdf>, 04 November 2013
- Ester, M., Kriegel, H.P., Jörg, S. and Xiaowei, X. 1996. *A Density-Based Algorithm for Discovering Clusters*, 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp.1-6
- Fan, P., Li, G., Yuan, L. and Li, Y. 2011. Vague continuous K-nearest neighbor queries over moving objects. *Journal of Information System, Volume 37, Issue 1, March 2012* pp.13-32
- Fan, J, Li, R., 2006. Statistical challenges with high dimensionality: feature selection in knowledge discovery, *International Congress of Mathematician, Madrid, , pp.595-622*
- Farquard, M.A.H., Ravi, V. And Raju, S.B. 2010. *Rule Extraction from Support Vector Machine Using Modified Active Learning Based Approach: An Application to CRM*. KES (1) pp.461-470
- Fayyad, U., Gregory, P.S. and Padhraic, S. 1996, From Data Mining to Knowledge Discovery in Database, *American Association for Artificial Intelligence*, Volume 17 Number 3, pp.37-54
- Felice, M.D. and Xin, Y. 2011. *Short-Term Load Forecasting with Neural Network Ensembles. A Comparative Study*. IEEE Computational Intelligence Magazine, August 2011, pp.47-56
- Finley, T. and Thorsten, J. 2010. *Supervised Clustering with Support Vector Machines*, International Conference on Machine Learning, Bonn, Germany, pp.1-8.
- Fodor, I.K. 2002. A Survey of Dimension Reduction Techniques. *LLNL Technical Report, UCRL-ID-148494*, pp.1-18.
- Fung, G.M. 2005. *Multicategory Proximal Support Vector Machine Classifiers*, Machine Learning, volume 59, pp.77-97.

- Gan, G., Jianhong, W. and Zijiang, Y. 2006. A Subspace Clustering Algorithm for High Dimensional Categorical Data. *International Joint Conference on In Neural Networks (IJCNN)*, pp.4406-4412.
- Gao, J. 2005, Clustered SVD strategies in latent semantic indexing, *Information Processing and Management 41*, Elsevier, pp.1051–1063.
- Gath, I. and A. B. Geva. (1989). “Unsupervised Optimal Fuzzy Clustering,” *IEEE Transaction on Patterns Analysis and Machine Intelligence* Volume 11 No.7, pp.773-780.
- Ghahramani, Z. 2004. Unsupervised Learning. In Bousquet, O., Raetsch, G. and von Luxburg, U. (eds), *Advanced Lectures on Machine Learning LNAI 3176*. Springer-Verlag. pp.1-32
- Globerson, A. and Naftali, T. 2003. Sufficient Dimensionality Reduction, *Journal of Machine Learning Research* 3, pp.1307-1331.
- Gradojevic, N. And Ramazan, G. 2011. *Financial Applications Of Nonextensive Entropy*. IEEE Signal Processing Magazine, September 2011, pp.116-141
- Grossman, R.L. 1996. Data Mining Challenges for Digital Libraries, *Journal ACM Computing Surveys (CSUR)*, Volume 28A Issues 4es Article No. 108 December (1996)
- Gunnemann, S., Hardy, K. and Thomas, S. 2009. *Subspace Clustering for Uncertain Data*, SIAM International Conference on Data Mining, pp.245-260.
- Halkidi, M., Yanis Batiskadis and Michalis Vazirgiannis, 2001. *On Clustering Validation Technique*, Journal of Intelligent Information System, Vol. 17:2/3, pp. 107-145
- Han, J. and Micheline, K. 2006. *Data Mining: Concepts and Techniques, 2nd Edition*, pp.25-26.
- Hamalainen, W., Suhonen, J., Sutinen, E., Toivonen, H., 2004, Data mining in personalizing distance education courses", *World Conference On Open Learning And Distance Education*, Hong Kong, 2004, pp.15-21
- Hanna, M., 2004, *Data mining in the e-learning domain*, Computers & Education Journal, Vol. 42, No. 3, pp. 267–287
- Hand, D., Heikki, M. and Padhraic, S. 2010. Principles of Data Mining. Massachusetts Institute of Technology Press, Cambridge, Massachusetts, London, England



- Hang, Z.Q. and Yang, J.Y. 1991. *Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane*, Pattern Recognition, Vol. 24, No. 4, pp. 317-324
- Herawan, T. and Mustafa, M.D. 2011. A soft set approach for Association Rules Mining, *Journal Knowledge-Based Systems* Volume 24 Issue 1, February, 2011, pp. 186-195
- Heuchan, A.J.F. 2003. Industry and Higher Education—Meeting the Needs of the Mining, *Engineering Sector 105th Annual General Meeting of the Canadian Mining, Metallurgy and Petroleum*, Montreal, Quebec, pp.1-15
- Horng, S.C., Feng, Y.Y. and Shieh, S.Lin. 2011. *Hierarchical fuzzy clustering decision tree for classifying recipes of ion implanter*. Expert Systems with Applications 38. pp 933–940
- Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E. and Zimek, A. 2010. Can Shared-Neighbour Distances Defeat the Curse of Dimensionality, *Proceedings of the 21th International Conference on Scientific and Statistical Database Management (SSDBM) 2010*, pp.482-500.
- Hsia, T.C., An, J.S., Li, C.C. 2008. *Course planning of extension education to meet market demand by using data mining technique – an example of Chinkuo Technology University in Taiwan*, Expert Systems with Applications Volume 34 (2008), pp.596–602
- Hsu, C.M. 2011. *Forecasting Stock/futures Prices by Using Neural. China Networks with Feature Selection*, IEEE. pp.1-7
- Hsu, J. 2002. Data Mining Trends And Developments 2002. The Key Data Mining Technologies and Applications for the 21st Century, *The Proceedings of ISECON, San Antonio*, pp.1542-1547
- Huang, W., Lifei, C. and Qingshan Jiang, 2010. *A Novel Subspace Clustering Algorithm with Dimensional Density*, IEEE, pp.71-75.
- Huang, Z., Xudong, L. and Huilong, D. 2011. *Expert Systems with Applications. Expert Systems with Applications* . 38. pp 9483–9490
- Ichihashi, H, A. Notsu, and K. Honda 2010, Semi-hard c-Means Clustering with Application to Classifier Design, *IEEE International Conference on Fuzzy Systems*, pp.1-8

- Ivancevic, Vladimir Ivančević, Milan Čeliković, Ivan Luković, 2012, The Individual Stability of Student Spatial Deployment and its Implications, *International Symposium on Computer in Education*, pp.1-4
- Jain, A.K., Murty., M.N. and Flynn, P. J. 1999. Data clustering: A review. *ACM Computing Surveys*, 31(3), pp.264–323.
- Jiang Z. and Yi, D. 2010. Improving KNN Based Text Classifications. *2nd International Conference on Future Computer and Communication, IEEE*, pp. 317-321
- Jiangsheng, Y. 2002. Method of k-Nearest Neighbours, *Institute of Computational Linguistics, Peking University, China*, 2002  
<http://www.nlp.org.cn/docs/20020903/36/kNN.pdf>, 15 April 2012
- Jones, M.K., Richard, K.J., Paul, L.L. and Peter, J.S. 2009. *Training, Job Satisfaction and Workplace Performance in Britain: Evidence from WERS, LABOUR, Volume 23*, pp.139–175.
- Kailing, K, Kriegel, H.P. and Kroger, K. 2004. *Density-Connected Subspace Clustering for High-Dimensional Data*. 4th SIAM International Conference of Data Mining, pp.246-257
- Kambhatla, N., Todd, K., Kambhatla, A. and Todd, K.L. 1994. *Fast. Non\_Linear Dimension Reduction. Advances in Neural Information Processing Systems 6*, San Francisco, CA, Morgan Kaufmann Publishers.
- Kantardzic, M. 2011. *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley & Sons, Hobokke, New Jersey, USA
- Khalid, S. 2010. *Motion-based behaviour learning, profiling and classification in the presence of anomalies*, *Journal Pattern Recognition*. vol 43 pp.173 – 186.
- Khalilian, M and Norwati Musthapa, 2012. Data Stream Clustering: Challenges and Issues, *International MultiConference of Engineers and Computer Scientists 2010*, Hongkong, pp.1-4
- Knauf, R., Sakurai, Y., Takada, K., Tsuruta, S., 2012, A Case Study on Using Personalized Data Mining for University Curricula, *IEEE International Conference on Systems, Man, and Cybernetics*, pp.3051-3056
- Kohonen, T., Kaski, S. and Lappalainen, H. 1997. *Self-organized formation of various invariant-feature filters in the adaptive-subspace SOM*. *Neural Computation*, Volume 9. pp.1321-1344.

- Koutri, M., Avouris, N., Daskalaki, S., 2005, *A survey on web usage mining techniques for web-based adaptive hypermedia systems*, Adaptable And Adaptive Hypermedia Systems, IRM Press, pp. 125-149.
- Krejcie, R.V. and Daryle W.M., 1970. *Determining Sample Size For Research, Educational and Psychological Measurement*, 30, pp.607-610
- Kriegel, H.P., Karsten, M.B, Kröger, P., Alexey, P., Matthias, S. and Arthur, Z. 2007. Future trends in data mining, *Data Mining Knowledge Discovery*, pp.87–97.
- Kriegel, H.P., Kroger, P. and Arthur, Z. 2009a. Clustering High-Dimensional Data : A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering, *ACM Transactions on Knowledge Discovery from Data*, Vol. 3, No. 1, pp.1-58.
- Kriegel, H.P., Kroger, P., Matthias, R. and Sebastian, W. 2009b. *A Generic Framework for Efficient Subspace Clustering of High-Dimensional Data*, Fifth IEEE International Conference on Data Mining (ICDM'05), pp.1-8
- Kulluk, S., Lale, O. and Adil, B. 2012. *Training neural networks with harmony search algorithms for classification problems*, Engineering Applications of Artificial Intelligence pp.11–19
- Last, M. 2004. *Multi-objective Classification with Info-Fuzzy Networks*, ECML 2004, LNAI 3201, Springer-Verlag Berlin Heidelberg, pp.239–249
- Layton, M. and Gales, M.J.F. 2004. *Maximum Margin Training for Generative Kernel*, Cambridge University Engineering Department, Cambridge, UK. pp.1-21
- Lee, Jeonghwa, Chi-Hyuk Jun, 2013, PCA-based high-dimensional noisy data clustering via control of decision errors, *Knowledge Based System*, Volume 37, January 2013, pp.338-345.
- Lee, L.H. and Dino, I. 2010. Automatically computed document dependent weighting factor facility for Naïve Bayes classification. *Journal Expert Systems with Applications* pp.8471–8478.
- Leibovici, D.G., Bastin, L and Jackson, M. 2011. *Higher-order co-occurrences for exploratory point pattern analysis and decision tree clustering on spatial data*. Computers & Geosciences pp.382–389
- Li, X., Yunming, Y., Mark, J.L. and Michael K.Ng. 2010. *On cluster tree for nested and multi-density data clustering*, Pattern Recognition pp.3130–3143

- Li, Y., Ming, D. and Jing, H. 2008. *Localized feature selection for clustering*. Pattern Recognition Letters 29(1) pp.10-18
- Liang, Y., Klemen, A., 2006. *Statistical advances and challenges for analyzing correlated high dimensional SNP data in genomic study for complex diseases*. Statistics Survey, Volume 2, pp.43-60
- Liaw, Y.C., Maw, L.L. and Chien, M.W. 2010. Fast exact k nearest neighbors search using an orthogonal search tree, *Journal of Pattern Recognition*. pp.2351–2358.
- Lopez, M.I., Luna, J.M., Romero, C., 2012, Classification via clustering for predicting final marks based on student participation in forums 5th International Conference on Educational Data Mining, pp.148-151.
- Luan, J. 2002. *Data Mining and Its Applications in Higher Education*, New Directions For Institutional Research, No. 113, Spring Wiley Periodicals, Inc. Wilmington, DE.
- Nasiri, M., Minaci, B., Vafaei, F., 2012, Predicting GPA and Academic Dismissal in LMS Using Educational Data Mining: A Case Mining, *6th National and 3rd International Conference of e-Learning and e-Teaching*, pp.53-58
- Magidson, J., Vermunt, J., 2004. Latent class models. In *The SAGE Handbook of Quantitative Methodology for the Social Sciences*, D. Kaplan, Ed. Sage Publications, Thousand Oaks, CA, 175-198
- Magnani, M. and Danilo, M. 2004. A New Reparation Method for Incomplete Data in the Context of Supervised Learning, ITCC, *International Conference on Information Technology: Coding and Computing (ITCC'04)* Volume 1, pp.471-480
- Manning, Christopher D., Prabhakar Raghavan, Hinrich Schutze, 2009, *An Introduction to Information Retrieval*, Cambridge University Press Cambridge, England
- Maimon, O. and Lior, R. 2005. *Decomposition Methodology For Knowledge Discovery And Data Mining*”, World Scientific Publishing Co. Pte. Ltd., Singapore
- Matthew, B. 1998. Pattern discovery via entropy minimization, *MERL - A Mitsubishi Electric Research Laboratory*, <http://www.merl.com/papers/docs/TR98-21.pdf> (30 March 2013), pp.1-12
- Maulik, U. and Sanghamitra, B. 2002. *Performance Evaluation of Some Clustering Algorithms and Validity Indices*, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 12, December , pp.1650-1654

- May, R.J., Maier, H.R. and Dandy, G.C. 2010. Data splitting for artificial neural networks using SOM-based stratified sampling, *Journal Neural Network* Volume 23 Issue 2, March 2010, pp.283-294.
- McEntire, L.E., Lesley, R.D., Holly, K.O. and Michael, D.M. 2006. Innovations in job analysis : Development and application of metrics to analyze job data, *Human Resource Management Review*, Volume 16, pp.310-323.
- Menniti, D., Nadia, S. and Nicola, S. 2011. *Forecasting Next-Day Electricity Prices by a Neural Network Approach*, 8th International Conference on the European Energy Market (EEM), 2011, pp. 209-215.
- Mehrotra, K., Chilukuri, M. And Sanjay, R. 1996. *Elements of Artificial Neural Networks*. The MIT Press. Bradford PA
- Michalski, R.S., Mozetic,I., Hong,J., and Lavrac,N. 1986. The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains. *In Proceedings of the Fifth National Conference on Artificial Intelligence, Philadelphia, PA: Morgan Kaufmann*, pp.1041-1045
- Mihael, A., Markus, M.B., Kriegel, H.P. and Sander, J. 1999. OPTICS: Ordering Points To Identify the Clustering Structure. *Proc. ACM SIGMOD'99 International Conference on Management of Data, 1999*, pp.1-12.
- Mileta, D., Zdenko, S. and Minea, S. 2011. Forecasting prices of electricity on HUPX. *8th International Conference on the European Energy Market (EEM) Zagreb, Croatia*, pp.25-27
- Moise, G. and Sander, J. 2008. Finding non-redundant, statistically significant regions in high dimensional data: A Novel Approach To Projected And Subspace Clustering. *Journal Knowledge Data Discovery*, pp. 533-541
- Moise, G., Arthur, Z., Kröger, P. Kriegel, H.P. and Sander., J. 2009. *Subspace and projected clustering: experimental evaluation and analysis*. *Knowl Inf Syst* 21: pp.299–326
- Moser, C., Urban, D.M. and Weder, B. 2008. *International Competitiveness, Job Creation and Job Destruction - An Establishment Level Study of German Job Flows*", [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=1141651](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1141651), (21 January 2013)
- Mostow, J., Beck, J., 2006, *Some Useful Tactics To Modify, Map And Mine Data From Intelligent Tutors*, *Natural Language Engineering*, Vol. 12 No. 2, 195-208.

- Olivas, E.S., José, D.M.G., Marcelino, M.S., Jose, R.M.B. and Antonio, J.S.L. 2010. Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, *IGI Global*, Hershey, New York
- Orfanidis, P., and David, J.R. 2008, Preprocessing Enhancement to Improve data Mining Technique, *International Journal of Business Intelligence and Data Mining*, Vol. 3, No. 2, pp. 196-211
- Oswald, A. 2011. *Coping with new Challenges in Clustering and Biomedical Imaging*, Ph.D Dissertation, Ludwig Maximilians University at Munchen
- Owen, C. 1997. Design Research: Building the Knowledge Base. *Journal of the Japanese Society for the Science of Design* 5(2) pp.36-45.
- Pahl, C., Donnellan, C., 2003, Data Mining Technology for The Evaluation Of Web-Based Teaching And Learning Systems. *Congress E-Learning*. Montreal, Canada, pp. 1-7
- Papadopoulos, A. and Manolopoulos, Y. 1997. Performance of nearest neighbor queries in R-trees. In *Proceedings of International Conference on Database Theory (ICDT)*, Delphi, Greece, Jan. pp.394-408
- Parson, L., Ehtesham, H. and Huan, L. 2004 *Subspace Clustering for High Dimensional Data: A Review*. ACM SIGKDD Explorations, Volume 6, Issue 1, pp.92-103
- Pechenizkiy, M., Calders, T., Vasilyeva, E., De Bra, P., 2009, Mining the Student Assessment Data: Lessons Drawn from a Small Scale Case Study, *The 1st International Conference on Educational Data Mining*, Montréal, Québec, Canada, pp.12-18
- Phyu, T.N. 2009. Survey of Classification Techniques in Data Mining, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009*, Hong Kong, pp.18-22
- Plasse, M., Ndeye, N., Gilbert, S. And Alexandre, V. 2007. Combined Use of Association Rules Mining and Clustering Methods to Find Relevant Links Between Binary rare Attributes in a large data sets. *Computational Statistic and Data Analysis*, pp.1-18
- Pripužic K., Ivana, P.Ž. and Karl, A. 2011. Distributed processing of continuous sliding-window k-NN queries for data stream filtering, *Journal World Wide Web*, Springer, pp.465-494.

- Prasad, T. K, Srinivasa, R.O., Prasad,M.H.M.K., 2012. *Exploration of Meaningful Information from Educational Data Using Clustering and Sequential Pattern Miner* International Journal on Advanced Computer Theory and Engineering (IJACTE), Vol.1, Issue 1, pp.29-34
- Purnami, S.W., Jasni, M.Z. and Abdullah, E. 2010. *Data Mining Technique for Medical Diagnosis Using a New Smooth Support Vector Machine*, Networked Digital Technologies - Second International Conference, NDT, Prague, Czech Republic. Proceedings, Part II, pp.15-27
- Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.
- Rawat, M.S, 2009, *Launching Knowledge Management Project in Higher Education*, <http://www.kmice.uum.edu.my/kmice08/Paper/CR112.doc> (online) 11-May-2011.
- Richardson, S. 2010. Undergraduates perceptions of tourism and hospitality as a career choice, *International Journal of Hospitality and Tourism Management*, Volume: 17, pp.382-388.
- Romero, C., Ventura, S., 2007, Educational data mining: A survey from 1995 to 2005, *Expert Systems with Applications*, Vol. 33 No. 1, pp.135-146.
- Saha, I., Ujjwal, M., Sanghamitra, B. and Dariusz, P. 2012. SVMeFC: SVM Ensemble Fuzzy Clustering for Satellite Image Segmentation, *IEEE Geoscience And Remote Sensing Letters*, Vol. 9, No. 1, January, pp.52-55
- Sang, M.L., Abbot, A.L. and Philip A.A. 2007. Dimensionality Reduction and Clustering on Statistical Manifolds. *Proceeding Computer Vision and Pattern Recognition 2007*, pp.1-7
- Schölkopf, B., Alexander, S. and Klaus, R.M. 1998. Non Linear Kernel Principal Component Analysis, *Journal Vision And Learning*, Volume 10, pp.1299-1319.
- Sequeira, K. and Zaki, M. 2004. SCHISM: A new approach for interesting subspace mining. *ICDM*, pp.186-193
- Siddiqui , M.A., Shehab G.Din., 2013, Evaluation of Academic Plans of Study Using Data Mining Techniques, *IEEE 13th International Conference on Advanced Learning Technologies*, pp.224–228.
- Sikonja, M.R. and I. Kononenko, 2003, *Theoretical and Emphirical Analysis of Relief and RRelief*, Machine Learning, vol 53, pp.23–69.

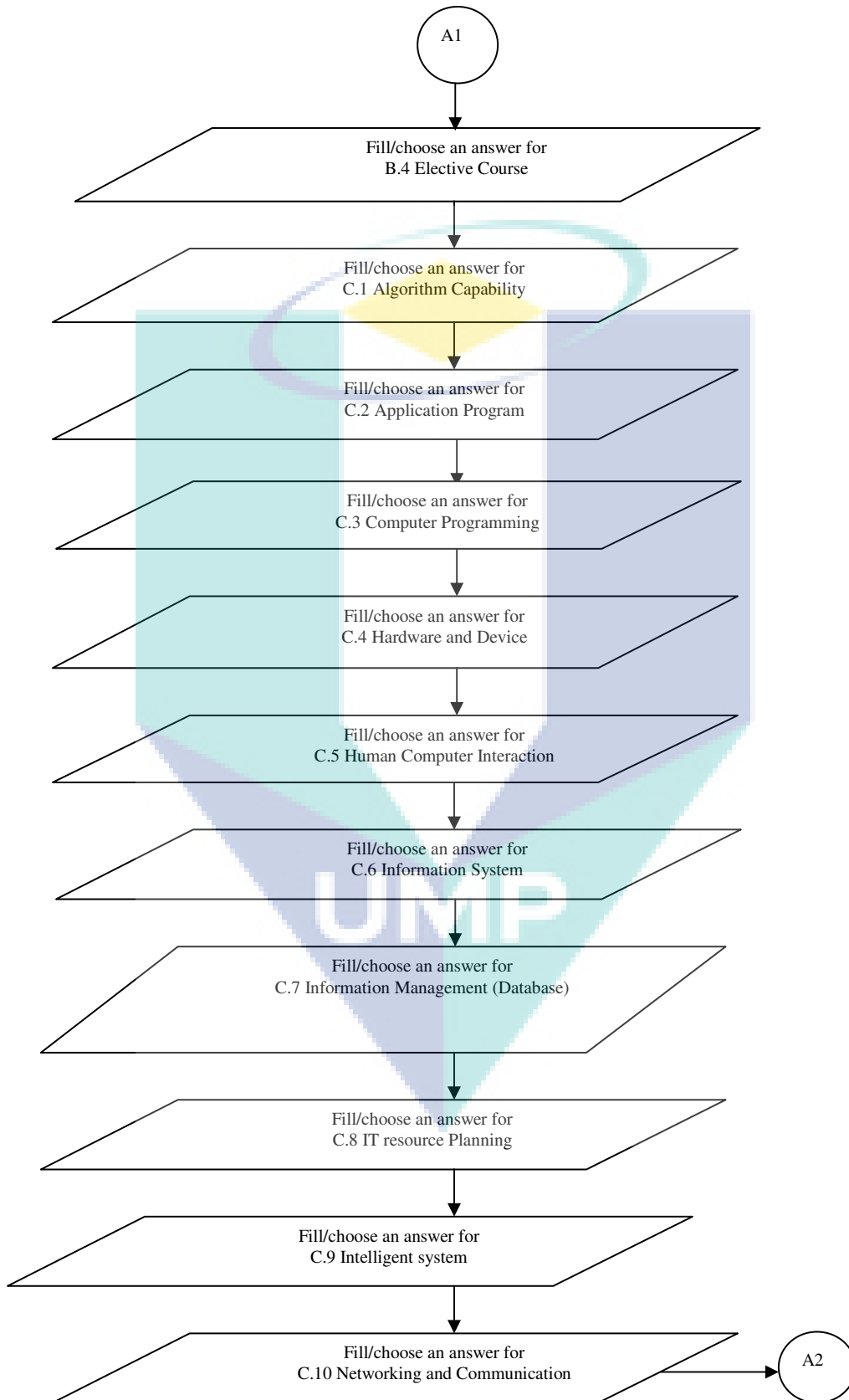
- Stearns, Jennifer C., Michael D. J. Lynch, Dilani B. Senadheera, Howard C. Tenenbaum, Michael B. Goldberg, Dennis G. Cvitkovitch, Kenneth Croitoru, Gabriel Moreno-Hagelsieb and Josh D. Neufeld, 2011, *Bacterial Biogeography Of The Human Digestive Tract*, Scientific Reports, Vol. 1 No. 170, pp.1-9
- Steinbach, M., Levent, E. and Vipin, K. 2003. *The Challenges of Clustering High Dimensional Data*, In *New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics, and Pattern Recognition*, pp.1-33
- Su, M.Y. 2011. Using clustering to improve the KNN-based classifiers for online anomalous network traffic identification, *Journal of Network and Computer Application*, Volume 34, pp.722-730.
- Tair, M.A., Alla, M.E., Mining Educational Data to Improve Students' Performance: A Case Study, *International Journal of Information and Communication Technology Research*, Vol. 2, No. 2, pp.140-146.
- Tan, M., and Eshelman, L. 1988. Using weighted networks to represent classification knowledge in noisy domains. *Proceedings of the Fifth International Conference on Machine Learning*, pp.121-134.
- Thakur, D., Nisarga, M. and Sharan, R.D. 2010. Re Optimization of ID3 and C4.5 Decision Tree. *Proceeding International Conference on Computer & Communication Technology 2010*, pp.448-450.
- Theodoridis, S. and Konstantinos, K. 2009, *Pattern Recognition*, Academic Press is an imprint of Elsevier, Burlington, MA USA
- Tong, S. and Edward, C. 2001. *Support Vector Machine Active Learning for Image Retrieval*, MM'01, Sept. 30-Oct. 5. Ottawa, Canada, pp.107-118
- Tomasev, N., Radovanovic, M., Mladenic, D., Ivanovic, M., 2013, The Role of Hubness in Clustering High-Dimensional Data, *IEEE Transactions on Knowledge and Data Engineering*, Revised January 2013, pp.1-14
- Trivedi, S., Zachary, P., Gábor, S., Heffernan, N., 2012, Spectral Clustering in Educational Data Mining, *4th International Conference on Educational Data Mining*, pp.129-138
- Vapnik, V. 1995. *The nature of statistical learning theory*. Springer, New York
- Visvanathan, M., Adagarla, B.S., Gerald, H and Lushington, P.S. 2009. Cluster Validation: *An Integrative Method for Cluster Analysis*, IEEE International Conference on Bioinformatics and Biomedicine Workshop, 2009, pp. 238 – 242.

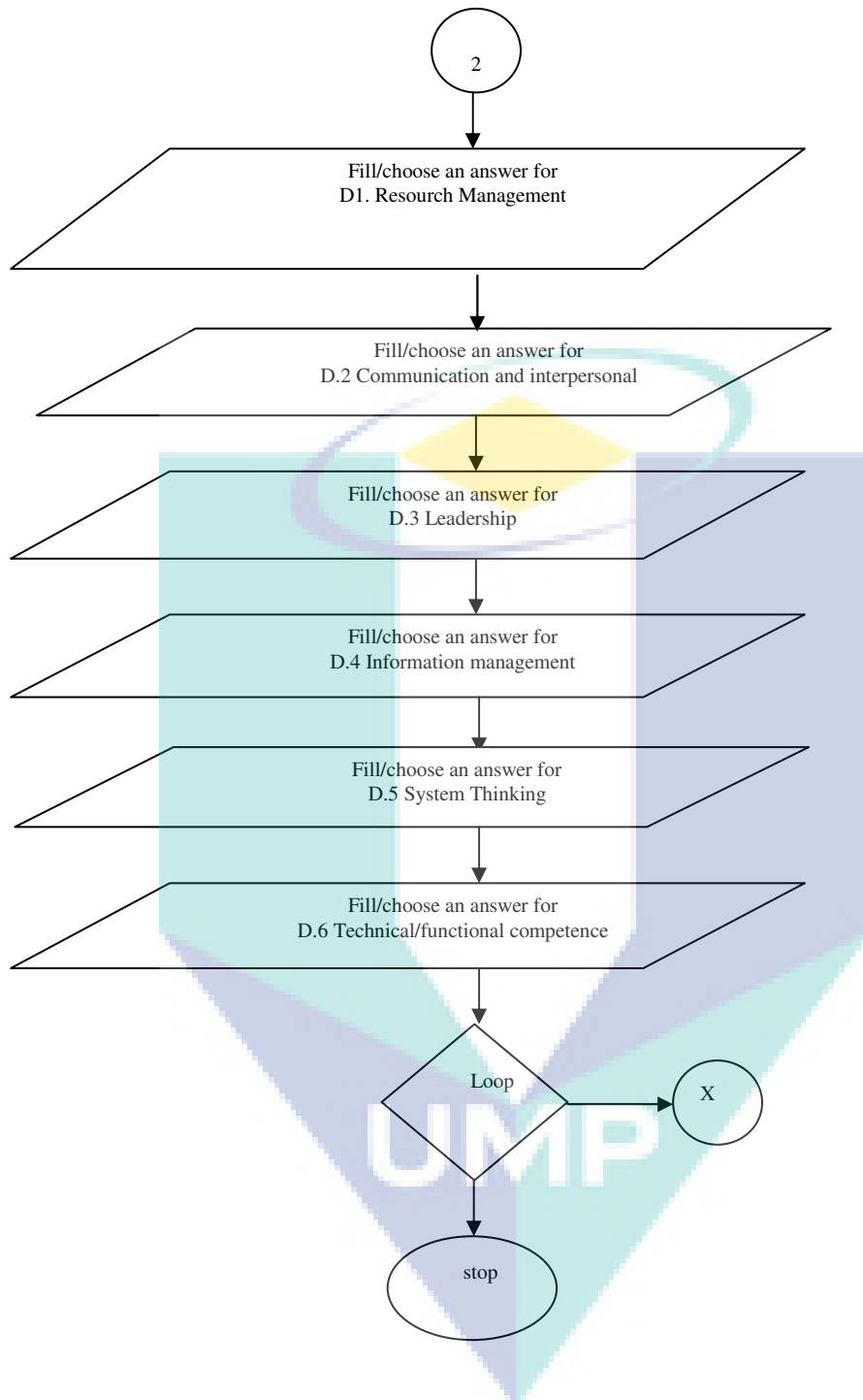


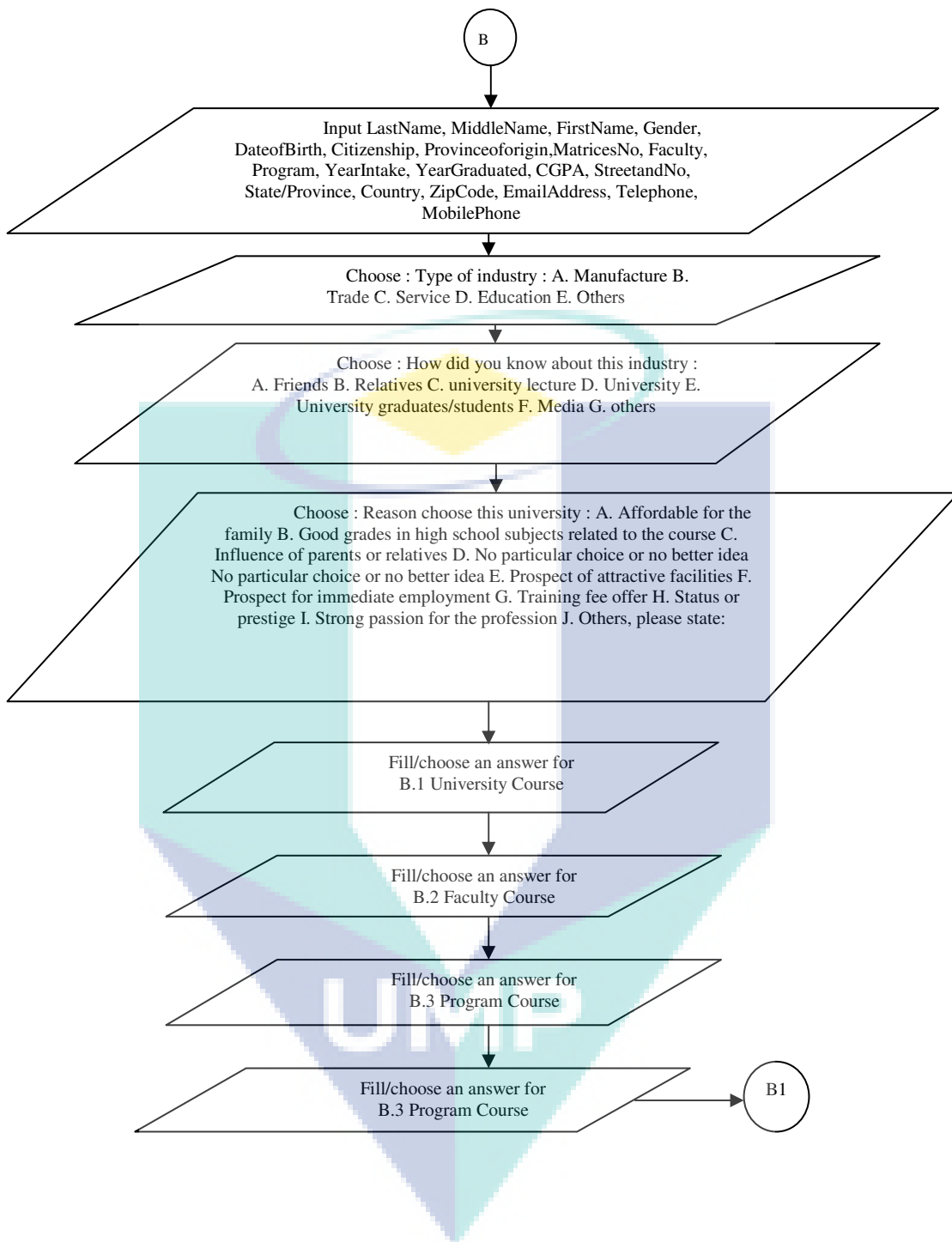
- Wai, R.J., Yi, C.C. and Yung, R.C. 2011. Short-Term Load Forecasting Via Fuzzy Neural Network With Varied Learning Rates. *Proceeding IEEE International Conference on Fuzzy Systems 2011*, pp.2426-2431
- Wang, C., Ya, Y.L., Yaoying, X. and Yan, W. 2007. Constructing the search for a job in academia from the perspective of self-regulated learning strategies and social cognitive career theory, *Journal of Vocational Behavior*, Volume 70, Issue 3, pp.574-589.
- Wang, G.A., Homa, A. and Hsinchun, C. 2011. A hierarchical Naïve Bayes model for approximate identity matching, *Journal of Decision Support Systems*, pp.413-423
- Wang, J. 2006, *Encyclopedia Of Data Warehousing And Data Mining*, Idea Group Reference, Hershey PA, pp.812-820
- Wang, X.J., 1999, *Data Mining and Knowledge Discovery for Process Monitoring and Control*, Springer, New York
- Weilin, L. and Lina, L. 2011. Neural network model for hydrological forecasting based on multivariate phase space reconstruction. *Seventh International Conference on Natural Computation. 2011*, pp.663-667
- Weston, J., and Herbrich, R. 2000. *Adaptive margin support vector machines*. In A. Smola, P. Bartlett, B. Scholkopf, & D. Schuurmans (Eds.), *Advances in large margin classifiers* Cambridge, UK MA: MIT Press. pp.281–295
- Witten, I.H. and Eibe, F. 2005. *Data Mining—Practical Machine Learning Tools and Technique*, 2nd edn, Morhan Kaufmann, San Fransisco
- Xie, N., Lin, C. and Aiping, L. 2009, Fault Diagnosis of Multistage Manufacturing System based on Rough Set Approach, *International Journal Advance Manufacturing Technology*, pp. 1-9
- Xu, B., Recker, M., Qi, X, Flann, N., Ye, L., 2009. *Journal of Educational Data Mining*, Volume 5, No 2, August, 2013
- Xu, R., Donald C.W.II. 2009. *Clustering*. New Jersey: John Wiley & Sons. Inc. pp.237-239
- Xulei, Y., Qing, S. and Aize, C. 2006. A New Cluster Validity For Data Clustering. *Neural Processing Letters* 23, pp.325–344
- Yang, L. and Teng, Z. 2011. *Prediction of Grain Yield Based on Spiking Neural Networks Model* “Beijing, IEEE Explorer. pp.171-174

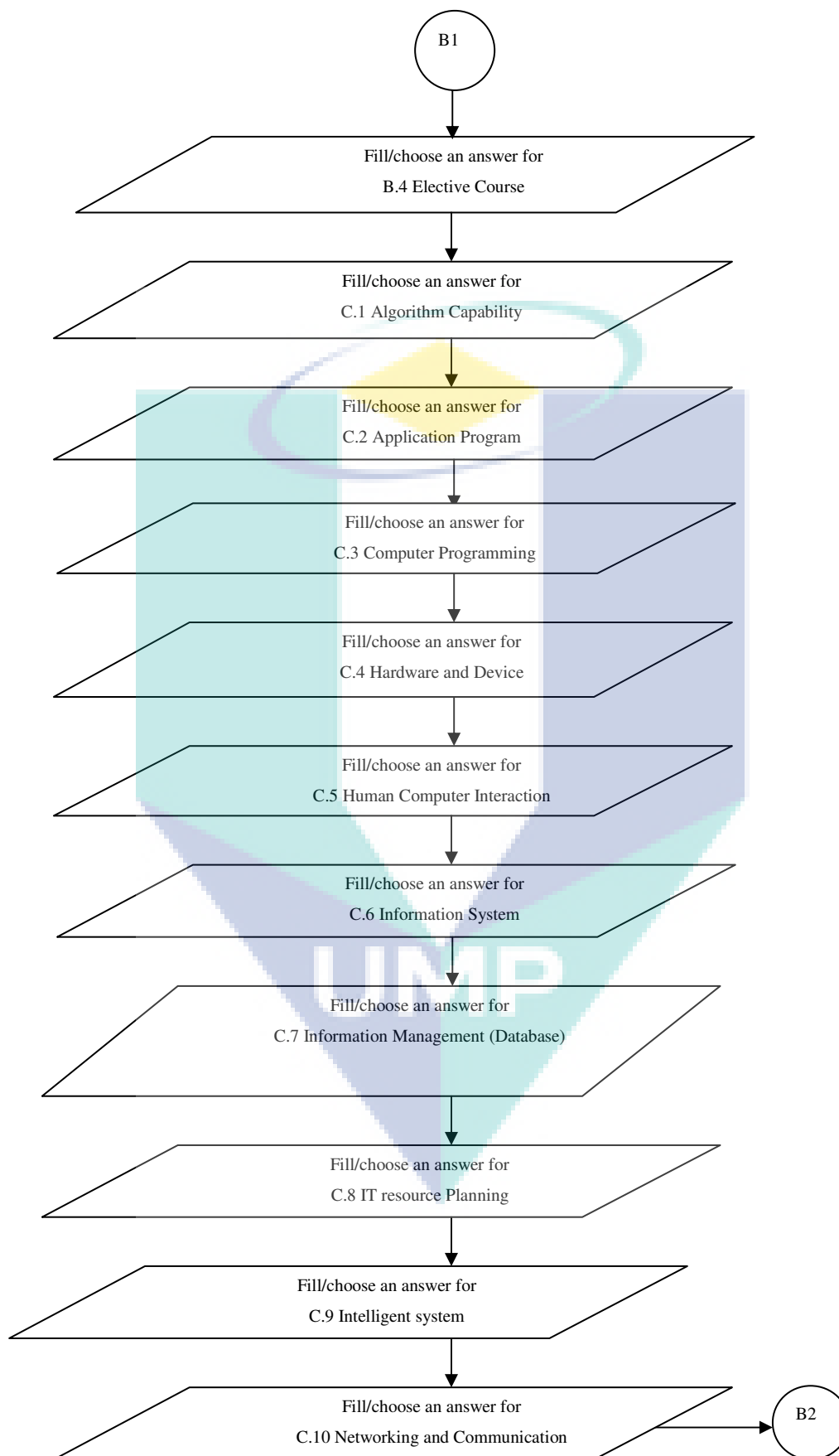
- Yang, Q. and Xindong, W. 2006. 10 Challenging Problem in Data mining Research. *International Journal of Information Technology & Decision Making*, Vol. 5, No. 4, pp.597–604.
- Yu, F., Zheming, L., Hao, L. and Pinghui, W. 2010. *Three-Dimensional Model Analysis and Processing*, Springer Heidelberg Dordrecht, New York
- Yucheng, L., Liu Y, 2010. Incremental Learning Method of Least Squares Support Vector Machine. *International Conference on Intelligent Computation Technology and Automation*, pp.529-532.
- Yun, J., Liping, J., Jian, Y. and Houkuan, H. 2012. *A multi-layer text classification framework based on two-level representation model*, *Expert Systems with Applications* 39 pp.2035–2046
- Yusof, Azwina M. and Rukaini Abdullah, 2005, *The Evolution of Programming Courses: Course curriculum, students, and their performance*, *ACM SIGCSE Bulletin archive*. Volume 37, Issue 4, pp.74 – 78
- Zang, X. and Jianli, Y. 2011. The Prediction of Index in Shanghai Stock Based on Genetic Neural Network., *Proceeding Second International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011*, pp. 5943 – 5946.
- Zhang, D. and Jeffrey J.P.T. 2005. Machine Learning Application in Software Engineering, *World Scientific Publishing Co. Pte. Ltd.*
- Zhang, D., Hua, Z.Z. and Songcan, C. 2008. Semi-Supervised Dimensionality Reduction, *7th SIAM International Conference on Data Mining*, pp.629-634.
- Zhao, Y., Chengqi, Z. And Longbing, C. 2009. Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction, *IGI Global*, Hershey PA.

## APPENDIXES









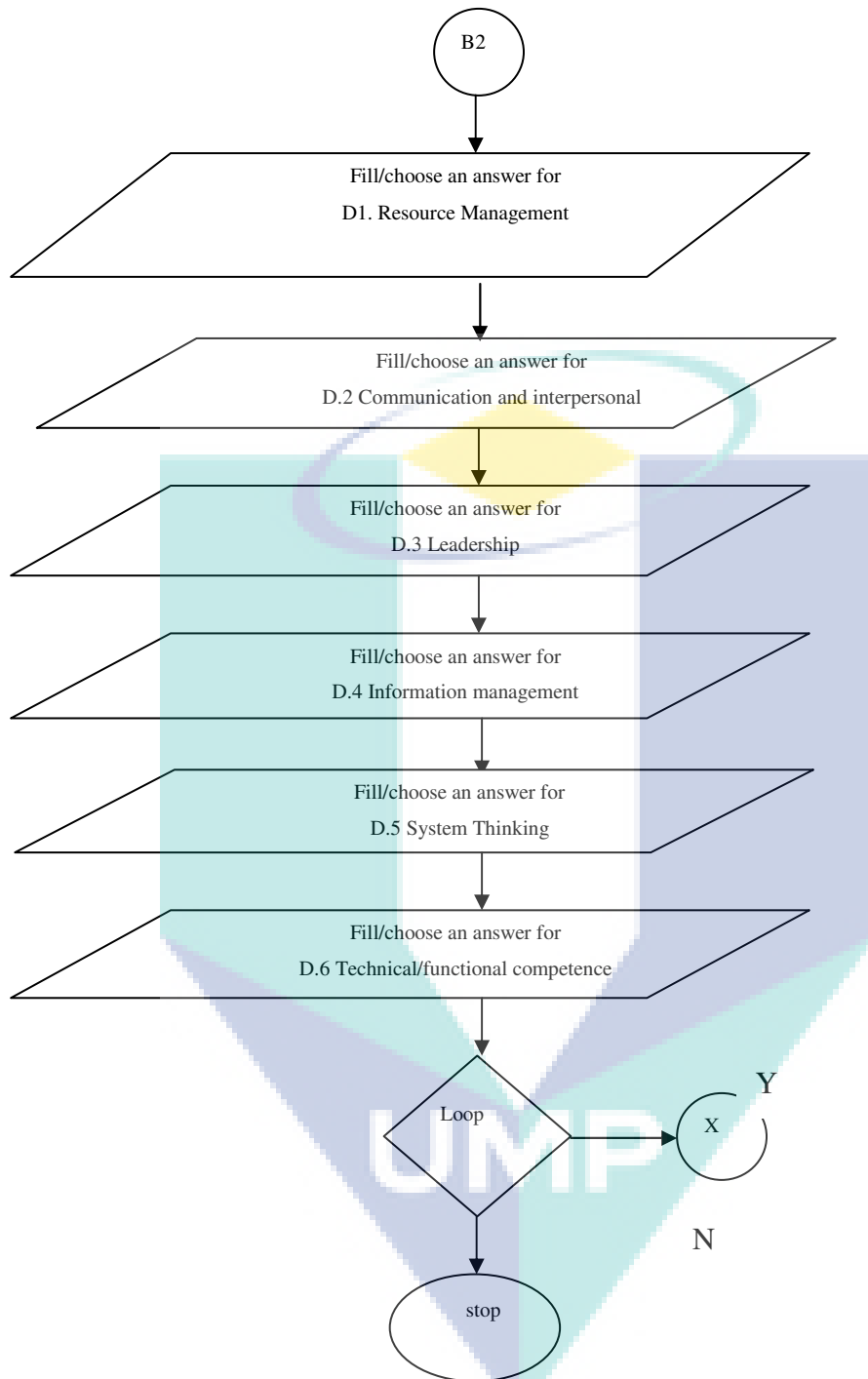


Figure 7.1. Flowchart of online questionnaire

## LIST OF PUBLICATION

## JOURNALS

- Sembiring, R.W., Jasni M.Z., Abdullah E., 2010, *Clustering High Dimensional Data Using Subspace And Projected Clustering Algorithms*, International Journal Of Computer Science & Information Technology (IJCSIT) Vol.2, No.4, August 2010
- Sembiring, R.W., Jasni M.Z., 2010, *Cluster Evaluation Of Density Based Subspace Clustering*, Journal Of Computing, Volume 2, Issue 11, November 2010
- Sembiring, R.W., Jasni M.Z., Abdullah E., 2010, *A Comparative Agglomerative Hierarchical Clustering Method To Cluster Implemented Course*, Journal Of Computing, Volume 2, Issue 12, December 2010
- Sembiring, R.W., Jasni M.Z., 2011, *The Design Of Pre-Processing Multidimensional Data Based On Component Analysis*, Computer And Information Science, Vol. 4, No. 3, May 2011
- Sembiring, R.W., Jasni M.Z., Abdullah E. 2011, *Dimension Reduction Of Health Data Clustering*, International Journal On New Computer Architectures And Their Applications (IJNCAA) 1(3)

## CONFERENCES

- Sembiring, R.W., 2010, *Evaluasi Klaster Pada Data Multidimensi Melalui Pendekatan Berbasis Densiti*, SNIKOM 2010, Indonesia
- Sembiring, R.W., Jasni M.Z., 2011, *Rancangan Pre-Processing Data Multidimensi Berdasarkan Analisa Komponen*, KNSI 2011, Indonesia
- Sembiring, R.W., Jasni M.Z., Abdullah E., 2011, *Alternative Model For Extracting Multidimensional Data Based-On Comparative Dimension Reduction*, ICSECS 2011, Malaysia
- Sembiring, R.W., Jasni M.Z., 2011, 2012, *Klaster Sub-Ruang Berdasarkan Kerapatan Data*, SNASTIKOM 2012, Indonesia