

**TREE-BASED CONTRAST SUBSPACE MINING  
METHOD**

**FLORENCE SIA FUI SZE**



**UMS**  
UNIVERSITI MALAYSIA SABAH

**FACULTY OF COMPUTING AND INFORMATICS  
UNIVERSITI MALAYSIA SABAH  
2020**

**TREE-BASED CONTRAST SUBSPACE MINING  
METHOD**

**FLORENCE SIA FUI SZE**



**UMS**

**THESIS SUBMITTED IN FULLFILMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF DOCTOR  
OF PHILOSOPHY**

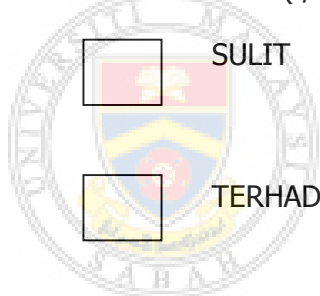
**FACULTY OF COMPUTING AND INFORMATICS  
UNIVERSITI MALAYSIA SABAH  
2020**

**UNIVERSITI MALAYSIA SABAH**  
BORANG PENGESAHAN STATUS TESIS

JUDUL : **TREE-BASED CONTRAST SUBSPACE MINING METHOD**  
IJAZAH : **DOKTOR FALSAFAH**  
BIDANG : **SAINS KOMPUTER**

Saya **FLORENCE SIA FUI SZE**, Sesi **2017-2020**, mengaku membenarkan tesis Doktor ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Tesis ini adalah hak milik Universiti Malaysia Sabah
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan ( / ):



SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

Disahkan Oleh,

---

**FLORENCE SIA FUI SZE**  
**DI1621004T**

---

(Tandatangan Pustakawan)

Tarikh : 30 June 2020

---

(Assoc. Prof. Dr. Rayner Alfred)  
Penyelia

## DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, excerpts, equations, summaries and references, which have been duly acknowledged.

17 March 2020

---

Florence Sia Fui Sze  
DI1621004T



UMS  
UNIVERSITI MALAYSIA SABAH

## CERTIFICATION

Name : **FLORENCE SIA FUI SZE**  
Matric No. : **DI1621004T**  
Title : **TREE-BASED CONTRAST SUBSPACE  
MINING METHOD**  
Degree : **DOCTOR OF PHILOSOPHY**  
Field : **COMPUTER SCIENCE**  
Date of Viva : **17 MARCH 2020**

**CERTIFIED BY;**



**SUPERVISOR**

Assoc. Prof. Dr. Rayner Alfred

Signature

**UMMS**  
UNIVERSITI MALAYSIA SABAH

## ACKNOWLEDGEMENT

I hereby would like to express my sincere gratitude to all those who have been supported me in completing this thesis. First and foremost, I would like to express my deepest appreciation to my supervisor, Assoc. Prof. Dr. Rayner Alfred for his guidance, patience, and knowledge. He helped a lot all the time when needed and gave the right direction towards completion of my research. His continuous support gave me strength in overcoming obstacles I have encountered during my research study and thesis writing.

Besides, I would like to thank to all laboratory assistants, Mrs. Diana, Mr. Fadzli, and Mr. Albert who have helped me with the laboratory computers for conducting experiments of my research.

I would like to extend my sincere thanks to all my friends and lab mates for their moral supports.

I would also like to take this opportunity to acknowledge the contribution from Universiti Malaysia Sabah (UMS) and Kementerian Pengajian Tinggi (KPT) Malaysia for granting a scholarship to me. With this scholarship, I am able to finish my research study without worrying about the financial burden.

Last but not least, I would like to express my heartfelt thanks to my parents, my late father Mr. Sia King Hai and my mother Mrs. Samilin Binti Miki, for their love, encouragement, moral and spiritual supports along the way to completing this research and thesis. I love you Papa. I would like also to thank my brother, Mr. Hendry Sia.

Florence Sia  
17 March 2020

## ABSTRACT

Mining contrast subspace finds subsets of features or subspaces where a query object is most likely similar to target class against other class in a multidimensional data set of two classes. Those subspaces are termed as contrast subspaces. All existing mining contrast subspace methods (i.e. CSMiner and CSMiner-BPR) use density-based likelihood contrast scoring function to estimate the likelihood of a query object to target class against other class in a subspace. Query object resides in the area that has high ratio of probability density of target class to probability density of other class with respect to query object in a contrast subspace. However, the probability density estimation of a class requires adjustment to the dimensionality or number of features in subspaces which may affect the performance of mining contrast subspace. Besides, the parameter setting and the subspace search strategy of all existing methods are not being optimized to mine contrast subspace. They also cannot be directly applied to mine contrast subspaces in categorical data. In this thesis, a novel tree-based contrast subspace mining method is introduced which employs tree-based likelihood contrast scoring function that is not affected by the dimensionality of subspaces. Tree-based likelihood contrast scoring function recursively partitions a subspace space in the way that query object fall in a group that has high ratio of probability of target class and probability of other class in a contrast subspace. The tree-based method begins with feature selection phase which finds relevant features and followed by contrast subspace search phase to search contrast subspaces from the relevant features, accordance to the tree-based likelihood contrast scoring function. Genetic algorithm has been widely used to find global solution to optimization and search problem. Hence, this thesis presents the optimization of parameters values for the tree-based method by genetic algorithm. This thesis also presents the optimization of contrast subspace search of the tree-based method by genetic algorithm. In addition, the tree-based method is extended to mine contrast subspaces of query object in categorical data. The research works involve first preparing the real world numerical and categorical data sets. Then, the tree-based method, the genetic algorithm based parameter values identification of tree-based method, and followed by the genetic algorithm based tree-based method, for numerical data sets are developed and evaluated. Lastly, the extended tree-based method for categorical data sets is developed and evaluated. The effectiveness of the tree-based method in mining contrast subspace is evaluated by the classification accuracy on the obtained contrast subspaces with respect to query object. The empirical results demonstrated that the tree-based method is capable to find relevant contrast subspace of the given query object while the tree-based method with the optimized parameter setting is the best for mining contrast subspace in numerical data. Furthermore, the results exhibited that the extended tree-based method is capable to find contrast subspace of query object in categorical data.

## **ABSTRAK**

### **KAEDAH TREE-BASED CONTRAST SUBSPACE MINING**

*Mining contrast subspace mencari subset-subset atribut atau subruang di mana objek pertanyaan adalah sama dengan kelas sasaran tetapi berbeza daripada kelas lain dalam data multidimensi dua kelas. Subruang tersebut dikenali sebagai subruang kontras. Semua kaedah-kaedah mining contrast subspace yang sedia ada (iaitu CSMiner dan CSMiner-BPR) menggunakan likelihood contrast scoring function berdasarkan kepadatan untuk menganggar persamaan objek pertanyaan dengan kelas sasaran serta perbezaan dengan kelas lain pada suatu subruang. Objek pertanyaan berada dalam kelompok yang mempunyai nisbah kepadatan kebarangkalian kelas sasaran kepada kepadatan kebarangkalian kelas lain yang tinggi pada suatu subruang kontras. Walau bagaimanapun, anggaran kepadatan kebarangkalian kelas memerlukan pelarasan berdasarkan bilangan atribut dalam subruang untuk mengelakkan penurunan kepadatan dengan penambahan bilangan atribut dalam subruang. Di samping itu, nilai parameter dan strategi pencarian subspace semua kaedah yang sedia ada adalah tidak dioptimumkan untuk mencari subruang kontras. Kaedah-kaedah yang sedia ada juga tidak dapat digunakan secara langsung untuk mencari subruang kontras bagi object pertanyaan dalam data kategori. Dalam tesis ini, kaedah baru tree-based contrast subspace mining diperkenalkan yang menggunakan tree-based likelihood contrast scoring function yang tidak terjejas oleh bilangan atribut dalam subruang, maka dengan itu tidak memerlukan sebarang pelarasan. Tree-based likelihood contrast scoring function membahagi data pada subruang secara berulang di mana objek pertanyaan dikumpulkan dengan objek yang mempunyai ciri yang sama. Kaedah tree-based bermula dengan fasa pemilihan atribut yang mencari atribut yang relevan berdasarkan tree-based likelihood contrast scoring function dan diikuti dengan fasa pencarian subruang kontras yang mencari subruang kontras dari atribut yang relevan berdasarkan tree-based likelihood contrast scoring function. Algoritma genetik telah digunakan secara meluas untuk mencari penyelesaian optimum kepada masalah pengoptimuman dan pencarian. Dengan itu, tesis ini membentangkan pengoptimuman nilai parameter terbaik untuk kaedah tree-based dengan menggunakan algoritma genetik. Tesis ini membentangkan pengoptimuman pencarian subruang kontras dengan menggunakan algoritma genetik. Seterusnya, kaedah tree-based diperluas untuk mencari subruang kontras dalam data kategori. Keberkesanan kaedah tree-based dinilai dari segi ketepatan klasifikasi pada subruang kontras yang diperolehi. Kajian ini bermula dengan menyediakan data berangka dan kategori. Selepas itu, kaedah tree-based, kaedah pencarian nilai parameter berdasarkan algoritma genetik, dan kaedah tree-based berdasarkan algoritma genetik untuk data berangka dibina dan dinilai. Akhir sekali, kaedah tree-based untuk data kategori dibina dan dinilai. Keberkesanan kaedah tree-based akan dinilai berdasarkan ketepatan klasifikasi pada subruang kontras yang diperolehi. Keputusan empirikal menunjukkan kaedah tree-based mampu mencari subruang kontras yang relevan bagi objek pertanyaan dan kaedah tree-based dengan tetapan parameter yang telah dioptimumkan adalah terbaik untuk pencarian subruang kontras dalam data berangka. Selain itu, keputusan empirikal menunjukkan kaedah tree-based yang diperluas tersebut mampu mencari contrast subspaces bagi objek pertanyaan dalam data kategori.*



# LIST OF CONTENTS

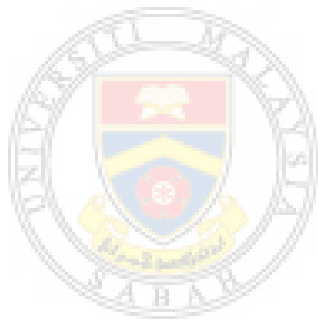
	Page
<b>TITLE</b>	i
<b>DECLARATION</b>	ii
<b>CERTIFICATION</b>	iii
<b>ACKNOWLEDGEMENT</b>	iv
<b>ABSTRACT</b>	v
<b><i>ABSTRAK</i></b>	vi
<b>LIST OF CONTENTS</b>	vii
<b>LIST OF TABLES</b>	xii
<b>LIST OF FIGURES</b>	xv
<b>LIST OF ABBREVIATIONS</b>	xvii
<b>LIST OF SYMBOLS</b>	xviii
<b>LIST OF APPENDICES</b>	xx
<b>CHAPTER 1: INTRODUCTION</b>	1
1.1 Background	1
1.2 Problem Statement	4
1.3 Operational Definition	5
1.4 Research Questions	6
1.5 Research Objectives	7
1.6 Research Scopes	7
1.7 Research Contributions	10
1.8 Thesis Organization	11

<b>CHAPTER 2: LITERATURE REVIEW</b>	13
2.1 Introduction	13
2.2 Data Mining	13
2.2.1 Classification	14
a. Classification Evaluation Metric	18
2.2.2 Outlier Detection	18
2.3 Finding Subspaces for Explaining Outlyingness of Object	21
2.4 Mining Contrast Subspace	23
2.4.1 Contrast Subspace Miner	24
2.4.2 Contrast Subspace Miner Bounding-Pruning-Refining	25
2.5 Approaches for Handling Dimensionality Biased Scoring Function	26
2.6 Subspace Search Approaches	29
2.6.1 Feature Selection	29
2.6.2 Evolutionary Algorithm	32
2.7 Conclusion	35
<b>CHAPTER 3: METHODOLOGY</b>	37
3.1 Introduction	37
3.2 Research Methodology	38
3.3 Step 1: Data Preparation	39
3.3.1 Numerical Data Sets	40
3.3.2 Categorical Data Sets	41
3.4 Step 2: Tree-Based Contrast Subspace Mining Method Development and Evaluation	42
3.5 Step 3: Genetic Algorithm based Parameters Values Identification for Tree-Based Contrast Subspace Mining Development and Evaluation	49
3.6 Step 4: Genetic Tree-Based Contrast Subspace Mining Method Development and Evaluation	51
3.7 Step 5: Extended Tree-Based Contrast Subspace Mining Method Development and Evaluation	53
3.8 Conclusion	56

<b>CHAPTER 4: TREE-BASED CONTRAST SUBSPACE MINING METHOD</b>	<b>58</b>
4.1 Introduction	58
4.2 Tree-Based Contrast Subspace Mining Method	59
4.2.1 Tree-Based Likelihood Contrast Scoring Function	59
4.3 Framework of Tree-Based Contrast Subspace Mining Method	65
4.3.1 Feature Selection Phase	65
4.3.2 Contrast Subspace Search Phase	68
4.3.3 Contrast Subspace Accuracy Evaluation Metric	71
4.4 Parameter Setting	72
4.4.1 Parameter Setting for Minimum Number of Objects <i>MinObjs</i> and Small Constant Value $\epsilon$	73
4.4.2 Parameter Setting for Number of Relevant Features <i>l</i> and Number of Random Subspaces <i>t</i>	91
4.5 Effect of <i>k</i> Values on the Contrast Subspace Accuracy	105
4.6 Sensitivity Analysis of Tree-Based Likelihood Scoring Function	111
4.7 Parameter Sensitivity Analysis of Density-Based Mining Contrast Subspace Method	117
4.7.1 Effect of $\nu$ Value on Contrast Subspace Accuracy	117
4.7.2 Effect of <i>k</i> Value on Contrast Subspace Accuracy	120
4.8 Experimental Setup and Analysis on Tree-based Contrast Subspace Mining Method	125
4.9 T-Test	129
4.10 Conclusion	131
<b>CHAPTER 5: OPTIMIZING PARAMETERS VALUES OF TREE-BASED MINING CONTRAST SUBSPACE METHOD USING GENETIC ALGORITHM</b>	<b>133</b>
5.1 Introduction	133
5.2 Genetic Algorithm Based Parameters Values Identification Method	134
5.2.1 Chromosome Representation	136
5.2.2 Initial Population	137
5.2.3 Fitness Evaluation	137
5.2.4 Parents Selection	137
5.2.5 Crossover	139
5.2.6 Mutation	140

5.3	Experimental Setup and Analysis	141
5.4	T-Test	153
5.5	Conclusion	156
<b>CHAPTER 6: OPTIMIZING TREE-BASED CONTRAST SUBSPACE MINING USING GENETIC ALGORITHM</b>		<b>158</b>
6.1	Introduction	158
6.2	Genetic Tree-Based Contrast Subspace Mining Method	159
6.2.1	Chromosome Representation	161
6.2.2	Initial Population	161
6.2.3	Fitness Evaluation	161
6.2.4	Selection	162
6.2.5	Crossover	163
6.2.6	Mutation	165
6.3	Experimental Setup and Analysis	165
6.4	T-test	169
6.5	Conclusion	172
<b>CHAPTER 7: EXTENDED TREE-BASED CONTRAST SUBSPACE MINING METHOD FOR CATEGORICAL DATA</b>		<b>173</b>
7.1	Introduction	173
7.2	Tree-Based Likelihood Contrast Scoring Function for Categorical Data	174
7.3	The Framework of Extended Tree-Based Contrast Subspace Mining Method	179
7.4	Experimental Setup and Analysis	182
7.5	T-Test	186
7.6	Conclusion	188
<b>CHAPTER 8: CONCLUSION AND FUTURE WORKS RECOMMENDATION</b>		<b>190</b>
8.1	Summary of Research Works	190
8.2	Future Works	193

8.3	Conclusion	195
	<b>REFERENCES</b>	197
	<b>APPENDICES</b>	229



**UMS**  
UNIVERSITI MALAYSIA SABAH

## LIST OF TABLES

		Page
Table 3.1:	The Characteristics of the Chosen Numerical Data Sets	40
Table 3.2:	The Characteristics of the Chosen Categorical Data Sets	42
Table 3.3:	Test Score for A Sample of Students	47
Table 4.1:	Filter Feature Selection Algorithm of Tree-Based Contrast Subspace Mining Method	68
Table 4.2:	Contrast Subspace Search Algorithm of Tree-Based Contrast Subspace Mining Method	70
Table 4.3:	Average Classification Accuracy (%) of J48 on (a)BCW and PID, (b)Wine and Glass, (c)CMSC and Wave	76
Table 4.4:	Average Classification Accuracy (%) of NB on (a)BCW and PID, (b)Wine and Glass, (c)CMSC and Wave	79
Table 4.5:	Average Classification Accuracy (%) of SVM on (a)BCW and PID, (b)Wine and Glass, (c)CMSC and Wave	82
Table 4.6:	Average Classification Accuracy (%) of RF on (a)BCW and PID, (b)Wine and Glass, (c)CMSC and Wave	86
Table 4.7:	Average Classification Accuracy (%) of J48 on (a)BCW, PID, Wine, and Glass, (b)CMSC and Wave	94
Table 4.8:	Average Classification Accuracy (%) of NB on (a)BCW, PID, Wine, and Glass, (b)CMSC and Wave	96
Table 4.9:	Average Classification Accuracy (%) of SVM on (a)BCW, PID, Wine, and Glass, (b)CMSC and Wave	99
Table 4.10:	Average Classification Accuracy (%) of RF on (a)BCW, PID, Wine, and Glass, (b)CMSC and Wave	102
Table 4.11:	Average Tree-Based and Density-Based Likelihood Scores of Target Class for Subspace Dimensionality $d$ on BCW, PID, Wine, and Glass	113
Table 4.12:	Average Tree-Based and Density-Based Likelihood Scores of Other Class for Subspace Dimensionality $d$ on BCW, PID, Wine, and Glass	114

Table 4.13:	Average Classification Accuracy (%) of J48, NB, SVM, and RF based on 5-fold, 10-fold, and 20-fold CV for BCW, PID, Wine, Glass, CMSC, and Wave	119
Table 4.14:	Average Classification Accuracy (%) on Contrast Subspace from Tree-Based and Density-Based Contrast Subspace Mining Methods	127
Table 4.15:	Paired T-Test for (a)J48, (b)NB, (c)SVM, and (d)RF, on BCW, PID, Wine, Glass, CMSC, and Wave Data Sets	129
Table 5.1:	An Example of Selection Probabilities of Chromosomes <i>Chr1-Chr5</i>	138
Table 5.2:	The Selected Chromosomes Based on Random Integers $r$	138
Table 5.3:	The Best Sets of Parameters Values Based on J48, NB, SVM, and RF for (a)BCW, (b)PID, (c)Wine, (d)Glass, (e)CMSC and (f)Wave	144
Table 5.4:	Summary of Best Set of Parameters Values of Tree-Based Contrast Subspace Mining Method for BCW, PID, Wine, Glass, CMSC, and Wave	151
Table 5.5:	Average Classification Accuracy (%) of J48, NB, SVM, and RF on BCW, PID, Wine, Glass, CMSC, and Wave	152
Table 5.6:	Paired T-Test on Average Classification Accuracy (%) for J48, NB, SVM, and RF, on (a)BCW, (b)PID, (c)Wine, and (d)Glass, (e)CMSC, (f)Wave Data Sets	154
Table 6.1:	An Example of Fitness Scores and Selection Probabilities of Chromosomes <i>Chr1-Chr5</i>	163
Table 6.2:	The Selected Chromosomes Based on Random Integers $r$	163
Table 6.3:	The <i>Minobjs</i> , $\varepsilon$ , and $l$ for BCW, PID, Wine, Glass, CMSC, and Wave Data Sets	166
Table 6.4:	Average Classification Accuracy (%) for Genetic and Tree-Based Baseline Contrast Subspace Mining Methods	168
Table 6.5:	Paired T-Test on Average Classification Accuracy (%) for J48, NB, SVM, J48, and RF, on (a)BCW, (b)PID, (c)Wine, (d)Glass, and (e)CMSC, (f)Wave Data Sets	170
Table 7.1:	Filter Feature Selection Algorithm of Extended Tree-Based Contrast Subspace Mining Method	180

Table 7.2:	Contrast Subspace Search Algorithm of Extended Tree-Based Contrast Subspace Mining Method	182
Table 7.3:	Average Classification Accuracy (%) of J48, NB, SVM, and RF for Baseline Classification and Extended Tree-Based Method on BC, Votes, TTT, Lymphography, Mushroom, and Chess	185
Table 7.4:	Paired T-Test on Average Classification Accuracy (%) for J48, NB, SVM, and RF, on (a)BC, (b)Votes, (c)TTT, (d)Lymphography, (e)Mushroom, and (f)Chess Data Sets	187



UMS  
UNIVERSITI MALAYSIA SABAH



## LIST OF FIGURES

	Page
Figure 3.1: Research Methodology of Tree-Based Mining Contrast Subspace	38
Figure 3.2: An Example of Decision Tree for Numerical Data Set	43
Figure 3.3: An Example of Decision Tree for Categorical Data Set	54
Figure 4.1: Half Binary Tree Construction for Subspace $S=\{f_1, f_2\}$ Using Tree-Based Likelihood Contrast Scoring Function	61
Figure 4.2: Half Binary Tree Construction For Subspace $S=\{f_1, f_2, f_3\}$ Using Tree-Based Likelihood Contrast Scoring Function	62
Figure 4.3: Framework of Tree-Based Contrast Subspace Mining Method	66
Figure 4.4: Filter Feature Selection Process of Tree-Based Contrast Subspace Mining Method	66
Figure 4.5: Contrast Subspace Search Process of Tree-based Contrast Subspace Mining Method	69
Figure 4.6: Average Classification Accuracy (%) of J48, NB, SVM, and RF for $k$ values on (a)BCW, (b)PID, (c)Wine, (d)Glass, (e)CMSC, and (f)Wave	109
Figure 4.7: Average Classification Accuracy (%) of J48, NB, SVM, and RF on Contrast Subspaces From The Density-Based Mining Contrast Subspace Method for (a)BCW, (b)PID, (c)Wine, (d)Glass,(e)CMSC, and (f)Wave	123
Figure 5.1: The Framework of Genetic Algorithm Based Parameters Values Identification Method for Tree-Based Contrast Subspace Mining Method	135
Figure 5.2: An Example of Chromosome Representation	136
Figure 5.3: The Cumulative Probabilities of Chromosomes <i>Chr1-Chr5</i>	138
Figure 5.4: One-Point Crossover Operation	139
Figure 5.5: An Example of Mutation Operation	140

Figure 6.1:	The Framework of Genetic Tree-Based Contrast Subspace Mining Method	160
Figure 6.2:	An Example of Chromosome Representation for Subset of Features $S=\{f_2, f_3, f_4, f_5\}$	161
Figure 6.3:	The Cumulative Probabilities of Chromosomes <i>Chr1-Chr5</i>	163
Figure 6.4:	One-Point Crossover Operation	164
Figure 6.5:	An Example of Mutation Operation	165
Figure 7.1:	Half Binary Tree Construction for Subspace $S=\{f_1, f_2\}$ using the Tree-Based Likelihood Contrast Scoring Function for Categorical Data	175
Figure 7.2:	Half Binary Tree Construction for Subspace $S=\{f_1, f_3, f_4\}$ using the Tree-Based Likelihood Contrast Scoring Function for Categorical Data	176



UMS  
UNIVERSITI MALAYSIA SABAH

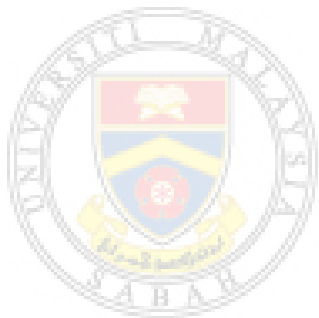
## LIST OF ABBREVIATIONS

<b>AD</b>	-	Alzheimer's Disease
<b>BC</b>	-	Breast Cancer
<b>BCW</b>	-	Breast Cancer Wisconsin
<b>CMSC</b>	-	Climate Model Simulation Crashes
<b>CSMiner</b>	-	Contrast Subspace Miner
<b>CSMiner-BPR</b>	-	Contrast Subspace Miner -Bounding-Pruning-Refining
<b>CV</b>	-	Cross Validation
<b>DM</b>	-	Data Mining
<b>DLB</b>	-	Dementia with Lewy Bodies
<b>Freq</b>	-	Frequency
<b>GA</b>	-	Genetic Algorithm
<b>Info Gain</b>	-	Information Gain
<b>k-NN</b>	-	k-Nearest Neighbour
<b>MinObjs</b>	-	Minimum Number of Objects
<b>NB</b>	-	Naïve Bayes
<b>PID</b>	-	Pima Indian Diabetes
<b>RF</b>	-	Random Forest
<b>SVM</b>	-	Support Vector Machine
<b>TTT</b>	-	Tic-Tac-Toe
<b>UCI</b>	-	University of California, Irvine
<b>WEKA</b>	-	Waikato Environment for Knowledge Analysis
<b>Wave</b>	-	Waveform

## LIST OF SYMBOLS

<b><math>C</math></b>	-	Class
<b><math>C_+</math></b>	-	Target class
<b><math>C_-</math></b>	-	Other class
<b><math>d</math></b>	-	Number of features
<b><math>\epsilon</math></b>	-	Small constant value
<b><math>f</math></b>	-	Feature
<b><math>F</math></b>	-	Set of features
<b><math>Fs</math></b>	-	List of one-dimensional subspaces
<b><math>h</math></b>	-	Number of highly scored random subspaces
<b><math>k</math></b>	-	Number of nearest neighbour
<b><math>l</math></b>	-	Number of relevant features
<b><math>n</math></b>	-	Number of objects
<b><math>LS</math></b>	-	List of subspaces
<b><math>\mu</math></b>	-	Number of iterations
<b><math>o</math></b>	-	Object
<b><math>O</math></b>	-	Set of objects
<b><math>O_+</math></b>	-	Target object
<b><math>O_-</math></b>	-	Other object
<b><math>p</math></b>	-	Population size/ Number of Chromosomes
<b><math>P_c</math></b>	-	Probability of crossover
<b><math>P_m</math></b>	-	Probability of mutation
<b><math>q</math></b>	-	Query object
<b><math>r</math></b>	-	Random integer

- $S$**  - Subspace
- $t$**  - Number of random subspaces
- $T$**  - Tree node
- $T_{leaf}$**  - Tree leaf node
- $v$**  - Number of fold for cross validation
- $x$**  - Feature value
- $X$**  - Data set
- $y$**  - Number of nodes



UMS  
UNIVERSITI MALAYSIA SABAH

## LIST OF APPENDICES

	Page
Appendix A: The Classification Accuracy of (a)J48, (b)NB, (c)SVM, and (d)RF for Minimum Number of Objects <i>Minobjs</i> Against Small Constant Value $\varepsilon$	207
Appendix B: The Classification Accuracy of (a)J48, (b)NB, (c)SVM, and (d)RF for Number of Relevant Features / Against Number of Random Subspaces $t$	216
Appendix C: The Classification Accuracy of J48, NB, SVM, and RF Against Number of Nearest Neighbour $k$ on (a)BCW, (b)PID, (c)Wine, (d)Glass, (e)CMSC, and (f)Wave for Tree-Based Method	225
Appendix D: The Tree-Based and Density-Based Likelihood Score Against Dimensionality of Subspace $d$	227
Appendix E: The Classification Accuracy of J48, NB, SVM, and RF Against Number of Nearest Neighbour $k$ on (a)BCW, (b)PID, (c)Wine, (d)Glass, (e)CMSC, and (f)Wave for Density-Based Method	229
Appendix F: List of Publications	230

# CHAPTER 1

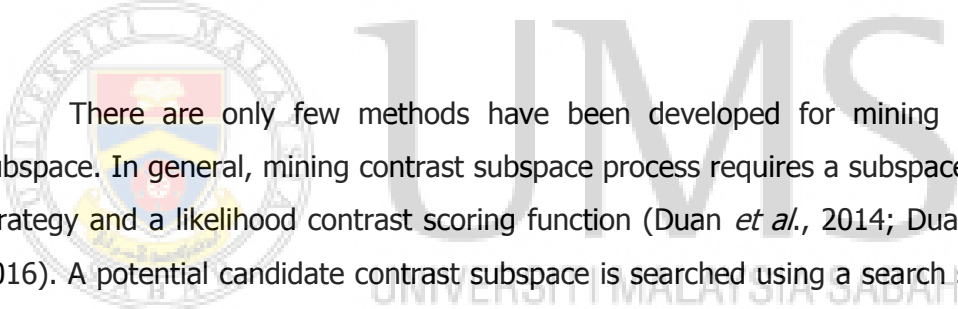
## INTRODUCTION

### 1.1 Background

Current advanced technology has made it possible to accumulate massive amount of data set in a database at lower cost. Data mining is essential to exploit this stored data in order to extract useful information by automatic means. Outlier detection is one of the data mining tasks which it aims to detect data objects whose behaviour deviates significantly from the remaining objects in a data set. Those objects are outliers that may be bad data that need to be removed or malicious data that urge to be tackled. Besides, finding explanation about why the detected outlier is different in a data set is also important to provide information necessary for interpreting the outlier. Accordingly, there are more and more attention is directed to identifying explanation about why and how an object differs in a data set (Duan *et al.*, 2015; Duan *et al.*, 2014; Dang *et al.*, 2013).

In recent years, mining contrast subspace has been introduced which finds explanation about how an object differs between two classes in a data set (Duan *et al.*, 2014). More specifically, given a multidimensional data set of two classes, a target class and a query object, mining contrast subspace finds contrast subspaces where the query object is most similar to the target class while most dissimilar to other class. Those contrast subspaces are subsets of features from the full feature set of the data set. A query object can be any object which its contrast subspaces want to be investigated.

Mining contrast subspace has many important real life applications. One of the examples, in the medical field, Dementia with Lewy Bodies (DLB) and Alzheimer's Disease (AD) are common neurodegenerative dementia happened to older people. DLB and AD share many similar conditions that include memory loss, difficulty in judgment and reasoning, which causes DLB is often misdiagnosed as AD (Surendranathan & O'brien, 2018; Walker *et al.*, 2007). When a doctor wanted to diagnose a patient against these two types of dementia, the doctor may want to know in what subspace the patient is most similar to the cases of DLB and different from AD at the same time. By knowing that subspace, it helps to ensure accurate diagnosis and right treatment to be given for the patient. Another example, in the insurance field, a fraudulent claim is suspected and need to be investigated. An analyst may want to know what subspace makes the claim is similar to the fraud cases but dissimilar to the normal cases. That subspace gives analyst useful information for deeper investigation so as to avoid claim misuse.



There are only few methods have been developed for mining contrast subspace. In general, mining contrast subspace process requires a subspace search strategy and a likelihood contrast scoring function (Duan *et al.*, 2014; Duan *et al.*, 2016). A potential candidate contrast subspace is searched using a search strategy from a collection of possible subspaces derived from the full feature set given in a data set. The similarity of a query object to a target class against other class in the searched subspace is estimated by using a likelihood contrast scoring function. After examining all candidate subspaces, the likelihood contrast score among the subspaces are compared to find the contrast subspaces for the query object.

All of the existing mining contrast subspace methods (i.e. CSMiner and CSMiner-BPR) employ a probability density based likelihood contrast scoring function (Duan *et al.*, 2014; Duan *et al.*, 2016). For a subspace, it estimates the ratio of probability density of a target class against probability density of other class, with respect to a query object. The probability density estimation of a class uses distance between a query object and other objects in the class to measure the similarity of