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

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

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2020

UMS
UNIVERSITI MALAYSIA SABAH

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 Doktoral 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	

UNIVERSITI MALAYSIA SABAH

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 Signature

Assoc. Prof. Dr. Rayner Alfred

UMS
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

UMS
UNIVERSITI MALAYSIA SABAH

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 susbpace 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 mengganggar persamaan objek pertanyaan dengan kelas sasaran serta perbezaan dengan kelas lain pada suatu subruang. Objek pertanyaan berada dalam kelompok yang menpunyai nisbah kepadatan kebarangkalian kelas sasaran kepada kepadatan kebarangkalian kelas lain yang tinggi pada suatu subruang kontras. Walau bagaimanapun, anggaran kepadatan keberangkalian 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 berulangan di mana objek pertanyaan dikumpulkan dengan objek yang mempunyai ciri yang sama. Kaedah tree-based bermula dengan fasa pemilihan atribut yang mencari atribut yang releven 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. Keberkesan 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

UMS
UNIVERSITI MALAYSIA SABAH

UMS
UNIVERSITI MALAYSIA SABAH

UMS UNIVERSITI MALAYSIA SABAH

UMS
UNIVERSITI MALAYSIA SABAH

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF TABLES

UNIVERSITI MALAYSIA SABAH

UMS UNIVERSITI MALAYSIA SABAH

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF FIGURES

UNIVERSITI MALAYSIA SABAH

UNIVERSITI MALAYSIA SABAH

LIST OF ABBREVIATIONS

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

UMS
UNIVERSITI MALAYSIA SABAH

LIST OF SYMBOLS

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

UMS UNIVERSITI MALAYSIA SABAH

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

UMS UNIVERSITI MALAYSIA SABAH

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).

UNIVERSITI MALAYSIA SABAH

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 Alzhemer'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