

**BALANCING EXPLOITATION AND
EXPLORATION SEARCH BEHAVIOR ON
NATURE-INSPIRED CLUSTERING
ALGORITHMS**

MOHAMMED Y. T. ALSWAITTI

UNIVERSITI SAINS MALAYSIA

2018

**BALANCING EXPLOITATION AND EXPLORATION SEARCH
BEHAVIOR ON NATURE-INSPIRED CLUSTERING ALGORITHMS**

by

MOHAMMED Y. T. ALSWAITI

**Thesis submitted in fulfilment of the
requirements for the degree of
Doctor of Philosophy**

August 2018

DEDICATION

To my parents for their endless love and unconditional support, and to my beloved wife who is my source of strength and inspiration.

M.Alswaitti
2018

ACKNOWLEDGEMENT



(Chapter Name: Ibrahim, Verse No: 7)

“If you give thanks (by accepting Faith and worshipping none but Allah), I will give you more (of My Blessings)”.

First and foremost, all praises go to the God, the Almighty, merciful and passionate, for granting me the capability and blessings throughout my research journey. Many people have contributed towards the completion of this thesis. It is an honor for me to place on record my sense of gratitude to one and all, who directly or indirectly have lent their hand in this venture.

Firstly, I would like to express my sincere gratitude to my advisor Professor Dr. Nor Ashidi bin Mat Isa for the continuous support of my study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study. I wish to extend my gratitude to Dr. Mohanad Albughdadi for his invaluable support and insightful suggestions.

Further thanks go to the Ministry of Higher Education (MOHE) which provided the financial support for my study through Malaysia International Scholarship (MIS) Scheme.

Lastly, my earnest thanks to my parents, wife, daughter, brothers, sisters, and friends for their unconditional love, support, understanding, and faith. The appreciation can hardly be expressed in words.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENT	ii
TABLE OF CONTENTS	iii
LIST OF TABLES	vii
LIST OF FIGURES	x
LIST OF ABBREVIATIONS	xii
LIST OF SYMBOLS	xv
ABSTRAK	xviii
ABSTRACT	xx
CHAPTER ONE: INTRODUCTION	
1.1 Introduction	1
1.2 Data Clustering Algorithms	2
1.3 Problem Statement	4
1.4 Research Objectives	7
1.5 Research Scopes	7
1.6 Thesis Outlines	9
CHAPTER TWO: LITERATURE REVIEW	
2.1 Introduction	11
2.2 Clustering	11
2.2.1 Partitional Clustering Algorithms	13
2.2.2 Hierarchical Clustering Algorithms	14
2.2.3 Density-based Clustering Algorithms	14

2.2.4	Optimization-based Clustering Algorithms	15
2.3	Gravitational-based Clustering Algorithms	20
2.3.1	Variations of Gravitational Clustering Algorithms	21
2.3.2	Gravity Clustering Algorithm (GC)	29
2.4	Particle Swarm Optimization-based Clustering Algorithms	34
2.4.1	Particle Swarm Optimization Algorithm	35
2.4.2	PSO Utilization in Cluster Analysis	37
2.4.2(a)	Stand-alone PSO Clustering Algorithms	37
2.4.2(b)	Hybrid PSO Algorithms	46
2.4.3	Improvement Aspects over the Canonical PSO Algorithm	48
2.5	Differential Evolution Optimization-based Clustering Algorithms	52
2.5.1	Differential Evolution Algorithm	53
2.5.1(a)	Population Initialization	54
2.5.1(b)	Mutation	54
2.5.1(c)	Crossover	54
2.5.1(d)	Selection	55
2.5.2	DE Utilization in Cluster Analysis	55
2.6	Summary	62

CHAPTER THREE: METHODOLOGY

3.1	Introduction	64
3.2	Optimized Gravitational-based Data Clustering Framework (OGC)	64
3.2.1	The Proposed OGC Framework	65
3.2.1(a)	Centroids Initialization	67
3.2.1(b)	Gravitational Force Formula	68

3.2.1(c)	Centroid Updating Formula	70
3.3	Density-Based Particle Swarm Optimization Framework for Data Clustering (DPSO)	72
3.3.1	The Proposed DPSO Framework	73
3.3.1(a)	Swarm Initialization	73
3.3.1(b)	Cognitive and Social Terms	75
3.3.1(c)	Learning Coefficients	80
3.4	Variance-based Differential Evolution Framework with an Optional Crossover for Data Clustering (VDEO)	83
3.4.1	The Proposed VDEO Framework	84
3.4.1(a)	Initialization and Problem Formulation	86
3.4.1(b)	Mutation	88
3.4.1(c)	Crossover	91
3.4.1(d)	Selection	92
3.5	Characteristics of the Proposed Frameworks	95
3.5.1	The OGC Framework	95
3.5.2	The DPSO Framework	99
3.5.3	The VDEO Framework	101
3.6	Datasets Descriptions	104
3.7	Evaluation Metrics	108
3.7.1	Classification Accuracy (CA)	109
3.7.2	F-Score	110
3.7.3	Purity	110
3.7.4	Average Distance of Data to the Cluster Centroid (ADDC)	111
3.7.5	The Dunn Index (DI)	111
3.7.6	Friedman Aligned-Ranks (FA) test	112

3.8	Summary	113
-----	---------	-----

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1	Introduction	116
4.2	Results of the Optimized Gravitational-based Data Clustering Framework (OGC)	117
4.3	Results of the Density-Based Particle Swarm Optimization Framework for Data Clustering (DPSO)	129
4.3.1	Comparison with the HPSO Algorithm	132
4.3.2	Performance with State-Of-The-Art Algorithms	140
4.3.3	Complexity Comparison	147
4.3.4	Discussion	150
4.4	Results of the Variance-based Differential Evolution Framework with an Optional Crossover for Data Clustering (VDEO)	152
4.5	Comparison of the Proposed Frameworks	167
4.6	Summary	171

CHAPTER FIVE: CONCLUSION AND FUTURE WORKS

5.1	Conclusion	174
5.2	Future Works	176

REFERENCES	178
-------------------	-----

APPENDICES

Appendix A	PCA of the final Clustering results of the selected datasets by the OGC framework and its competing algorithms
------------	--

LIST OF PUBLICATIONS

LIST OF TABLES

		Page
Table 2.1	Summary of the advantages and disadvantages of the clustering algorithms categories.	18
Table 2.2	Comparison of the existing gravitational-based clustering algorithms.	27
Table 2.3	Comparison of the existing PSO-based clustering algorithms.	44
Table 2.4	Comparison of the existing DE-based clustering algorithms.	60
Table 2.5	Datasets descriptions.	105
Table 4.1	Experimental parameter settings for the competing algorithms.	118
Table 4.2	Average classification accuracy (CA), and standard deviation (Std) of 50 runs of the competing algorithms.	121
Table 4.3	Average F-score experimental results of 50 runs.	124
Table 4.4	Average Purity experimental results of 50 runs.	126
Table 4.5	Average ADDC experimental results of 50 runs.	128
Table 4.6	Parameter settings for experimentation.	130
Table 4.7	Variations of the proposed DPSO framework with different learning coefficients and bandwidth estimation methods.	131
Table 4.8	Average classification accuracy (CA), standard deviation (Std), and Dunn index (DI) of 50 runs using the normalized datasets to compare variations (A-D) with the HPSO algorithm.	135
Table 4.9	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for CA using the normalized datasets to compare variations (A-D) with the HPSO algorithm.	136
Table 4.10	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for DI using the normalized datasets to compare variations (A-D) with the HPSO algorithm.	137
Table 4.11	Average classification accuracy (CA), standard deviation (Std), and Dunn index (DI) of 50 runs using the un-normalized datasets to compare variations (E-H) with the HPSO algorithm.	138

Table 4.12	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for CA the un-normalized datasets to compare variations (E-H) with the HPSO algorithm.	139
Table 4.13	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for DI the un-normalized datasets to compare variations (E-H) with the HPSO algorithm.	139
Table 4.14	Average classification accuracy (CA), standard deviation (Std), and Dunn index (DI) of 50 runs using the un-normalized datasets of the competing algorithms.	142
Table 4.15	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for CA using the un-normalized datasets of the competing algorithms.	143
Table 4.16	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for DI using the un-normalized datasets of the competing algorithms.	143
Table 4.17	Average classification accuracy (CA), standard deviation (Std), and Dunn index (DI) of 50 runs using the normalized datasets of the competing algorithms.	145
Table 4.18	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for CA using the normalized datasets of the competing algorithms.	146
Table 4.19	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for DI using the normalized datasets of the competing algorithms.	146
Table 4.20	Experimental parameter settings for the DE-based clustering algorithms.	153
Table 4.21	Average objective function values and standard deviation (Std) among the competing DE clustering-based algorithms for 50 runs on the 15 datasets.	155
Table 4.22	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for objective function values using the 15 datasets of the competing algorithms.	157
Table 4.23	Average classification accuracy and standard deviation (Std) among the competing DE clustering-based algorithms for 50 runs on the 15 datasets.	158
Table 4.24	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for CA using the 15 datasets of the competing algorithms.	161

Table 4.25	Average classification accuracy and standard deviation (Std) of the proposed frameworks for 50 runs on the 15 datasets.	168
Table 4.26	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for CA using the 15 datasets of the proposed frameworks.	169
Table 4.27	Average objective function values and standard deviation (Std) of the proposed frameworks for 50 runs on the 15 datasets.	170
Table 4.28	Average ranking obtained by Friedman Aligned-Ranks and Holm's test for objective function values using the 15 datasets of the proposed frameworks.	170

LIST OF FIGURES

		Page
Figure 2.1	The classification scheme of clustering algorithms.	12
Figure 2.2	The applied gravitational forces from multiple objects on a single object.	20
Figure 2.3	Iris petals clustering by the GC algorithm through iterations: (a) iteration 1, (b) iteration 50, (c) iteration 100, and (d) iteration 200.	34
Figure 3.1	An overview of the proposed Optimized Gravitational-based Data Clustering Framework (OGC).	66
Figure 3.2	An overview of the proposed Density-based Particle Swarm Optimization Framework for Data Clustering (DPSO).	74
Figure 3.3	The effect of the bandwidth σ value on the KDE. (a) fixed σ value ($\sigma=0.01$), (b) fixed σ value ($\sigma=0.5$), (c) estimated σ value by Wu et al. (2016) ($\sigma=0.166$), (d) estimated σ value by the proposed approach ($\sigma=0.0952$). The solid black line, dashed blue curves and red dots represent the KDE, individual kernels and data points, respectively.	78
Figure 3.4	(a) The properties of the normal distribution curve, (b) Overlapping region of two data points with one σ distance.	79
Figure 3.5	An overview of the proposed Variance-based Differential Evolution Framework with an Optional Crossover for Data Clustering (VDEO).	85
Figure 3.6	Dataset splitting steps into K subsets	88
Figure 3.7	Ruspini dataset clustering by the proposed OGC framework through iterations without initialization method. (a) Iteration 0, (b) Iteration 10, (c) Iteration 50, (d) Iteration 100, and (e) the final solution.	97
Figure 3.8	Ruspini dataset clustering by the proposed OGC framework through iterations with the variance-based initialization method. (a) Iteration 0, (b) Iteration 10, (c) Iteration 50, (d) Iteration 100, and (e) the final solution.	99
Figure 3.9	Ruspini dataset clustering by the proposed DPSO framework through iterations. (a) Iteration 0, (b) Iteration 10, (c) Iteration 50, (d) Iteration 100, and (e) the final solution.	100
Figure 3.10	Ruspini dataset clustering by the proposed VDEO framework. (a) initial setups of the framework, (b) random	111

solutions generation using the first mutation scheme, (c) random solutions generation using the second mutation scheme, (d) neither the mutant nor the trial solutions are fitter than the current one, (e) the trial solution is fitter than the current, and (f) the final solution.

Figure 4.1	Iris petals clustering by the proposed OGC framework through iterations: (a) iteration 1, (b) iteration 10, (c) iteration 100, and (d) iteration 200.	120
Figure 4.2	(a) WDBC-Int original dataset and the PCA clustering results by: (b) the proposed OGC, (c) K means++, (d) FCM, (e) SGC, and (f) GC algorithms.	123
Figure 4.3	(a) Lung Cancer original dataset and the PCA clustering results by: (b) the proposed OGC, (c) K means++, (d) FCM, (e) SGC, and (f) GC algorithms.	125
Figure 4.4	3-D scatter of the original Balance dataset using the principle component analysis.	127
Figure 4.5	The effect of the number of observations on the execution time using a synthetic dataset.	148
Figure 4.6	The effect of the number of dimensions on the execution time using a synthetic dataset.	148
Figure 4.7	The execution time of the proposed DPSO framework and the HPSO algorithm when applied to the 15 datasets from the UCI.	149
Figure 4.8	The convergence rate of the competing algorithms on the 15 datasets. (a) iris, (b) Haberman, (c) New Thyroid, (d) Seeds, (e) Lung Cancer, (f) Glass, (g) Wine, (h) Balance, (i) Vowel, (j) BSTCD, (k) Heart, (l) WDBC-Int, (m) Dermatology, (n)WDBC, (o) Landsat.	165

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony Algorithm
ACDE	Automatic Clustering Using an Improved Differential Evolution Algorithm
ACO	Ant Colony Optimization Algorithm
ACPSO	Accelerated Chaotic Particle Swarm Optimization Algorithm
ACROA	Artificial Chemical Reaction Optimization Algorithm
ADDC	Average Distance of Data to The Cluster Centroid
BBO	Biogeography-based Optimizer
BH	Black Hole Optimization Algorithm
CA	Classification Accuracy
CLARANS	Clustering Objects for Spatial Data Mining
CPSO	Chaotic Particle Swarm Optimization Algorithm
CS	Cuckoo Search Optimization Algorithm
DBSCAN	Density Based Spatial Clustering of Applications with Noise
DE	Differential Evolution Algorithm
DEMM	Data Clustering with Differential Evolution Incorporating Macromutations
DI	Dunn Index
DPSO	Density-based Particle Swarm Optimization Framework for Data Clustering
DSDE	Dynamic Shuffled Differential Evolution Algorithm for Data Clustering
ECPSO	Extended Chaotic Particle Swarm Optimization Algorithm
EPSO	An Evolutionary Particle Swarm Optimization Algorithm for Data Clustering
ES	Evolution Strategy
FA	Friedman Aligned-Ranks

FCM	Fuzzy C-means Algorithm
FFA	Firefly Optimization Algorithm
FN	False Negatives
FP	False Positives
FSDE	Forced Strategy Differential Evolution Algorithm for Data Clustering
GA	Genetic Algorithm
GC	Gravitational Clustering Algorithm
GKPSOCA	Gaussian Kernel PSO Algorithm
GP	Genetic Programming
GPSO	Gravity-based Particle Swarm Optimization algorithm
GRIN	An Incremental Hierarchical Clustering Algorithm for Numerical Datasets Based on the Gravity Theory in Physics
GSA	Gravitational Search Algorithm
GSA-HS	Gravitational Search Algorithm with A Heuristic Search for Clustering Problems
GSA-KM	A Combined Approach for Clustering Based On K-means and Gravitational Search Algorithms
GSOM	Gravitational Clustering of the Self-Organizing Map
HPSO	Particle Swarm Optimization Based Hierarchical Agglomerative Clustering
KDE	Kernel Density Estimation
MEPSO	Automatic Kernel Clustering with A Multi-Elitist Particle Swarm Optimization Algorithm
MOA	Magnetic Optimization Algorithm
OGC	Optimized Gravitational-based Data Clustering Framework
OPTICS	Ordering Points to Identify the Clustering Structure
PAM	Partitioning Around Medoids
PCA	Principal Component Analysis

pdf	Probability Density Function
PLDC	PSO-based Local Density Model
PSC	Particle Swarm Clustering Algorithm
PSO	Particle Swarm Optimization Algorithm
RAIN	Clustering Algorithm Based on the Randomized Interactions of Data Points
RO	Ray Optimization Algorithm
SGC	Simplified Gravitational Clustering Algorithm
TP	True Positives
TRW	Trace Within Criterion
UCI	Machine Learning Repository
VCR	Variance Ratio Criterion
VDEO	Variance-based Differential Evolution Framework with an Optional Crossover for Data Clustering

LIST OF SYMBOLS

a	The Acceleration
C	Cluster
$ C $	Cluster Cardinality
C_r	Crossover Probability
D	Euclidean Distance
d	Number of Dimensions
d_f	Particles Controlling Parameter
$dist_{max}$	Largest Distance Among a Number of Clusters
$dist_{min}$	Distance Between the Closest Two Data Points
e	Distance Controlling Parameter
F	Gravitational Force
F_g	Gravitational Learning Coefficient
$f(\cdot)$	Objective Function
G	Gravitational Constant
G_0	Initial Value of the Gravity Constant
gB	The Global Best Position
h	Mutation Controlling Factor
K	Number of Clusters
$KE(\cdot)$	Kernel Function
l	Learning Coefficient
M	Object Mass
N	Number of Data Points
N_p	Number of Population
O	Swarm of Particles

P	Precision
pB	The Personal Best Position
$pD_{j,d}$	The Personal Dense Position
$P_{i,d}$	Position of A Particle
P_{mm}	Probability of Macro-Mutations
Q	Trial Solution
R	Recall
r	Randomly Generated Number by A Uniform Distribution in an Interval [0,1]
S	Target Solution
S_{best}	Best Solution Found at The Current Iteration
t	The Current Iteration
$T(\cdot)$	Similarity Function
U	Centroid
v	The Velocity
var	The Variance
W	Maximum Number of Iterations
X	Data Point
X_{bmax}	The Maximum Bound of the Subset Space
X_{bmin}	The Minimum Bound of the Subset Space
X_{max}	The Maximum Bound of the Search Space
X_{min}	The Minimum Bound of the Search Space
Y	Mutant Solution
Z	Ground Truth of a Cluster
Δt	Step of Time
α	A Small Positive Number to Avoid Division by Zero
δ	Step of Distance

ε	Distance Threshold
μ	Mutation Factor
σ	Kernel Bandwidth
φ	Threshold Value
ω	The Inertial Weight

**PENYEIMBANGKAN SIFAT PENCARIAN EKSPLOITASI DAN
EKSPLOKASI KE ATAS ALGORITMA PENGELOMPOKAN
BERINSPIRASIKAN SEMULAJADI**

ABSTRAK

Teknik-teknik pengelompokan berasaskan pengoptimuman yang diinspirasi daripada alam semulajadi adalah berkuasa, teguh dan lebih canggih daripada kaedah-kaedah pengelompokan konvensional disebabkan ciri stokastik dan heuristik teknik-teknik tersebut. Namun demikian, algoritma-algoritma ini mempunyai beberapa kelemahan seperti kecenderungan untuk terperangkap dalam optima tempatan dan kadar penumpuan yang lambat. Kelemahan yang kedua adalah akibat daripada kesukaran dalam mengimbangi proses eksplorasi dan eksploitasi yang mana telah mempengaruhi secara langsung kualiti akhir proses pengelompokan. Oleh itu, penyelidikan ini telah mencadangkan tiga kerangka kerja yang ditambah baik iaitu Pengoptimuman berasaskan Graviti (OGC), Pengoptimuman Kawanan Zarah berasaskan Ketumpatan (DPSO), dan Evolusi Kebezaan berasaskan Varians dengan Lintas Pilihan (VDEO) untuk proses pengelompokan data. Dalam kerangka kerja OGC, sifat pencarian penerokaan algoritma Penggugusan Graviti (GC) telah ditambahbaik dengan (i) menghapuskan penumpukan halaju ejen, dan (ii) mengintegrasikan kaedah pememulaan agen-agen menggunakan varians dan median untuk menyusun proses eksplorasi. Selain itu, keseimbangan antara proses eksplorasi dan eksploitasi dalam kerangka kerja DPSO dipertimbangkan dengan menggunakan gabungan (i) teknik penganggaran ketumpatan inti yang berkaitan dengan kaedah penganggaran lebar jalur baharu dan (ii) anggaran pekali pembelajaran graviti pelbagai dimensi. Akhir sekali, (i) perwakilan penyelesaian berasaskan tunggal, (ii) skim mutasi

boleh ubah, (iii) anggaran berasaskan vektor bagi faktor mutasi, dan (iv) strategi lintas pilihan dicadangkan dalam kerangka kerja VDEO. Prestasi keseluruhan ketiga-tiga kerangka kerja yang dicadangkan ini telah dibandingkan dengan beberapa algoritma-algoritma pengelompokan terkini menggunakan 15 set data daripada repositori UCI. Keputusan-keputusan eksperimen juga dinilai dengan teliti dan disahkan dengan analisis statistik tak berparameter. Berdasarkan keputusan-keputusan eksperimen yang diperolehi, kerangka OGC, DPSO, dan VDEO masing-masing telah mencapai peningkatan purata sehingga 24.36%, 9.38%, dan 11.98% untuk kejituan klasifikasi. Semua kerangka kerja juga telah mencapai kedudukan pertama dalam ujian Pangkat Sejajar Friedman (FA) dalam semua metrik penilaian. Selain itu, ketiga-tiga kerangka kerja tersebut telah menghasilkan penumpuan pencapaian dari segi keboleholangan. Kerangka kerja OGC telah menghasilkan prestasi yang ketara dari segi kejituan klasifikasi, manakala kerangka kerja VDEO telah menunjukkan prestasi yang ketara dari segi kepadatan kelompok. Dalam hal lain, kerangka kerja DPSO mempunyai kelebihan dari segi keseimbangan keadaan dengan menghasilkan keputusan yang sangat kompetitif berbanding OGC dan DPSO dalam kedua-dua metrik penilaian. Sebagai kesimpulan, mengimbangi kelakuan pencarian telah dengan jelasnya meningkatkan prestasi keseluruhan ketiga-tiga kerangka kerja yang dicadangkan dan menjadikan setiap kerangka kerja tersebut sebagai alat yang sangat baik untuk pengelompokan data.

BALANCING EXPLOITATION AND EXPLORATION SEARCH BEHAVIOR ON NATURE-INSPIRED CLUSTERING ALGORITHMS

ABSTRACT

Nature-inspired optimization-based clustering techniques are powerful, robust and more sophisticated than the conventional clustering methods due to their stochastic and heuristic characteristics. Unfortunately, these algorithms suffer with several drawbacks such as the tendency to be trapped or stagnate into local optima and slow convergence rates. The latter drawbacks are consequences of the difficulty in balancing the exploration and exploitation processes which directly affect the final quality of the clustering solutions. Hence, this research has proposed three enhanced frameworks, namely, Optimized Gravitational-based (OGC), Density-Based Particle Swarm Optimization (DPSO), and Variance-based Differential Evolution with an Optional Crossover (VDEO) frameworks for data clustering. In the OGC framework, the exhibited explorative search behavior of the Gravitational Clustering (GC) algorithm has been addressed by (i) eliminating the agent velocity accumulation, and (ii) integrating an initialization method of agents using variance and median to subrogate the exploration process. Moreover, the balance between the exploration and exploitation processes in the DPSO framework is considered using a combination of (i) a kernel density estimation technique associated with new bandwidth estimation method and (ii) estimated multi-dimensional gravitational learning coefficients. Lastly, (i) a single-based solution representation, (ii) a switchable mutation scheme, (iii) a vector-based estimation of the mutation factor, and (iv) an optional crossover strategy are proposed in the VDEO framework. The overall performances of the three

proposed frameworks have been compared with several current state-of-the-art clustering algorithms on 15 benchmark datasets from the UCI repository. The experimental results are also thoroughly evaluated and verified via non-parametric statistical analysis. Based on the obtained experimental results, the OGC, DPSO, and VDEO frameworks achieved an average enhancement up to 24.36%, 9.38%, and 11.98% of classification accuracy, respectively. All the frameworks also achieved the first rank by the Friedman aligned-ranks (FA) test in all evaluation metrics. Moreover, the three frameworks provided convergent performances in terms of the repeatability. Meanwhile, the OGC framework obtained a significant performance in terms of the classification accuracy, where the VDEO framework presented a significant performance in terms of cluster compactness. On the other hand, the DPSO framework favored the balanced state by producing very competitive results compared to the OGC and DPSO in both evaluation metrics. As a conclusion, balancing the search behavior notably enhanced the overall performance of the three proposed frameworks and made each of them an excellent tool for data clustering.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Recently, the vast advancements in data storage technologies and internet applications have resulted in a massive growth of data quantity of all types. This diversity of the data is an outcome of an endless sequence of daily life interactions while accessing, recording, and transferring information (such as text, images, and videos) among humans. The increase in both the volume and the variety of this data induced the need for an advanced technology that is automatically capable of summarizing these huge amounts of data into meaningful, comprehensible, and useful information.

To meet this requirement, data mining has emerged as a powerful technique to extract the valuable hidden information and knowledge from the large databases. Cluster analysis is one of the simplest data mining tools that used to categorize the data objects based on their features into a set of natural and similar clusters without prior knowledge of the data. Naturally, the grouped objects within the same cluster share a high degree of similarity while being dissonant to other objects belonging to other clusters. In other words, the formed clusters should satisfy a high degree of homogeneity within their members and a high degree of heterogeneity to other clusters.

Grouping patterns into meaningful clusters in an unsupervised manner is done using clustering algorithms where they play an outstanding role in machine learning due to their capabilities in exploring data without having any prior information about them, i.e., there are no labels associated with these data. These algorithms aim at modeling the underlying structure or distribution in the data, which can be used for