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. ALSWAITTI

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

بِسْمِ اللَّهِ الرَّحْمَٰنِ الرَّحِيمِ

{ لَئِن شَكَرْتُمْ لأَزِيدَنَّكُمْ }

(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

LIST OF TABLES

LIST OF FIGURES

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 EKSPLORASI KE ATAS ALGORITMA PENGELOMPOKAN BERINSPIRASIKAN SEMULAJADI

ABSTRAK

Teknik-teknik pengelompokan berasaskan pengoptimuman yang diinspirasikan 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 kebolehulangan. 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