ENHANCED MEMETIC DIFFERENTIAL EVOLUTION OPTIMISATION ALGORITHMS FOR DATA CLUSTERING PROBLEMS

HOSSAM MOHS'D JABR MUSTAFA

UNIVERSITI KEBANGSAAN MALAYSIA

ENHANCED MEMETIC DIFFERENTIAL EVOLUTION OPTIMISATION ALGORITHMS FOR DATA CLUSTERING PROBLEMS

HOSSAM MOH'D JABR MUSTAFA

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

ALGORITMA PENGOPTIMUMAN EVOLUSI PEMBEZAAN MEMETIC DIPERTINGKAT UNTUK MASALAH PENGGUGUSAN DATA

HOSSAM MOH'D JABR MUSTAFA

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI
IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2019

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have been duly acknowledged.

15 August 2019

HOSSAM MOH'D JABR MUSTAFA
P81300

ACKNOWLEDGEMENT

First and foremost, all praises and thanks to the mighty Allah for giving me enough strength and blessing to complete this thesis.

Surely, many wonderful people have contributed significantly to the completion of this thesis. I owe a great deal to them.

I would like to express my sincere gratitude and appreciation to my supervisor *Professor Dr. Masri Ayob* for her excellent guidance, support, and kindness during my research and the completion of this thesis. It was a great honour and opportunity to work with her throughout my PhD program. In fact, this thesis is the result of her fantastic assistant, which directed my efforts toward accomplishing the objectives of this research.

Additionally, I would like to express my sincere appreciation to my co-supervisor *Dr. Mohd Zakree Ahmad Nazri*, for continuous guidance in the academic suggestions and good advices.

Further, I would like to express genuine appreciation to my family for their extreme love and kindness. I would like to give special thanks to my parents for their spiritual assistance. My success expresses and follows the support and love from both of them. My sincerest appreciation belongs to my beloved wife for being so understanding hearted patient throughout PhD journey. She has been my inspiration and motivation for proceeding to enhance my education. I also thank my wonderful children: *Mohammad, Nada*, *Adam* and *Elena* for continuously getting me to smile. Their love motivated me and granted me spirit from overseas. I love you all.

Last but not least, I am really thankful to my friends who have supported me whenever I needed them, to all UKM staff and members of the Faculty of Information Science and Technology for assistance. It is my pleasure to have a special appreciation to all of the DMO research group members for being a member of my second home.

ABSTRACT

The performance of data clustering algorithms depends mainly on their ability to balance between the exploration and exploitation of the search, and the effectiveness of outlier detection techniques. Although recent single criterion data clustering algorithms have achieved reasonable quality solutions for some datasets, their performance across real-life datasets could be improved. Moreover, most of these data clustering algorithms adopt a single criterion optimisation approach, which often fails to find good data clustering solutions for a wide diversity of datasets with different cluster characteristics. A multi-objective meta-heuristic approach is sometimes been utilised to address this issue, which seeks an optimal clustering solution by maximising or minimising more than one objective functions. Some of these data clustering algorithms (e.g. differential evolution (DE), particle swarm optimisation (PSO) or non-dominated sorting genetic algorithm (NSGA-II)) find good quality solutions for some datasets, but fail to attain good results across all datasets. These shortcomings could be caused by the challenges of balancing exploration and exploitation, which may lead to premature convergence, stagnation or weak diversity in the pareto-front solutions. Moreover, the design of these clustering algorithms is usually developed using distance measures. These algorithms may experience challenges in identifying data points that are either noise or outlier. Three memetic differential evolution algorithms are proposed to overcome the shortcomings mentioned above. The research first proposes a single criterion memetic differential evolution optimisation algorithm (MADE). The memetic algorithm (MA) employs an adaptive DE mutation operator. Such a combination expected to improve the convergence and gain a better balance between exploration and exploitation. The experimental results, based on several real-life benchmark datasets taken from the UCI repository, show that MADE outperformed other competing algorithms. Next, the research introduces a multi-objective memetic differential evolution algorithm (MOMDE) for data clustering. The MOMDE combines the memetic differential evolution algorithm with the dominance-based multi-objective approach, in order to improve the search for optimal clustering by maximising or/and minimising two cluster quality measures for many datasets. Finally, the research proposes an enhanced MOMDE algorithm (eMOMDE) based on the local outlier factor (Conn_LOF), which aims to improve the performance of the connectivity measure of objective function by eliminating the outliers. The experiments based on real-life datasets from the UCI machine learning repository and synthetic two-dimensional datasets showed that the MOMDE and eMOMDE algorithms outperformed other compared data clustering algorithms. The external validity is evaluated using the F-measure to evaluate the accuracy of the obtained clustering, whilst the multi-objective performance assessment metrics is used to evaluate the quality of Pareto-optimal sets such as convergence, diversity, coverage, and overall non-dominant vector generation. Generally, in most of the cases, the proposed algorithms significantly outperformed recent researches when tested on standard benchmark datasets. This indicates that the combination between the adaptive DE mutation strategy, local search, multi-objective optimisation, and handling outliers within the clustering criterion can enhance the performance of the MA in solving the data clustering problems for different kinds of datasets.

ABSTRAK

Prestasi data kelompok algoritma bergantung terutamanya pada keupayaan mereka untuk mengimbangi antara penerokaan dan eksploitasi carian, dan keberkesanan teknik-teknik pengesanan outlier. Walaupun data tunggal kriteria terkini kelompok algoritma telah mencapai penyelesaian kualiti yang munasabah bagi sesetengah datasets, prestasi mereka merentasi pelbagai datasets masih boleh dipertingkatkan. Selain itu, kebanyakan algoritma kelompok data menggunakan pendekatan pengoptimuman kriteria tunggal, yang sering gagal untuk mencari penyelesaian kelompok data yang baik bagi kepelbagaian datasets yang mempunyai ciri-ciri kelompok yang berbeza. Pendekatan berbilang objektif meta-heuristik seringkali digunakan untuk menangani isu ini, untuk menghasilkan penyelesaian pengelompokan optimum dengan memaksimumkan atau meminimumkan fungsi objektif yang lebih daripada satu. Antara algoritma kelompok data (cth: evolusi pembezaan (DE), zarah GI pengoptimuman (PSO) atau algoritma genetik sisihan tanpa dominasi (NSGA-II)) yang mencari kualiti penyelesaian yang baik bagi sesetengah datasets, tetapi gagal untuk mencapai keputusan yang baik di semua datasets. Kekurangan ini mungkin disebabkan oleh cabaran mengimbangi penerokaan dan eksploitasi, yang boleh membawa kepada pertembungan pramatang, stagnasi atau lemah dalam penyelesaian pareto-hadapan. Selain itu, rekabentuk algoritma kelompok ini, biasanya dibangunkan menggunakan ukuran jarak. Algoritma ini mungkin mengalami cabaran dalam mengenal pasti antara titik data, bunyi bising atau outlier. Tiga algoritma pembezaan evolusi memetic dicadangan untuk mengatasi kelemahan yang tersebut. Kajian pertama mencadangkan algoritma pengoptimuman yang menggabungkan kriteria tunggal memetic dalam algorithma pembezaan evolusi (MADE). Algoritma memetic (MA) menggunakan mutasi DE mudah suai. Gabungan itu dijangka meningkatkan eksploitasi dan mengimbangi antara penerokaan dan eksploitasi. Keputusan eksperimen, berdasarkan beberapa datasets tanda aras sebenar yang diambil daripada repositori UCI, menunjukkan bahawa prestasi algoritma MADE adalah setanding dengan algorithma lain. Seterusnya, kajian ini memperkenalkan algoritma memetic pembezaan evolusi pelbagai objektif (MOMDE) untuk pengelompokan data. Next, the research introduces a multi-objective memetic differential evolution algorithm (MOMDE) for data clustering. MOMDE menggabungkan algoritma memetic pembezaan evolusi dengan algorithma dominasi pelbagai objektif, untuk memperbaiki carian bagi kelompok yang optimum dengan memaksimakan dan/atau meminimakan dua ukuran kualiti untuk pelbagai datasets. Akhir sekali, kajian ini mencadangkan untuk dipertingkatkan MOMDE algoritma (eMOMDE) berdasarkan faktor outlier tempatan (Conn_LOF), yang bertujuan untuk meningkatkan prestasi dalam ukuran objektif fungsi perhubungan dengan menghapuskan outliers. Eksperimen ke atas set data nyata daripada repositori pembelajaran mesin UCI dan set data sintetik dua dimensi menunjukkan bahawa algoritma MOMDE dan eMOMDE menandingi prestasi algoritma kelompok yang lain. Validasi luaran adalah dinilai menggunakan ukuran-F untuk menilai ketepatan kelompok yang diperolehi, manakala metrik penilaian prestasi pelbagai objektif digunakan untuk menilai kualiti set optimum Pareto seperti penumpuan, kepelbagaian, lingkungan, dan penjanaan vektor bebas-dominan. Umumnya, dalam kebanyakan kes, algoritma cadangan jelas menandingi kajian terdahulu apabila diuji ke atas set data tanda aras. Ini menunjukkan bahawa gabungan antara strategi mutasi DE mudah suai, carian tempatan, pengoptimuman pelbagai objektif dan pengendalian outliers dalam kriteria pengelompokan boleh meningkatkan prestasi MA dalam menyelesaikan masalah pengelompokan data bagi pelbagai jenis set data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony algorithm
ACO	Ant Colony Optimisation algorithm
BEA	Bacteria Evolutionary Algorithm
BH	Black Hole algorithm
CS-DBSCAN	Improved Density-Based Spatial Clustering
CSO	Cat Swarm Optimisation algorithm
DE	Differential Evolution Algorithm
DSDE	Dynamic Shuffled Differential Evolution algorithm
EA	Evolutionary Algorithm
ER	Error Rate
FA	Firefly Algorithm
GA	Genetic Algorithm
GSA	Gravitational Search Algorithm
GWO	Grey Wolf Optimiser
H-KHA	Hybrid of KH with HS algorithms
HS	Harmony Search Algorithm
ICA	Imperialist Competitive Algorithm
ICMPKHM	Combined KHM with ICS and PSO algorithms
ICS	Improved Cuckoo Search
KHA	Krill herd algorithm
KHM	K-Marmonic Means algorithm
KSC-LCA	Hybrid of k-means and Chaotic LCA
LCA	League Championship Algorithm
LOF	Local Outlier Factor
MA	Memetic Algorithm
MADE	Memetic Differential Evolution Algorithm
MOAC	Magnetic Optimization Algorithm for data Clustering

MOMDE	Multi-Objective Memetic Differential Evolution Algorithm
MOO	Multi-Objective Optimisation
NSABC	Non-dominated sorting based multi-objective ABC
NSGA-II	Non-dominated Sorting Genetic Algorithm
ONVG	Overall Non-dominant Vector Generation
PI	Performance assessment Indices
P-metaheuristics	Population-based Metaheuristics
PSO	Particle Swarm Optimizer
PSOAG	Age-based Particle Dwarm Optimization algorithm
SA	Simulated Annealing
S-metaheuristics	Single-solution based Metaheuristics
SN	Signal to Noise
SPEA-II	improved Strength Pareto Evolutionary Algorithm
TS	Tabu Search
TSMPSO	Hybrid multiobjective particle swarm optimization

CHAPTER I

INTRODUCTION

1.1 OVERVIEW

This thesis investigates data clustering problem that aims to produce discriminatory clustering methods, which can enhance the quality of clustering solutions. The thesis has seven chapters including the current one. This chapter presents the main research components; including the background of the research, the problem statement, the research scope and objectives. Thus, the organisation of the thesis structure is described at the end of this chapter.

1.2 RESEARCH BACKGROUND

The data mining learning algorithms could be grouped into supervised algorithms, and unsupervised algorithms (Lantz 2013; Tan et al. 2006). Supervised learning algorithms are used to train predictive models on the instruction of what and how to learn. Whilst, unsupervised learning algorithms are used to train descriptive models that are used for tasks that would benefit from the insight of summarised data. The most widely used data mining tasks include pattern discovery, numeric prediction, classification, and cluster analysis. Figure 1.1 illustrates the most popular data mining tasks.

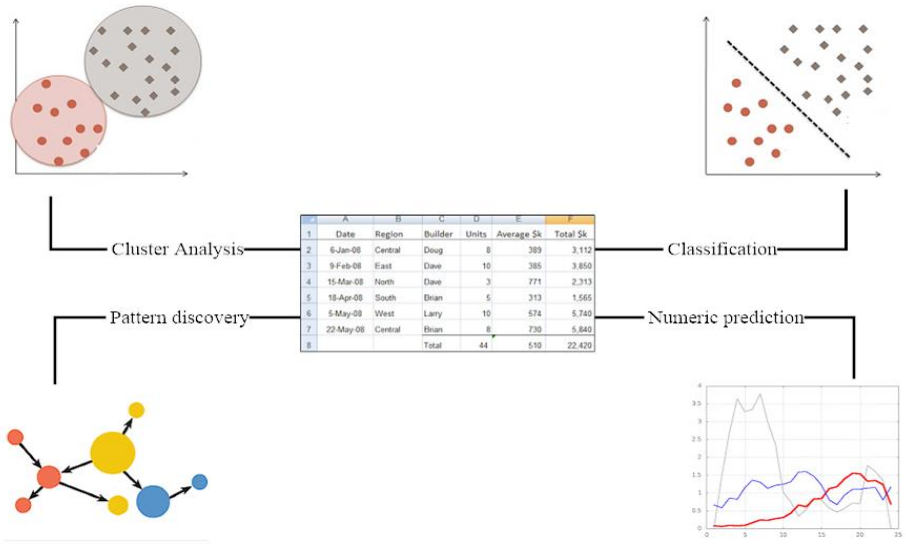


Figure 1.1 The most popular data mining tasks

In brief, the classification is a supervised learning task that consists of learning and classification steps. In the learning step, data with known class labels are used to build a predictive model and the classification algorithm analyses the training data. In the classification step, the label of new data is unknown, and will be predicted using the predictive model. The numeric prediction analysis can be used to model the relationship between independent variables and dependent variables. The independent variables are known as attributes, and the response variables are what to be predicted. The pattern discovery is an unsupervised learning algorithm that predict the models of such dependencies variables, and the objective of these learning algorithms is to build a predictive model for estimating the next values of the series based on the earlier observed values.

Clustering is a common descriptive task that seeks to identify a finite set of categories or cluster to describe the data. The clustering has no training stage; it is usually used when both class labels and the number of classes are not known in advance. Clustering is widely used in different application to gain insight into the structure of the data, to focus on a specific set of clusters for further analysis, and to detect the characteristics of each cluster. Clustering has been developed and used as a basic tool for different disciplines and fields such as Information Retrieval (Wu et al. 2013), Internet of Things (Abbasi & Younis 2007; Tsai et al. 2014), Business (Müller & Hamm 2014), Medicine (Esfandiari et al. 2014; Nahar et al. 2013b, 2013a), Image

segmentation (Gonzalez & Woods 2002; Kumar et al. 2014a; Pratt 2000), and Climate (Dowd et al. 2017; Steinbach et al. 2003).

In recent decades, clustering methods have been extensively studied (Abul Hasan & Ramakrishnan 2011; Aggarwal & Reddy 2013; Esmin et al. 2015; Kushwaha et al. 2017; Saxena et al. 2017), especially for distance based cluster analysis. Moreover, much effort has been focused on finding efficient and effective methods for clustering high dimensional datasets and complex shape clusters (e.g., non-convex shapes). Most of the clustering methods in the literature partition data into a predefined number of clusters are based on the used fitness (objective) function. The fitness function serves a major function in partitioning data. Thus, we need to choose the fitness function carefully and ensure that it is suitable for the used dataset. In most cases, the fitness function is unsuitable for all kinds of datasets; it can only be suitable for certain kinds of datasets. For example, the intra-cluster distance fitness function is suitable for spherical distributed data, whereas the connectivity fitness function is suitable for well-separated datasets regardless of their shape (Das et al. 2009; Gan et al. 2007; Kaufman & Rousseeuw 1990).

The clustering methods can be broadly classified based on the fitness function such as partitioning methods (Celebi 2015; Jain 2010; Wu et al. 2013), hierarchical methods (Das et al. 2009; Tan et al. 2006), density-based methods (Daszykowski & Walczak 2010; Shamshirband et al. 2014), grid-based methods (Aggarwal & Reddy 2013; Das et al. 2009), and graph-based methods (Gallardo & Cotta 2015; Schaeffer 2007). The internal and external quality measures can be used to identify the quality of the clustering. Well-known internal quality measures include intra-cluster distance, connectivity, Dunn index, Silhouette index, and separation index. Thus, well-known external quality measures include F-measure, purity, Rand index, and error rate. As a summary, Table 1.1 describes the common data mining tasks and their evaluation criteria.

Table 1.1 Brief description and evaluation criteria of the popular data mining tasks

Data mining task	Brief description	Evaluation criteria
Classification	The supervised algorithm that predicts the classes of unclassified data by using the prediction model build from known classes. Examples of classifiers are Nearest Neighbour, Naive Bayes, Decision Trees, Neural Networks, and Support Vector Machines.	Accuracy, computational time, scalability, and robustness.
Numeric prediction	The supervised algorithm that predicts (forecast) numeric data by using the numeric prediction model build from known classes. Examples of numeric predictors are Linear Regression, Regression Trees, Neural Networks, and Support Vector Machines.	Accuracy, computational time, scalability, and robustness.
Pattern discovery	Unsupervised learning algorithms used to identify frequent associations within data. Example of pattern detectors is Association Rules.	Support and confidence
Clustering	Unsupervised learning algorithms use the descriptive modelling task for dividing a dataset into homogeneous groups. Examples of clustering algorithms are K-means, K-medoids, hierarchical clustering, and meta-heuristic clustering.	Internal and external quality measures.

1.3 PROBLEM STATEMENT

Although there is a large number of sophisticated clustering algorithms in a wide range of applications and fields, clustering remains a complex task due to the wide variety of applications, and different types and volumes of data. Finding a single clustering algorithm that can fulfil all requirements of clustering is still unrevealed. Additionally, clustering is a non-deterministic polynomial-time hardness problem (Figueiredo et al. 2019; Jain 2010), which generates a huge search space that grows exponentially with the data volume and leads to unexplored search space regions even with medium sizes of datasets.

One of the popular clustering methods is classified as partitioning clustering methods, which attempt to divide the dataset into a set of disjoint clusters and try to optimise specific criterion function that may emphasise the local structure of the data. The most popular partition clustering algorithms are k-means, k-medoids, expectation maximisation, clustering large applications (Figueiredo et al. 2019; Jain 2010). The K-means algorithm, recognised as being efficient and straightforward, is one of the popular for centre-based clustering (Figueiredo et al. 2019; Jain 2010). However, K-

means can detect only well-separated, compact or spherical clusters (Everitt et al. 2011a). It is sensitive to noise due to the use of squared Euclidean distance, where any point in the cluster can significantly influence the centre of clusters. The performance of K-means is also highly sensitive to the selection of initial centres (Jain 2010). Improper initialisation may lead to empty clusters, weak convergence and a high possibility of getting trapped in a local optima (which is best clustering solution within the neighbouring possible solutions) rather than finding the global optima (which is the optimum solution across of all possible solutions) (Jain 2010). Some researchers overcome these issues by using metaheuristics, such as Genetic algorithms (Mustafi et al. 2017), Particle Swarm Optimization (Niu et al. 2017), Ant Colony Optimization (İnkaya et al. 2015), Black Hole Algorithm (Chandrasekar & Krishnamoorthi 2014), Gravitational Search Algorithm (Han et al. 2017) and Krill Herd algorithm (Abualigah et al. 2017).

In clustering problems, the balance between exploration and exploitation and preserving population diversity can affect the ability of the clustering algorithm in finding good clusters among the datasets being used (Dowlatshahi & Nezamabadi-Pour 2014; Kumar et al. 2015; Liu et al. 2012). Some of the earlier proposed clustering algorithms, based on metaheuristics, managed to find good clustering solutions for particular datasets. However, these algorithms were unable to find good solutions across all clustering problem datasets, or their results were inconsistent (Aggarwal & Reddy 2013; Celebi 2015). These deficiencies in the results might be due to the imbalance between exploration and exploitation and inappropriate diversity preservation mechanism of the metaheuristic algorithm that may lead to premature convergence or stagnation (Bouyer & Hatamlou 2018; Dowlatshahi & Nezamabadi-Pour 2014; Figueiredo et al. 2019). Some researchers have proposed a hybrid approach by combining a global search with a local search (which searches the neighbourhood solutions to find better solution) to achieve a better balance and diversity. The global search handles exploration, while exploitation is handled by the local search (Jaradat et al. 2016; Talbi 2012; Yassen et al. 2015, 2017). The Memetic Algorithms (MAs) are one type of hybrid evolutionary algorithms (EA) (which are based on biological evolution, such as mutation, reproduction, selection, and recombination, and utilised to capture global solutions) that offers an efficient optimisation framework by

combining perturbation mechanisms, local search strategies, population management (Sörensen & Sevaux 2006) and learning strategies (Kheng et al. 2012). MAs can adopt the strength of other optimisation algorithms by combining them within the same framework, which can provide better performance and overcome the weakness of other algorithms. MAs comprise evolutionary phases that gained its success in complex optimisation problems (Lin et al. 2015; Rezapoor Mirsaleh & Reza Meybodi 2016; Sabar et al. 2013). More specifically, mutation, improvement and restart phases are effectively responsible for the effectiveness of a MAs performance (Krasnogor et al. 2006; Li et al. 2014). The differential evolution (DE) algorithm can be hybridised with the MA in the mutation phase, where DE offers a superior mutation performance across many combinatorial and continuous domains' problems (Sabar et al. 2017).

However, the DE algorithm is subject to stagnation problems (Neri & Tirronen 2010; Chunmei Zhang et al. 2013). Many researchers tried to use the adaptation approach with the DE mutation operator, where two trends were mainly focusing in the control parameter adaptation strategy (Venkatakrishnan et al. 2018) and adaptive strategy control (Wang et al. 2016). The mutation strategy capable to guide the search process to global optimum (Tanabe & Fukunaga 2013). Therefore, global and local mutation operators can balance between the global and local search throughout the evolutionary processes.

Recently, many clustering algorithms have been proposed in the literature (Abualigah et al. 2017; Han et al. 2017; İnkaya et al. 2015; Mustafi et al. 2017; Niu et al. 2017), where the existing clustering criterion that has been used in these algorithms notably affected the quality of the final solutions (Das et al. 2009; Garza-Fabre et al. 2017a; Jain 2010; Maulik et al. 2011). Therefore, the quality of the obtained cluster solutions depends on the selection of suitable clustering criterion (Garcia-Piquer et al. 2017; Martínez-Peñaloza et al. 2017; Mukhopadhyay et al. 2014, 2015). Many clustering algorithms that were introduced recently focused on employing the single criterion optimisation, which could be inappropriate for cluster characteristics of the vast diversity of datasets (Mukhopadhyay et al. 2014, 2015; Wang 2018; Zhou & Zhu 2018). Moreover, the single objective (mono-objective) clustering algorithms are in practice fail to find good data clustering solutions in such datasets (Mukhopadhyay et

al. 2014, 2015; Wang 2018; Zhou & Zhu 2018). The majority of real-life economics, engineering, computing, or management sciences optimisation problems are considered as multi-objective problems, which should be solved using more than one conflicting objectives be minimised or/and maximised (Maulik et al. 2011; Talbi 2009). Moreover, the data clustering as one task of data mining is considered as a multi-objective optimisation problem (Das et al. 2009; Maulik et al. 2011; Mukhopadhyay et al. 2015). Therefore, the multi-objective metaheuristic approach may be applicable to find optimal clustering solutions, by maximizing or/and minimizing more than one objective functions for different types of real-life and synthetic two-dimensional datasets with different cluster shapes and characteristics (Das et al. 2009; Maulik et al. 2011; Mukhopadhyay et al. 2015).

The recent data clustering algorithms, which are based on the multi-objective metaheuristic, were successful in finding good solutions for some datasets, whilst they failed to provide good results across all datasets, shown unstable results or caused an insufficient diversity (Aggarwal & Reddy 2013; Celebi 2015; Figueiredo et al. 2019; Prakash & Singh 2017). These deficiencies might occur because of the unbalanced mechanisms between exploration and exploitation capabilities in the data clustering algorithms, which might cause weak convergence, stagnation, or weak diversity in the optimal Pareto-front solutions (Bouyer & Hatamlou 2018; Dowlatshahi & Nezamabadi-Pour 2014; Prakash & Singh 2017, 2015).

The data clustering validity measures considered as a significant part in the design of the clustering algorithms. These algorithms experience challenges in recognising data points that are either noise or outlier (Aggarwal 2015). Basically, the outliers appear as data points that fail to follow the general pattern of the majority of points, and can significantly hinder the performance of many data clustering algorithms (Aggarwal 2015). In recent decades, outlier detection techniques have attracted cluster analysis researchers to overcome main deficiencies of recent cluster analysis techniques such as K-means (Gan & Ng 2017), hierarchical clustering (Gagolewski et al. 2016), and density-based clustering (Abid et al. 2017). However, further improvements is needed to tackle the rapid growth of data complexity with the consideration of preserving the accuracy of the clustering algorithm (Aggarwal 2015). Although the

majority of the clustering algorithms attempt to detect outliers during the clustering analysis stage (Aggarwal 2015), few algorithms offer validity measures that can tackle detection of these outliers (De Morsier et al. 2015; Todeschini et al. 2013). Nevertheless, connectivity measure of the cluster can measure level of the connectedness of the neighbour data objects that are located in the same cluster (Handl & Knowles 2004; Kishor et al. 2016; Mukhopadhyay et al. 2015), and may measure the amount of connectedness based on non-reliable data objects that can be a form of outliers (Aggarwal 2015). Therefore, the selection of suitable neighbour data objects mechanism can be modified to exclude such outliers, and consequently improve the performance of the connectivity measure.

1.4 RESEARCH QUESTIONS

In general, finding research questions requires comprehensive reading and knowledge corresponding state of the art. An accurate approach is to start from a research hypothesis to focus on the specific perspectives of the research problem to investigate. In this thesis, we are interested in answering the following research hypothesis:

> *If adaptive DE mutation strategy, local search, multi-objective optimisation, and handling outliers within the clustering criterion can be combined within the MA, then employing these components will enhance the performance of the MA that can work well in solving the data clustering problems for different kinds of datasets with different characteristics.*

Furthermore, the general research question is then identified to get a good indication of the significant gaps in the current data clustering research field. When the gaps have been distinguished, a particular research question will be raised, and the ultimate goal of this thesis is to find the answer for these particular research questions. In this thesis, we are interested in answering the following general research question:

> *Can we develop an enhanced memetic differential evolution optimisation algorithms that work well in solving the data clustering problems for different kinds of datasets with different characteristics?*

Finding the answer to the above research question requires us to conduct a comprehensive literature review regarding existing data clustering problems, MA, DE, and outlier detection techniques. The literature review is provided in the next chapter which gives us a clear indication about the current gaps in the data clustering literature. Based on the research issues that have been addressed in Section 1.3, this thesis attempts to answer the following research questions:

RQ1. Does the hybridization between MA and DE algorithms can balance between the exploration and exploitation and improve the population diversity to solve the data clustering problem?

RQ2. How to further enhance the results obtained from the hybridization between MA and DE algorithms, by employing the adaptive DE mutation strategy and the local search?

RQ3. Can we enhance the performance of the hybrid MA and DE algorithm, by using two conflicting objectives at the same time?

RQ4. Does the multi-objective hybrid MA and DE algorithm can be an appropriate approach for handling different kinds of datasets with different characteristics?

RQ5. Does the performance of the multi-objective hybrid MA and DE algorithm can be enhanced by handling the outliers within the clustering criterion?

1.5 RESEARCH OBJECTIVES

The thesis intends to demonstrate that the memetic differential evolution algorithms can be successful in finding good solutions for data clustering problems. Besides, the research aims to enhance the balance between the exploration and exploitation capabilities and avoid falling into local optima and premature convergence problems. The research also employs multiple criterion optimisation approaches to deal with the vast diversity of datasets. Further, this research uses an outlier detection mechanism within the clustering criterion to enhance the quality of clustering solutions. These goals can be achieved through the following objectives:

RO1. To enhance the balance between the exploration and exploitation and improve the population diversity by hybridization of the MA, adaptive DE mutation strategy, and local search for solving the data clustering problems.

RO2. To enhance the performance of the hybrid MA and DE using the multi-objective approach to solve the data clustering problems for different kinds of datasets with different characteristics.

RO3. To enhance the performance of the multi-objective hybrid MA and DE algorithm by handling the outliers within the clustering criterion.

Table 1.2 provides a summary of the mapping between research questions, objectives, and contributions. The first two research questions are answered in the first objective of the thesis, which can be achieved by an enhanced adaptive memetic differential evolution optimisation algorithms for data clustering problems are investigated in this research.

The second research objective answers the third and fourth research questions, which can be achieved by a multi-objective memetic differential evolution optimisation algorithms for data clustering problems. The third research objective answers the last research question, which can be achieved by an enhanced multi-objective memetic differential evolution algorithm using a modified connectivity validity measure based on outlier detection approach.

Table 1.2 Summary of mapping between research issues, questions, objectives and contributions of this thesis.

Chapter	Research Issue	Research Question	Research Objectives	Contribution
Chapter IV	• Current clustering algorithms are unable to find good solutions across all clustering problem datasets and the results were inconsistent. • MA can handle exploration, while exploitation can be handled by the local search. • Adaptive DE mutation strategy can balance between the global and local search.	• Does the hybridization between MA and DE algorithms can balance between the exploration and exploitation and improve the population diversity to solve the data clustering problem? • How to further enhance the results obtained from the hybridization between MA and DE algorithms, by employing the adaptive DE mutation strategy and the local search?	• To enhance the balance between the exploration and exploitation and improve the population diversity by hybridization of the MA, adaptive DE mutation strategy, and local search for solving the data clustering problems.	• An enhanced memetic differential evolution optimisation algorithm for data clustering problems using adaptive DE mutation strategy, and local search for solving the data clustering problems.
Chapter V	• Single criterion optimisation is inappropriate for cluster characteristics of the vast diversity of datasets. • Recent multi-objective clustering algorithms failed to provide good results across all datasets and shown unstable results.	• Can we enhance the performance of the hybrid MA and DE algorithm, by using two conflicting objectives at the same time? • Does the multi-objective hybrid MA and DE algorithm can be an appropriate approach for handling different kinds of datasets with different characteristics?	• To enhance the performance of the hybrid MA and DE using the multi-objective approach to solve the data clustering problems for different kinds of datasets with different characteristics.	• A multi-objective memetic differential evolution optimisation algorithm to solve the data clustering problems for different kinds of datasets with different characteristics.
Chapter VI	• Recent multi-objective clustering algorithms attempt to detect outliers during the clustering analysis stage but did not offer validity measures that can tackle detection of these outlier.	• Does the performance of the multi-objective hybrid MA and DE algorithm can be enhanced by handling the outliers within the clustering criterion?	• To enhance the performance of the multi-objective hybrid MA and DE algorithm by handling the outliers within the clustering criterion.	• An enhanced multi-objective memetic differential evolution algorithm using an outliers detection mechanism within the clustering criterion.

1.6 EXPECTED CONTRIBUTIONS

The following contributions are planned in order to achieve the objectives:

1. An enhanced memetic differential evolution optimisation algorithm for data clustering problems using adaptive DE mutation strategy, and local search for solving the data clustering problems.

2. A multi-objective memetic differential evolution optimisation algorithm to solve the data clustering problems for different kinds of datasets with different characteristics.

3. An enhanced multi-objective memetic differential evolution algorithm using an outliers detection mechanism within the clustering criterion.

1.7 RESEARCH SCOPE

This research is focused on developing appropriate metaheuristic algorithms to achieve high quality of polythetic and hard data clustering solutions. Additionally, the research aims to improve the balance between the exploration and exploitation capabilities of the search and prevent the algorithm from falling into local optima and premature convergence. Furthermore, the research is focused on adopting the multiple criterion optimisation approaches, which could be suitable for the vast diversity of datasets with different cluster characteristics. The research employs an outlier detection mechanism within the clustering criterion to enhance the quality of clustering solutions of the proposed approach.

In this work, twelve real-life benchmark datasets from the UCI machine learning repository (Dheeru & Karra Taniskidou 2017), and fourteen synthetic two-dimensional benchmark datasets from (Bandyopadhyay & Pal 2007; Franti & Sieranoja 2018; Fu & Medico 2007; Gionis et al. 2005; Handl & Knowles 2004, 2007; Jain & Law 2005; Veenman et al. 2002) are used as standard datasets to evaluate the performance of the proposed approaches. Many researchers in the literature have used these datasets such as (Abualigah et al. 2017; Bouyer & Hatamlou 2018; Handl & Knowles 2012; Hatamlou 2013; Jiang et al. 2013; Kishor et al. 2016; Li & Liu 2017;

Prakash & Singh 2015; Wang 2018; Zhou & Zhu 2018). Note that Chapter III presents further details about these datasets. The proposed approaches performance has been evaluated and compared against available results of different approaches available in the literature, based on the same datasets and using the same clustering criterion. Moreover, the algorithm performance is evaluated based on the F-measure function of the obtained results to evaluate the accuracy of the obtained clustering. The quality of the solutions produced by the multiple criterion approaches additionally evaluated using performance assessment matrices such as convergence, diversity, coverage, and overall non-dominant vector generation. Statistical tests are then performed to identify the significant difference in the results obtained from the proposed approaches. Note that Chapter III presents further details about these performance evaluation measures.

1.8 OVERVIEW OF THE THESIS ORGANIZATION

This thesis contains seven chapters, including the current chapter that contains the introduction. Chapter I presents the background, problem statement, research questions, research objective, expected contributions, and research scope. The remainder of this thesis is organised as follows:

In data clustering problems, the literature review of the related research objectives is given in Chapter II. The chapter begins with essential definitions and background to data clustering methods. The related studies on traditional data clustering are then reviewed, and the recent metaheuristic based data clustering approaches are reviewed and analysed. Thus, the chapter presents the background of MA and DE optimisation algorithms, and reviews and analyses the related studies. The chapter also provides the background and the related studies of the multi-objective metaheuristics for data clustering problems. Finally, the chapter ends with primary definitions and background to outlier detection techniques.

The methodology of the research is demonstrated in chapter III. In this chapter, the research framework and different phases are presented to achieve the research objectives. The chapter then explains the datasets and the evaluation criteria used in the research and ends with an overview of the proposed and applied methods as presented at the end of Chapter III.

Chapter IV contains the details of the proposed memetic differential evolution algorithm for solving data clustering problems. In this chapter, solution representation, objective function and details of the proposed algorithm are explained. The results of the method are analysed and compared with recent data clustering techniques in the literature using benchmark datasets. Finally, statistical analysis tests are performed to find the significant difference between the results obtained from the proposed approach when compared with current techniques.

Chapter V introduces a memetic differential evolution algorithm that is based on the multi-objective approach for solving data clustering, to improve the search for optimal clustering by maximising or/and minimising two cluster quality measures. In this chapter, the performance of the proposed algorithm is further analysed using performance assessment matrices, to evaluate the performance of the multi-objective approach. Finally, statistical analysis is performed to find the significant difference between the results obtained from the proposed approach compared data clustering techniques in the literature.

Chapter VI proposes an outlier detection mechanism within connectivity validity measure to improve the clustering solutions quality obtained by the multi-objective memetic differential evolution algorithm. In this chapter, the performance of the proposed mechanism is compared with the current connectivity validity measure using different algorithms.

Chapter VII presents the summary and conclusions of the work, contributions, and research directions for future work.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter offers the background and review of the current state of the data clustering area by discussing the breadth and depth of work that has been done so far, and it can identify the gaps where this thesis is focused. The chapter begins with the background and related work of the data clustering problems. Also, the chapter reviews the relevant metaheuristic-based methods in data clustering. The chapter also discusses the MA and DE optimisation algorithms, and reviews and analyses the related studies. Additionally, the chapter provides the background and the related studies of the multi-objective metaheuristics for data clustering problems. Finally, the chapter ends with primary definitions, background, and reviews to outlier detection techniques. This chapter demonstrates reasons on why this thesis framed the specific interests and the research question and show how related research have influenced the proposed methods.

2.2 DATA CLUSTERING

The rapid development of Information Technology has increasingly generated large data in various areas and industries. This data may contain medical information, shopping habits, and criminal records, thus, it has given attention to the storing and manipulating approach of the data for knowledge discovery, and mainly for decision making. Data mining is an essential component in knowledge discovery, and it can extract knowledge and useful information by finding patterns from the vast amounts of the raw data. Data clustering is a common task that aims to identify a finite set of groups or cluster to find a proper description for the data. It usually used when both class labels and the number of classes are not known in advance. The clustering problem has been

discussed extensively, although there is no uniform definition for data clustering. Several researchers in the field of data mining give a different description for data clustering. Some researchers such as Arabie et al. (1996) defined clustering as those methods concerned in some ways with the identification of homogeneous groups of objects. Everitt et al. (2011) provided other definitions as a set of entities that are alike, and entities from different clusters are not alike.

Data clustering also is known as cluster analysis, taxonomy analysis, segmentation analysis, unsupervised classification, or Q-analysis, which is widely used in different applications to understand the structure of the data deeply, to focus on a specific set of clusters for further analysis, and to detect the characteristics of each cluster. It has been enhanced and utilised as an essential tool for different disciplines and fields such as Information Retrieval (Wu et al. 2013), Internet of Things (Abbasi & Younis 2007), Business (Müller & Hamm 2014), Medicine (Esfandiari et al. 2014; Nahar et al. 2013b), Image segmentation (Twinkle Gupta & Dharmender Kumar 2014), and Climate (Dowd et al. 2017). Numerous books have been published on data clustering, such as those by (Gan et al. 2007), (Everitt et al. 2011b), (Aggarwal & Reddy 2013), (Celebi 2015), and (Everitt et al. 2011b). Thus, clustering has been examined in the data mining books by (Aggarwal 2015) and (Tan et al. 2006), as well as in machine learning books by (Bishop 2006) and (Lantz 2013). Articles on cluster analysis can be published in a wide range of 75 technical journals and more than 45 conferences related to data clustering (Gan et al. 2007). These journals and conferences from diverse fields of knowledge like Computing, Statistics, Bioinformatics, Bioinformatics, Marketing, Knowledge Discovery, Image Processing, Machine Learning, and Operations Research.

2.2.1 Applications of Data Clustering

Data clustering methods have been employed in a vast number of applications and disciplines. Some disciplines and fields utilised clustering are shown below:

- *Information Retrieval*: Clustering methods have been used in diverse applications in information retrieval like clustering Massive dataset, finding topics in the collection of documents, and textual document indexing (Wu et al. 2013).

- *Internet of Things*: Some studies on clustering for the IoT focus on partitioning the incoming patterns, which are based on predefined proximity measures, into three different groups; a set of unlabelled input patterns that can consist of various data, such as the behaviour of the user captured by the sensors. Another focus is to find out the behaviour of the user that provides the user needs for the services. Thus, clustering techniques used for distributed clustering which is the essential demand for the wireless sensor network (Abbasi & Younis 2007; Tsai et al. 2014).

- *Business*: Businesses Collect vast volumes of knowledge on existing and prospective customers. Customers can be segmented into small groups to perform additional analysis (Müller & Hamm 2014).

- *Climate*: Recognising the global climate demands detecting patterns in the oceans and atmosphere. Therefore, data clustering aims to see patterns in the atmospheric pressure that has a significant influence on land climate (Dowd et al. 2017; Steinbach et al. 2003).

- *Medicine*: Cluster analysis is used to distinguish the various subcategories of diseases. It can further be employed to detect patterns in the temporal or spatial disease distribution (Esfandiari et al. 2014; Nahar et al. 2013b, 2013a).

- *Image Segmentation*: cluster analysis is employed to discover the edges or borders of the objects in the images (Gonzalez & Woods 2002; Kumar et al. 2014a; Pratt 2000).

- *Gene Expression and biology*: In Gene Expression, the gene expression data characteristics become meaningful when clustering both genes and samples. It can be grouped based on their expression patterns into clusters (Kerr et al. 2008).

2.2.2 Types of clusters

The clusters can be categorised based on their shape into the following types (Tan et al. 2006):

1) *Well-Separated cluster*: This clusters type contains a set of data points where every point is nearby to all other points within the cluster than other points outside the cluster.

2) *Prototype-Based cluster*: This cluster type consists of a set of data points in which every point is nearby the cluster prototype than other clusters prototypes. The centroid or the medoid are often representing the prototype of a cluster.

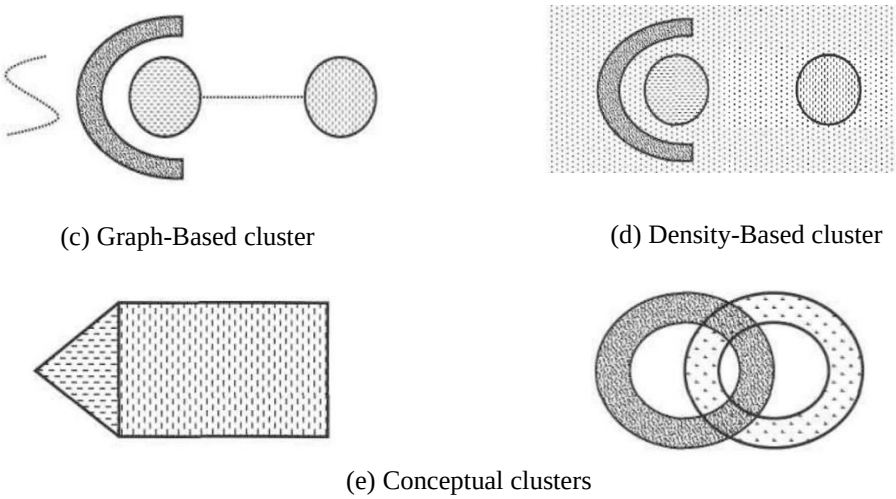
3) *Graph-Based cluster*: This cluster type can be distinguished as connected points, but that has no connection to points outside the cluster.

4) *Density-Based cluster*: This cluster type is a dense points region that is surrounded by low-density regions. The density-based cluster is used with the datasets that consist of noise, random shapes, or outliers.

5) *Conceptual Cluster*: This cluster type can be described as a set of data points are sharing common characteristics and offers a consistent shape such as triangle, ellipse, or spiral.

Figure 2.1 illustrates various cluster types that are represented by sets of two-dimensional data points (Tan et al. 2006).

(a) Well-separated clusters (b) Prototype-Based cluster



(c) Graph-Based cluster (d) Density-Based cluster

(e) Conceptual clusters

Figure 2.1 Different types of clusters as illustrated by sets of two-dimensional points

2.2.3 Data Types in data clustering

Data clustering methods are strongly correlated with the data types within the dataset. Accordingly, understanding normalisation, proximity, and scale are important in investigating the results of clustering methods. (Aggarwal & Reddy 2013; Das et al. 2009; Gan et al. 2007), a particular attribute can be categorised as discrete, continuous or binary. Binary attributes have exactly two values, and the discrete attributes consist of values a finite number of values. Thus, the continuous types include an infinite number of values. Furthermore, one dataset may hold various types of data such as categorical and numerical, which depends upon the discipline of the dataset, for example, the DNA data is related to biology; also time series data is relevant to the finance or weather forecasting. Consequently, this thesis is concerned with investigating data clustering of the continuous numerical data type.

2.2.4 Taxonomies of data clustering methods

The clustering methods classification is based on several independent perspectives such as various methodologies, beginning points, the criteria of clustering, the algorithmic viewpoints, and the output representation. The clustering algorithms properties are expressed as the following (Tan et al. 2006):

– *Monothetic and Polythetic Clustering*: When attributes are respectively or simultaneously utilised in clustering algorithms, issues such as monothetic and polythetic may occur. Usually, algorithms are considered as polythetic, where entire attributes are employed in the calculation of the distances between data points and the result. Nevertheless, the algorithms are considered as monothetic deals with the attributes separately.

– *Hard and Fuzzy Clustering*: In hard clustering, the data points can be attached only to a single cluster, while it can belong to multiple clusters in the fuzzy clustering.

Hence, this thesis is concerned with the polythetic and hard data clustering approaches, which include the calculation of the distances using the entire attributes.

2.2.5 Mathematical formulation of the clustering problem

Data Clustering is a process of partitioning a set of n points into some K clusters, based on a specific similarity measure. The set of n points are represented by the set $X = \{x_1, x_2, \ldots, x_n\}$, the K clusters are denoted by $C = \{C_1, C_2, \ldots, C_K\}$, such that data points in the same clusters are similar, and other data points are dissimilar. In the data clustering problem, clusters must maintain the following three hard constraints (Das et al. 2008):

i. All clusters should not be empty and contain at least one data point:

$$C_i \neq \phi, \forall\, i \in \{1, 2, \ldots, K\}, \tag{2.1}$$

ii. Different clusters should not have data points in common:

$$C_i \cap C_j = \phi, \forall\, i \neq j \; and \; i, j \in \{1, 2, \ldots, K\}, \tag{2.2}$$

iii. All data points should be contained in a cluster:

$$\bigcup_{i=1}^{k} C_i = X \tag{2.3}$$

It is important to mention that data clustering problems are entirely associated with data types of the data points; therefore, understanding normalisation, proximity, and scale is essential in interpreting the results of clustering algorithms (Gan et al. 2007). Moreover, an adequate partitioning is influenced by determining a suitable fitness function with the similarity/dissimilarity measure. The Euclidean distance measure is one of the regularly chosen similarity measures in data clustering problems. Hence, The data clustering problem look for an optimal clusters C^* concerning complete feasible solutions set $C^*=\{C^1, C^2, \ldots, C^{N(n,K)}\}$ such that $C_i \neq C_j$, $i \neq j$. The number of feasible clusters *N(n, K) is* given by Equation 2.4:

$$\mathrm{N(n,K)} = \frac{1}{K!}\sum_{i=1}^{K}(-1)^{k-i}\binom{k}{i}(i)^n \tag{2.4}$$

The data clustering problem can be expressed by Equation 2.5:

$$\underset{C}{\boldsymbol{Optimize}}\ f(X,C) \tag{2.5}$$

The *f (X, C)* denotes the fitness function to evaluate the quality of clusters produced from the clustering algorithm. Accordingly, the fitness function can be maximised or minimised depending on the similarity/dissimilarity measure used. Further similarity/dissimilarity measures will be discussed in Section 2.2.6. The mathematical representation of multi-objective data clustering problems with *M*-objectives is given in Equation 2.6 (Maulik et al. 2011):

$$\boldsymbol{Optimize}\ \ f(X,C) = \big(f_1(X,C), f_2(X,C), \ldots\ f_M(X,C)\big) \tag{2.6}$$

$$\boldsymbol{subject\ to}\ \begin{cases} g_i(X,C) \leq 0, i = 1,2,\ldots,p, \\ h_j(X,C) = 0, j = 1,2,\ldots,q \end{cases}$$

Where $g_i(X,C)$ indicates the p inequality constraints, and $h_j(X,C)$ indicates the q equality constraints. Assuming that $v = [v_1, v_2, \ldots, v_d]$ and $u = [u_1, u_2, \ldots, u_d]$ in the

dimension space of size d, if: $\forall i \in \{1, \ldots, d\}$, $F(u_i) \leq F(v_i) \wedge \exists j \in \{1, \ldots, d\}$, $F(u_j) < F(v_j)$, then u Pareto dominates v, and it can denoted as $u \prec v$. The set $x*$ can be called non-dominated solution (Pareto-optimal), if there no vector x such that $x \prec x*$. The set named as Pareto-front P_F if consisted of all objective function values of $x*$ (Deb et al. 2002). Figure 2.2 illustrates an example of non-dominated solutions. The figure shows seven solutions [*A*, *B*, *C*, *D*, *E*, *F*, *G*] with two objectives [$Objective_i$, $Objective_j$] that are illustrated in the objective space (Talbi 2009). The set *P** = [*A*, *B*, *C*, *D*] is considered as an optimal pareto set because it is not dominated (solution x_1 is strictly better than solution x_2 in at least one objective) by any solution in the non-optimal pareto set *P* = [*E*, *F*, *G*].

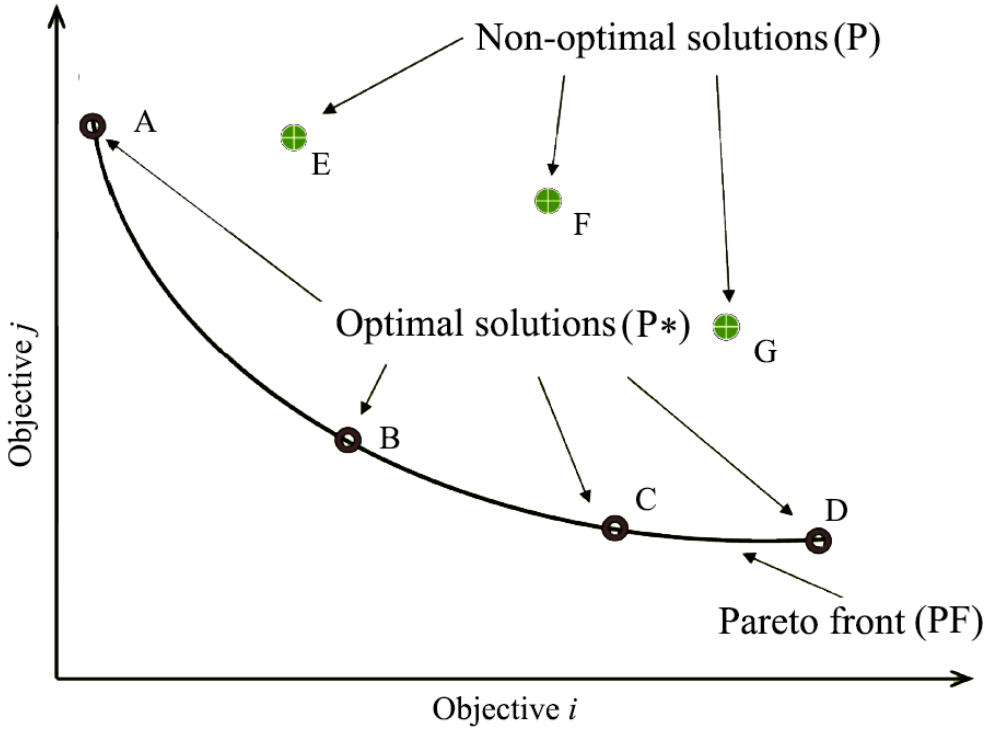


Figure 2.2 Example of non-dominated solutions of the optimal pareto set

2.2.6 Similarity/dissimilarity measures in data clustering algorithms

Similarity/dissimilarity measures recognised as one of the primary components when similar data points are grouped, where they can identify related points. The similarity distance functions, also known as dissimilarity functions, are investigated in several comprehensive reviews such as (Aggarwal & Reddy 2013; Everitt et al. 2011b; Gan et

al. 2007; Tan et al. 2006). The computation of the similarity between two points is achieved by particular distance functions, where these two data points are represented by two data vectors O_i and O_j in the d-dimensional space.

One of most popular similarity metric that were discussed in the literature is the Minkowski distances, which are general class of distance functions that are defined in Equation 2.7. Minkowski distance is one of the most widely used distances in the partitional clustering.

$$\text{Minkowski distance } (O_i, O_j) = \left(\sum_{m=1}^{d} \left(O_i^m - O_j^m \right)^r \right)^{1/r} \tag{2.7}$$

Where the value of r index represents the infinite number of distances. Two distance functions Manhattan distance (r=1) and Euclidean distance (r=2) are examples of Minkowski distance, as shown in Equations 2.8 and 2.9 respectively:

$$\text{Manhattan distance } (O_i, O_j) = \sum_{m=1}^{d} \left| O_i^m - O_j^m \right| \tag{2.8}$$

$$\text{Euclidean distance } (O_i, O_j) = \sqrt{\sum_{m=1}^{d} \left(O_i^m - O_j^m \right)^2} \tag{2.9}$$

Several clustering algorithms employ various distance functions depending on the type of data of the data points (as discussed in section 2.2.3) to determine the precise similarity between two points, or among two clusters. For example, partitioning clustering algorithms use Euclidean distance to measure the similarity of the continuous numerical data types. Consequently, this thesis employs the Euclidean distance as the similarity/dissimilarity measures (Further details about the objective function is mentioned in Section 3.2.3).

2.2.7 Criteria for clustering evaluation

Cluster evaluation also is known as cluster validation, is associated with the assessment steps of data clustering outcomes to discover partitioning. Cluster validity can be employed to determine the number of clusters and identifies the corresponding best partition. Thus, this section explains the fundamental background in this field and demonstrate various cluster validity methods offered in the literature. Accordingly, the validity index should consider the following perspectives by the partitioning methods:

- *Cohesion*: Patterns inside a particulate cluster must maintain similarity with each other as much as possible. The patterns fitness variation within a cluster indicates the compactness of the cluster.
- *Separation*: Distance between the clusters' centres indicates the separation of the clusters. Well-separated clusters reveal the efficient performance of the clustering method.

The results of several clustering methods need to be investigated. As a consequence, the best clustering method should be chosen by employing useful quality measures to demonstrate the clusters effectiveness. Cluster validity usually concerned with two perspectives. In the first perspective, the clusters quality is computed according to the homogeneity inside the clusters. The data points related to a particular cluster are mostly similar than other points associated with a different cluster. Therefore, cluster validity has been categorised into internal quality measures, which are also known as "*internal evaluation functions*" or "*unsupervised cluster evaluation*", are utilised to assess various cluster sets without any external knowledge. In the second category, external quality measures use the correct class labels that are determined based on the external knowledge provided by the experts. The external knowledge is employed to measure the correspondences between the obtained cluster and correct classification. The external quality measure is also known as "*external quality measures*", "*external evaluation functions*", or "*supervised cluster evaluation*".

2.2.8 Internal Evaluation Functions

Internal evaluation functions determine the clusters structure quality without any external knowledge or information. This subsection presents the internal validity measures that have been adopted in this thesis, and explained in the literature (Aggarwal & Reddy 2013; Das et al. 2009; Gan et al. 2007; Maulik et al. 2011; Tan et al. 2006).

a. Intra-Cluster Distance

The intra-cluster distance is one of the essential internal measures broadly applied in the estimation of the quality of the clustering solutions. (Aggarwal & Reddy 2013), the mathematical formulation of the intra-cluster distance is defined in Equation 2.10, Where $d(O_i, Z_l)$ defines the distance separating point O_i and cluster centre Z_l. Several functions can be employed to compute the distance between points in the data clustering problem such as Euclidean distance and Manhattan distance.

$$f(C,O) = \sum_{l=1}^{k} \sum_{Oi \in Cl}^{n} d(O_i, Z_l) \tag{2.10}$$

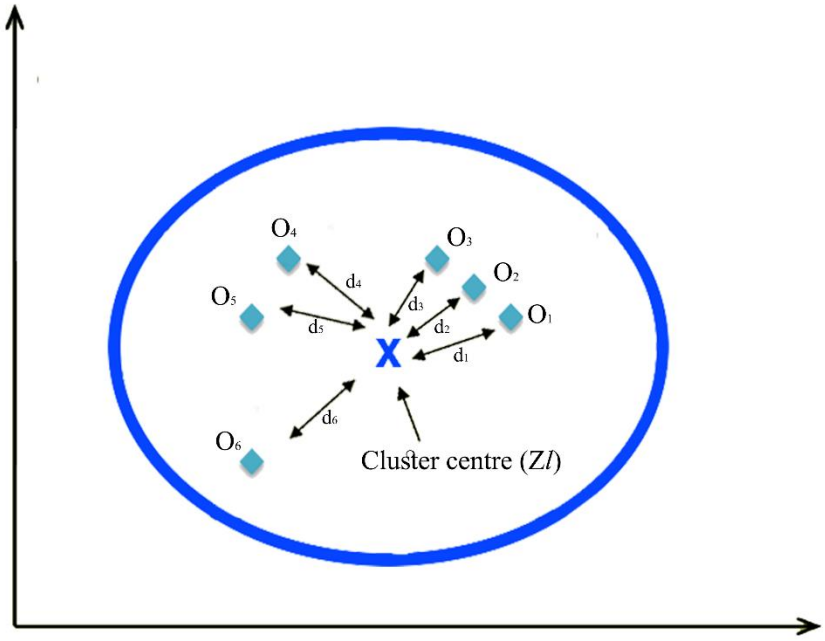


Figure 2.3 Example of the intra-cluster distance internal measures computation

Figure 2.3 presents an example of the intra-cluster distance compution of six data points group in a cluster with cluster centre (Z_l). The intra-cluster distance is computed by the summation of all distances between each data point and the cluster centre. According to Figure 2.3, the intra-cluster distance $F(C,O) = d_1 + d_2 + d_3 + d_4 + d_5 + d_6$.

The Euclidean distance (as shown in Equation 2.9) is an example of the traditional applied distance functions. Furthermore, the cluster centre Z_l is calculated by determining the average value for the entire data points related to the cluster, as shown in Equation 2.11. The n_l denotes the number of data points related to cluster centre Z_l. Equation 2.11 is used in this research to initialise and compute the cluster centres of the population (see Section 3.2.3 for more details).

$$Z_l = \frac{1}{n_l} \sum_{\forall O_i \in Z_l} (O_i) \tag{2.11}$$

b. Connectivity of Cluster

The connectivity of the cluster (Kishor et al. 2016; Mukhopadhyay et al. 2015) is used to measure the amount of the neighbour data points that are located in the same cluster and should be minimised. The mathematical formulation of the connectivity of the cluster is shown in Equations 2.12 and 2.13. Where N denotes the number of data points, and L defines the number of neighbours that contribute to the connectivity measure.

$$connectivity(C) = \sum_{i=1}^{N} \sum_{j=1}^{M} nn_i(j) \tag{2.12}$$

where

$$nn_i(j) = \begin{cases} \frac{1}{j}, & \text{if } point\ i\ is\ not\ in\ the\ same\ cluster\ of\ point\ j\ , \\ 0\ , & otherwise. \end{cases} \tag{2.13}$$

Figure 2.4 demonstrates the flowchart of computing the cluster connectivity. The computation starts by finding L nearest neighbourhood data point of each data point in the dataset (see Section 3.2.2 for more details). The computation is then iterates N

for each data points *i* in the dataset. Next it gives a penalty of 1 / *j* (as shown in Equation 2.13) for each *M* data points that reside outside the cluster of *i*.

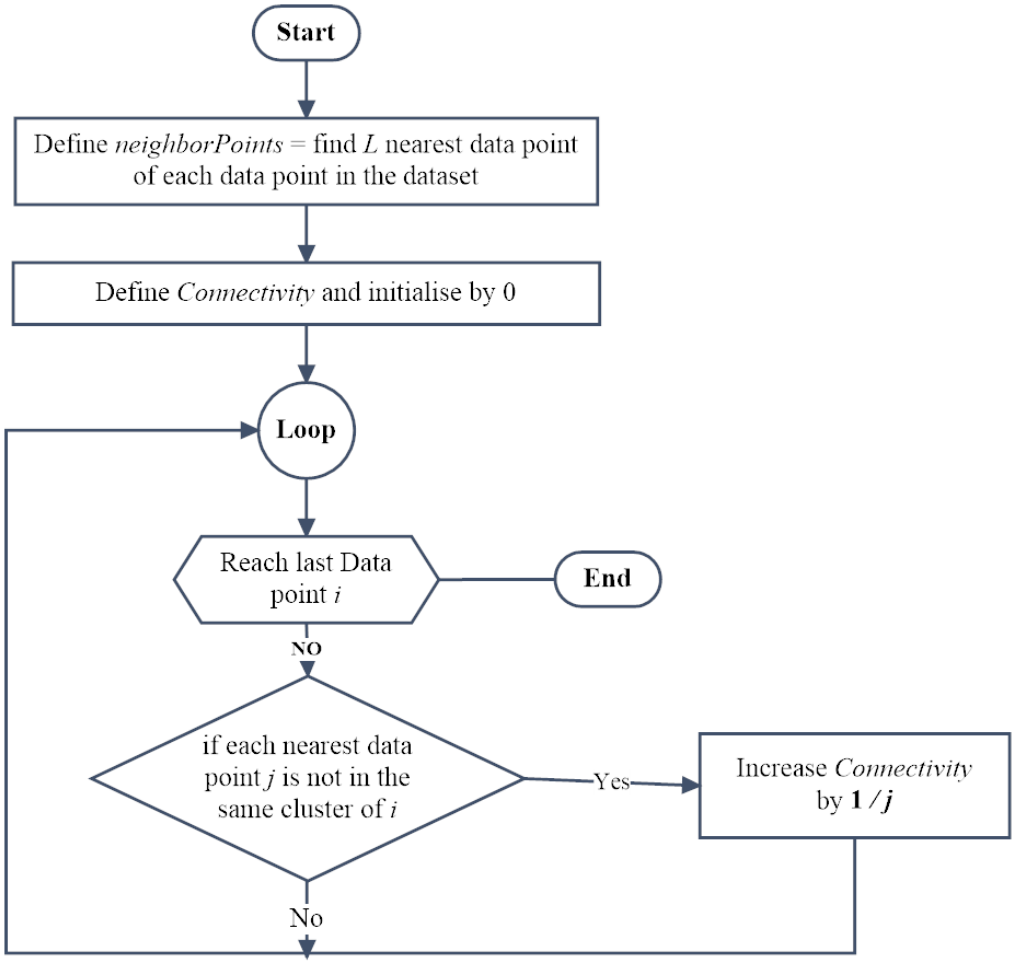


Figure 2.4 Flowchart of the computation of the cluster connectivity

Figure 2.5 presents an example of computing the cluster connectivity. The figure shows six neighbourhood data points {O_1, O_2, O_3, O_4, O_5, O_6} of the data point O_i. The calculation gives a penalty of 1 / *j* for data points {O_1, O_2, O_3} that reside outside the cluster (*Cluster 1*) that O_i belongs. The computation does not give any penalty for data points { O_4, O_5, O_6} that reside within the same cluster of O_i.

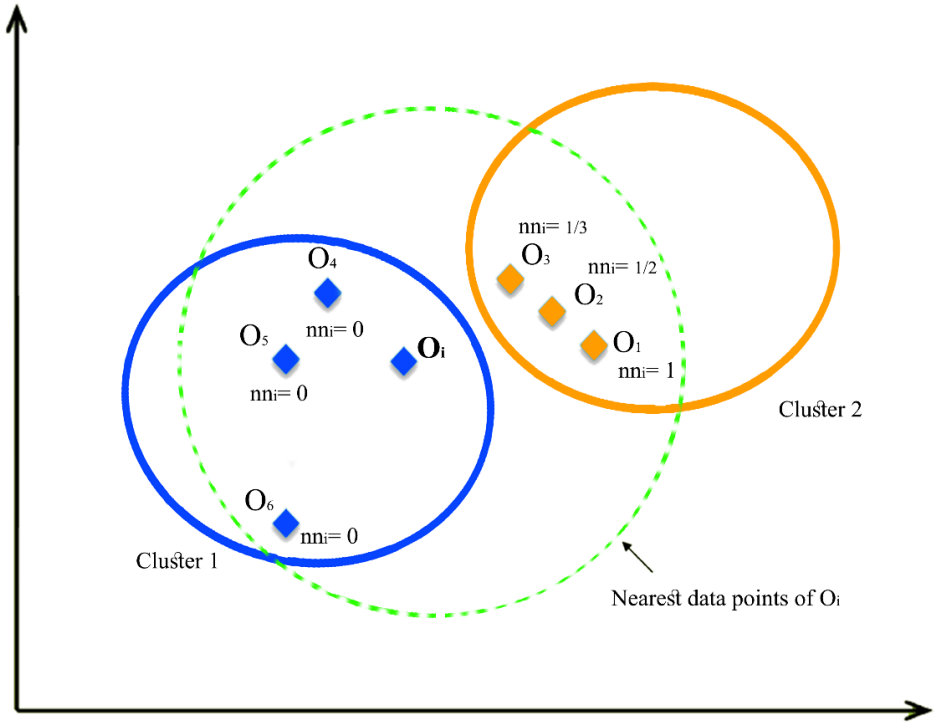


Figure 2.5 Example of the computation of the connectivity of the cluster

2.2.9 External Evaluation Functions

An external knowledge provided by the experts is adopted to estimate the level of correspondence between the class labels and the cluster labels. Cluster analysis employs methods from classification, such as F-measure and accuracy to assess each cluster validity in this thesis. The details regarding other external quality measures are further discussed in (Aggarwal & Reddy 2013; Gan et al. 2007; Tan et al. 2006).

a. F-measure

The F-measure is an external measure that compares the ground truth with the obtained clusters to calculate the similarity between them. The high percentage of the F-measure value indicates a better the clustering quality. The precision and recall of cluster S_j, and class R_i, i, j=1, 2, ..., k are shown in Equations 2.14 and 2.15, Where $|R_i|$ is the number of points in class R_i, and $|S_j|$ is the number of data points in cluster S_j, and L_{ij} is the number of data points of class R_i in cluster S_j. The F-measure of a class R_i is defined in equation 2.16. The overall F-measure is the weighted average of all classes is given in Equation 2.17 :

$$precision\ (R_i, S_j) = \frac{L_{ij}}{|S_j|} \tag{2.14}$$

$$recall\ (R_i, S_j) = \frac{L_{ij}}{|R_j|} \tag{2.15}$$

$$F\ (R_i) = \frac{2 \times\ precision\ (R_i, S_j) \times recall\ (R_i, S_j)}{precision\ (R_i, S_j) + recall\ (R_i, S_j)} \tag{2.16}$$

$$F - measure(k) = \frac{\sum_{i=0}^{k-1}(|R_i| \times F\ (R_i))}{\sum_{i=0}^{k-1}|R_i|} \tag{2.17}$$

b. Accuracy

The accuracy is an external measure that indicates the proportionate number of data points that correctly placed by the predictive model to match the class (ground truth) in the dataset. The mathematical formulation of the accuracy measure is shown in Equation 2.18:

$$Accuracy\ (k) = \frac{number\ of\ correct\ data\ objects\ identified}{total\ number\ of\ data\ Objects} \tag{2.18}$$

2.2.10 Review of traditional data clustering Algorithms

Generally, data clustering methods are classified into six clustering algorithms types that include partitioning, hierarchical, density-based, grid-based, subspace clustering, and metaheuristic clustering (Celebi 2015). Thus, the first five clustering methods are considered traditional clustering methods. Figure 2.6 demonstrates the main data clustering types with their popular methods. Briefly, partitioning methods attempt to divide the dataset directly into a set of disjoint *K* clusters, while the hierarchical methods attempt to construct a hierarchy of clusters. The density-based algorithms are concerned with clusters being dense areas of data points in the data space. The grid-based methods attempt to divide the possible number of values of each attribute into some contiguous intervals. At last the subspaces methods attempt to detect clusters by utilising the attributes.

c. Hierarchical Clustering

The hierarchical clustering seeks to partition the data by a sequence of partitions, which could be applied from only a single cluster containing the entire data points to *k* clusters, where each cluster includes a single data point. The hierarchical clustering is categorised into agglomerative or divisive clustering. In the agglomerative hierarchical clustering, the algorithm starts with single clusters, and at each iteration, it merges every two smallest distant clusters, where the number of clusters is decreased by one. There are three techniques employed for computing the distance among clusters: Single, average and complete linkage agglomerative algorithms (Tan et al. 2006). In Divisive hierarchical clustering follows the reverse process, it starts from a single cluster containing all the points. Each step, the largest cluster is divided into two clusters until the target number of clusters is achieved (Das et al. 2009; Tan et al. 2006).

The most popular hierarchical clustering algorithms are divisive analysis (DIANA) (Kaufman & Rousseeuw 1990) that follow a top-down strategy algorithm that operates as the reverse of agglomerative hierarchical clustering by beginning with whole points in one cluster; then it breaks the clusters into smaller portions of points until every object forms a cluster or when satisfying a termination condition. Agglomerative Nesting (AGNES) (Kaufman & Rousseeuw 1990) follows bottom-up strategy algorithm that begins by putting a point in a single cluster, and then it starts joining these atomic clusters into larger clusters until whole points joined in a single cluster, or when termination conditions are reached. Clustering using Representatives (CURE) (Guha et al. 2001; Guha et al. 1998) describes a cluster by points produced by choosing well-scattered points, then it shrinks them near the cluster centroid by a particular fraction. The Hierarchical Clustering with Dynamic Modelling (CHAMELEON) (Karypis et al. 1999) merges an initial partitioning of the data, using an effective graph algorithm, with a novel hierarchical clustering scheme that employs the concepts of closeness and interconnectivity, concurrently with the local modelling of clusters. The balanced iterative reducing and clustering using hierarchies (BIRCH) (Zhang et al. 1997; Zhang et al. 1996) utilised a hierarchical data structure called CFtree for the incremental and the dynamic segmenting of the incoming data points.

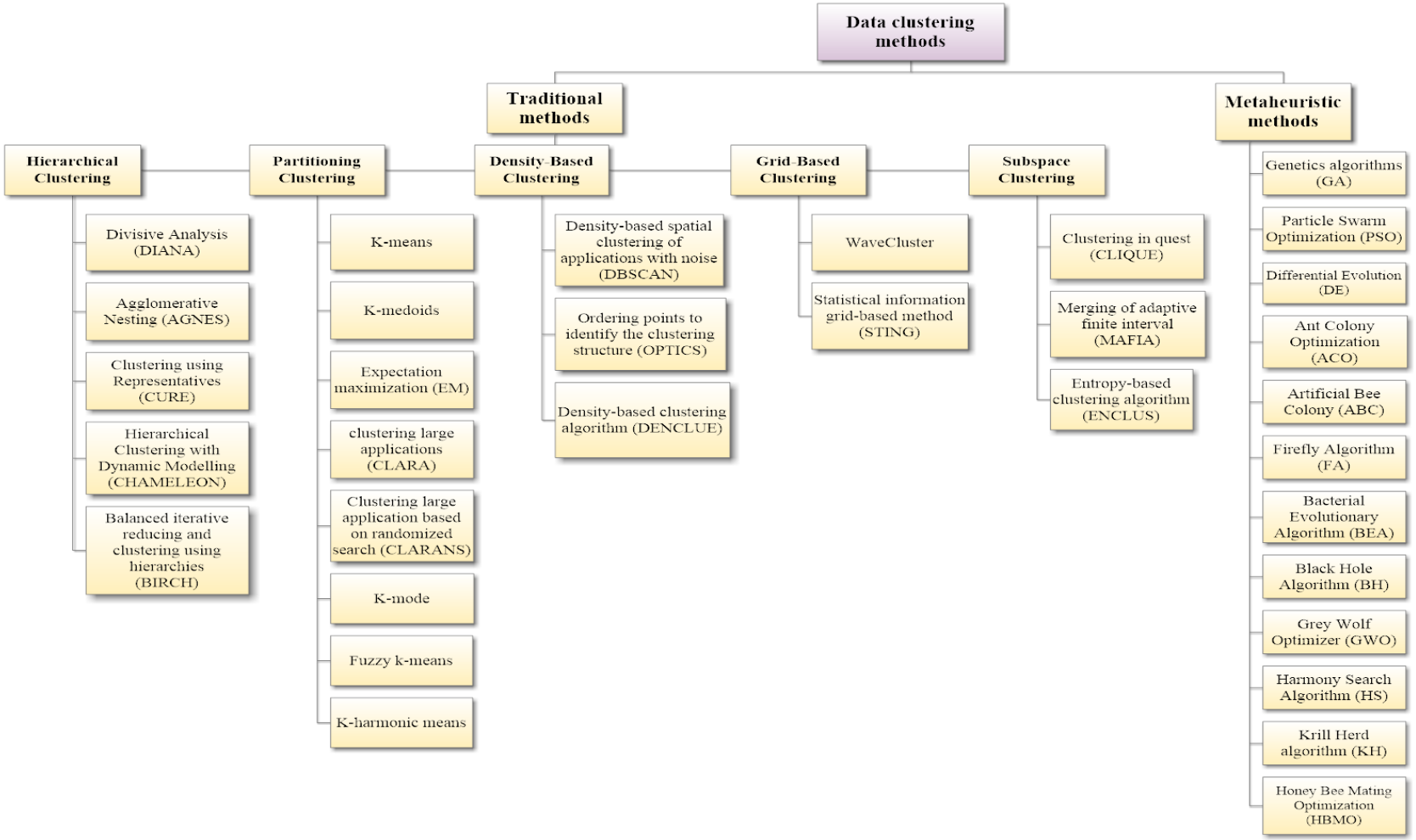


Figure 2.6 Main data clustering types with their popular methods

d. Partitional data clustering

The partitional data clustering methods try to split the dataset into a set of disjointed clusters. They seek to optimise a particular criterion function. The partitional methods identify clusters with convex shapes. However, they are not a good choice for discovering clusters of arbitrary shapes. The most popular partitional clustering algorithms are K-means, expectation maximisation, K-medoids, clustering large application based on randomised search, and clustering large applications.

i) K-means

The k-means algorithm is one of the most widely used partitional clustering algorithms (Forgy 1965; Jain 2010; Tan et al. 2006). The k-means algorithm assumes that the number of clusters k is fixed. The algorithm allocates the data points to the nearest clusters and next it keeps adjusting the cluster's membership concerning the distance function. The algorithm repeats this process until no significant change in distance function, or there are no longer changes in the membership of the clusters. K-means algorithm tries to minimise the intra-cluster distance that is defined as in Equation 2.19. Where the distance between cluster centre Z_l and object O_i is defined by $d(O_i, Z_l)$. Nevertheless, the k-means performance is highly sensitive to the initial centroids of clusters and may be trapped in the local optimal solution (Forgy 1965; Gan et al. 2007; Jain 2010; Jain & Dubes 1988).

$$f(O, C) = \sum_{l=1}^{k} \sum_{Oi \in Cl}^{n} d(O_i, Z_l) \tag{2.19}$$

ii) K-medoids

The partitioning around medoids (PAM) (Celebi 2015; Das et al. 2009; Maulik et al. 2011; Reynolds et al. 2004) is one of the most popular K-medoid algorithms. The PAM algorithm symbolises each cluster individual by one of the corresponding points in the cluster located nearby to the cluster centre. In general, PAM is a not efficient clustering algorithm for large size datasets (Jain 2010). Additionally, PAM is a costly algorithm

while seeking for the medoids, because it compares each data points with the entire dataset.

iii) Expectation-Maximisation

Expectation-Maximization (EM) algorithm is a broadly adopted method in partitional clustering (Das et al. 2009; Jain 2010). The EM groups a dataset into clusters by determining a group of Gaussian probability density functions (PDFs) relevant to the data.

iv) Clustering Large Applications

Clustering Large Applications (CLARA) (Kaufman & Rousseeuw 1990), is the execution of PAM in a subset of the dataset. It selects some samples from the dataset, implements PAM on the samples obtained from the best clustering. CLARA algorithm is inspired by the data sampling technique, where the representative of the data is chosen from a limit part of the real data. Thus, the PAM algorithm is employed to chose proper medoids. The CLARA algorithm can deal with numerous large volume datasets.

v) Clustering Large Applications based on Randomized Search

Clustering Large Applications based on Randomized Search (CLARANS) (Ng & Han 1994), extends the sampling method from the PAM algorithm. Clustering is performed by searching a graph that consists of the entire nodes of the k-medoids set. CLARANS algorithm substitutes a medoid in the obtained cluster that is called the current clustering neighbour.

vi) Other Partitional Clustering Algorithms

There are several other partitional clustering algorithms such as Fuzzy k-means algorithm, k-prototype algorithm, and k-harmonic means. The k-prototype algorithm is also as known k-mode algorithm (Das et al. 2009; Z. Huang 1998; Huang & Ng 2003) is inspired from the k-means algorithm and focuses in clustering the categorical data. The Fuzzy k-means algorithm also called fuzzy c-means (FCM) algorithm (Das et al.

2009; Jain 2010; Nasraoui & Krishnapuram 1996; Wang et al. 2004). FCM utilises the least square error criterion with fuzzy extension. The FCM is better than k-means algorithm when dealing with the overlapping clusters. Similar to k-means, the number of clusters has to be provided. The k-harmonic means (KHM) algorithm (Zhang et al. 1999) computes the harmonic mean of each cluster centre for all the clusters. Compared to k-means, The KHM algorithm considered less sensitive to the initial conditions (Zhang 2003).

e. Density-Based Clustering Algorithms

Generally, the density-based algorithms are concerned with clusters as dense areas of data points, which are surrounded by low-density regions. The density-based approach aims to find high-density and low-density areas based on the search space distribution, where both low-density and high-density regions are separated. The conventional approach is to split these high-dimensional regions into density-based grid units. Examples of density-based clustering methods are density-based spatial clustering of applications with noise (DBSCAN) that begins searching for ε-neighbourhood points, and if it holds at fixed least number of points, the cluster creation will be started. Otherwise, these points are recognised as noise. In the Ordering Points to Identify the Clustering Structure (OPTICS) (Ankerst et al. 1999), the algorithm extends the DBSCAN algorithm by manipulating more various local densities; OPTICS builds an augmented ordering for the data points. Moreover, the density-based clustering algorithm (DENCLUE) offers a method for clustering large-scale multimedia databases. The essential concept of DENCLUE is to analytically represent the entire point's density using the influence functions summation, which demonstrates the influence of the surrounding neighbourhood.

f. Grid-Based Clustering Algorithms

A grid-based clustering algorithm is an efficient method to organise a low dimensional dataset. The concept of grid-based data clustering algorithms is to split the possible number of values of each attribute into some contiguous intervals and creating a set of grid cells which contains the values of the points (Das et al. 2009; Tan et al. 2006). Points can be attached to grid cells in a single pass throughout the data, and the knowledge concerning

every cell like the number of points in the cell can additionally be collected at the same time. There are some techniques to perform clustering using a grid; the most popular approaches is WaveCluster (Sheikholeslami & Zhang 1998) that is inspired by the wavelet transform signal processing methods. The statistical information grid-based method (STING) (Wang et al. 1997) breaks the spatial area into rectangular cells by utilising the hierarchical structure.

g. Subspace Clustering Methods

The previous traditional clustering techniques try to found clusters by using the attributes. Therefore, if only subspaces of the data of the features are considered, then the clusters found can be entirely different from one subspace to another. Some subspace clustering methods such as clustering in a quest (CLIQUE) (Aggarwal & Reddy 2013; Agrawal et al. 1998; Das et al. 2009) aims to find subspace clusters by checking each subspace for clusters. The merging of the adaptive finite interval (MAFIA) (Aggarwal & Reddy 2013; Nagesh et al. 2001) extends CLIQUE algorithm by constructing adaptive grids to improve subspace clustering and also uses parallelism on a shared-nothing architecture to handle massive data sets. The entropy-based clustering algorithm (ENCLUS) (Cheng et al. 1999) extends the adaptive CLIQUE method by using a different approach of entropy-based criterion for subspace selection.

h. Clustering Using Metaheuristics

Although heuristic algorithms use domain knowledge to speed up the convergence that can provide quick solutions, these algorithms can easily be stuck at local optima and cannot be easily employed to solve other clustering problems. Therefore, metaheuristic algorithm can be employed in tackling complex problems by achieving adequate solutions within an appropriate computation time. The metaheuristic algorithms do not ensure finding global optimal solutions but probably can find good solutions. The recent research in the literature showed that some metaheuristic algorithms perform better for a particular type of optimisation problems, which also can perform better across different problem instances. Although other heuristics such as hyper-heuristic approaches can provide an alternative way to integrate multiple (domain blind) heuristic

algorithms into a single search algorithm, in practice, these domain-blind heuristics approach has not produced satisfactory quality of sdolutions (Burke et al. 2019).

The metaheuristic approaches have gained an outstanding reputation over the recent decades. Thus, metaheuristic algorithms were utilised in various problem domains and shown efficient performance and sufficient to solve optimisation problems that either large or complex. The following section discusses the data clustering algorithms related metaheuristics which are classified as single-solution based (S-metaheuristics) and population-based (P-metaheuristics) metaheuristics (Talbi 2009).

Table 2.1 demonstrates the summary of the traditional data clustering types. It presents various algorithms related to these types, and the common characteristics of the algorithms such as the handled types of data, suitable cluster shapes, robustness to the outliers and noise, and the advantages and disadvantages. These related issues have been investigated in the literature such as (Aggarwal 2015; Aggarwal & Reddy 2013; Celebi 2015; Das et al. 2009; Jain 2010; Maulik et al. 2011; Tan et al. 2006).

The traditional partitional clustering methods like K-means employs a greedy search method throughout the search space. These algorithms try to optimise the clusters compactness. Although most traditional clustering algorithms have an efficient computational time and straightforward implementation, they experience the following shortcomings:

1) Traditional clustering algorithms may fall into local optimum concerning the selection of the initial centres of the clusters.
2) Traditional clustering algorithms seek to optimise a single cluster criterion; hence, they may not include the various datasets characteristics.
3) They require a pre-given fixed number of clusters.

Table 2.1 Summary of the traditional data clustering types with their related algorithms

Clustering Technique	Example	Type of Data	Shape of clusters	Handle outliers and noise	Advantages	Disadvantages
Partitioning -based	K-means, K-medoids, Fuzzy C-Means, EM, K-harmonic, CLARANS, CLARA, K-mode	Numerical and Categorical (k-mode)	Non-convex	No	• Simple and relatively scalable. • Appropriate for datasets with well-separated and spherical clusters.	• Can be easily trapped into the local optima. • Not suitable for arbitrary-shapes and dense clusters. • The number of clusters (k) is provided. • Highly sensitive to the initial centroids. • Suitable only with numerical type data.
Hierarchical-based	DIANA AGNES CURE BIRICH CHAMELEON	Numerical and categorical	Arbitrary and Non-convex	Yes	• Flexible regarding the level of granularity. • Do not require the number of clusters to be known in advance. • Suitable for solving problems that involve point linkages (e.g. taxonomy trees)	• Cannot be corrected as soon as the dividing/combining decision is made. • Require high computational time for large and high-dimensional datasets. • Have degraded performance in high-dimensional spaces.
Density-based	DBSCAN OPTICS DENCLUE	Numerical	Arbitrary	Yes	• Discover clusters with different sizes and arbitrary shapes. • Efficient for low-dimensional data.	• Highly sensitive to the input parameters setting. • Unsuitable for high-dimensional datasets. • high computational time for high-dimensional datasets
Grid-based	STING WaveCluster	Spatial	Arbitrary	Yes	• Fast processing time in low-dimensional datasets. • Insensitive to the initialisation.	• High computational time for high-dimensional datasets. • All cluster boundaries are either horizontal or vertical; no diagonal boundary exists.
Subspace	CLIQUE MAFIA ENCLUS	Spatial	Arbitrary	No	• Simple and relatively scalable. • Suitable for high-dimensional datasets.	• Known number of subspaces and dimensions.

2.3 METAHEURISTIC ALGORITHMS FOR DATA CLUSTERING PROBLEMS

This section discusses the data clustering algorithms related metaheuristics which are classified as single-solution based (S-metaheuristics) and population-based (P-metaheuristics) metaheuristics (Talbi 2009). In brief, the single-solution based algorithms, such as simulated annealing and local search, handle and modify a single solution throughout the search process. In the population-based algorithms, such as evolutionary algorithms (EA) and particle swarm optimisation (PSO), the entire solutions in the population are considered. The single-solution based and population-based metaheuristics provide complementary features; the single-solution based metaheuristics are focused on the exploitation capability, which narrows the search over local areas. In contrast, the population-based metaheuristics employ the exploration that allows a sufficient diversification in the entire search space. To develop metaheuristic methods, two significant criteria have to be considered concerning the search space: the exploration or diversification and the exploitation or intensification. Figure 2.7 demonstrates both conflicting criteria of exploration and exploitation behaviour. Where single-based metaheuristics are more focused on exploitation, and the population-based metaheuristics concentrate more on exploration.

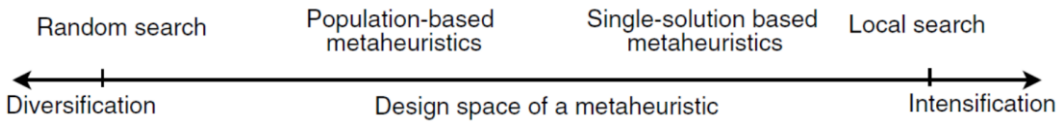


Figure 2.7 The conflict between exploration and exploitation (Talbi 2009)

The clustering methods intend to obtain a proper grouping of the input dataset so that some criterions are optimised. Subsequently, the data clustering problem can be represented as an optimisation problem (Das et al. 2009; Jain 2010). The objective is to optimise various characteristics of the clusters, like separation, compactness, and connectivity. The direct method to represent data clustering as an optimisation problem by optimising a cluster validity measure, which considers as the clustering solutions quality (Das et al. 2009). Any possible partitioning of the dataset can determine the search space of the optimisation problem and the associated values of the validity measure.

2.3.1 Single-solution based metaheuristics algorithms for data clustering

The algorithmic steps of every single-solution based metaheuristic (S-metaheuristics), also named single-based or local search, consist of the following steps:

- The search begins with an initial solution.
- At every iteration, The neighbourhood operator is utilised to produce a neighbouring solution.
- the chosen solution is adopted as the current solution based on acceptance criteria.
- These steps are repeated until a termination condition is satisfied.

Single-solution based metaheuristics employ the generation and replacement approaches iteratively on the current solution. The generation step produces a candidate solutions set from the current solution. The replacement phase a selects new solution from the candidate solution set to replace the current solution. The most commonly used S-metaheuristics are tabu search and simulated annealing. The typical search concepts for all S-metaheuristics are the definition of the neighbourhood structure and the determination of the initial solution. Some single-solution based a metaheuristics methods have been offered to solve the data clustering problems such as tabu search (TS) and simulated annealing (SA). However, this section focuses on the popular algorithms which achieved the best results.

a. Tabu Search Based Clustering Algorithms

The tabu search algorithm (TS) transforms the current solution to obtain a better solution by searching of the neighbourhood space. The objective function of the optimisation problem is employed to assess the quality of the solution, and the purpose is to obtain an optimal solution while exploring the search space. The original tabu search algorithm is utilised in solving the data clustering problem in (Al-Sultan 1995). A later study in tabu search based clustering method proposed by (Liu et al. 2008). The TS-Clustering is developed to investigate the clustering results; three neighbourhood methods are adopted to discover the nearby solutions. Recent research in tabu search

based clustering is introduced by Gyamfi et al. (2017), where they suggested tabu search optimisation method for K-Means clustering with an alternative, low-complexity formulation. The purposed algorithm intends to obtain the cluster centres using a neighbourhood structure that utilises the objective function gradient information. In the research of (Lu et al. 2017), the authors introduced an enhanced K-means clustering algorithm that is extended by a tabu search strategy, and which is modified to satisfy the demands of the applications of big data.

b. Simulated Annealing Algorithm Based Clustering Algorithms

Simulated annealing (SA) is inspired from the manner of heat and cooling controlling mechanism of the materials. The primary approach of the algorithm is replacing the current solution by a random neighbouring solution. The neighbouring solution is selected based on a probability between sequential solutions and a global parameter (temperature). The temperature is adjusted gradually depending random moves chosen. This acceptance is considered significant because the enables the SA algorithm to escape from local optimal solutions. Selim & Alsultan (1991) applied the simulated annealing to solve the data clustering problem. Since then, several other researchers have employed and enhanced SA to solve data clustering problem (Abdi et al. 2012; Duczmal & Assuncáo 2004; Kangping et al. 2016; Maulik & Mukhopadhyay 2010).

2.3.2 Population-based metaheuristics algorithms for data clustering

The population-based algorithms employ stochastic local search algorithms to maintain the population of the candidate solutions for the provided problem. In every search step, some individuals of the population may be transformed into new individuals. This mechanism supports the algorithm to obtain sufficient diversification over the search process. The population-based metaheuristics consist of the following algorithmic steps:

- The search begins with an initial population of solutions.
- At every iteration, apply the generation of a new population.

- Perform the replacement of the current population, where the selection is carried out from the current and the new populations.
- These steps are repeated until a termination condition is satisfied.

Many of the population-based metaheuristics are considered as nature-inspired algorithms. Moreover, several population-based metaheuristic methods have been proposed as novel clustering methods. In this section, various famous clustering techniques based on population-based metaheuristic algorithms are comprehensively discussed, such as Genetics algorithms (Mustafi et al. 2017), Particle Swarm Optimisation (Niu et al. 2017), Differential Evolution (Ramadas et al. 2016), Ant Colony Optimisation (Tao et al. 2016), Artificial Bee Colony (Kumar et al. 2017), Honey Bee Mating Optimisation (Chakaravarthy & Kalyani 2015), Firefly Algorithm (Nayak et al. 2017), Imperialist Competitive Algorithm (Emami & Derakhshan 2015), Bacterial Evolutionary Algorithm (Niu et al. 2013), Black Hole Algorithm (Chandrasekar & Krishnamoorthi 2014), Grey Wolf Optimiser (Kumar et al. 2017), Harmony Search Algorithm (Senthilnath et al. 2016), Cat Swarm Optimisation (Razzaq et al. 2016), Gravitational Search Algorithm (Han et al. 2017), Cuckoo Search Algorithm (Pandey et al. 2017), Krill Herd algorithm (Abualigah et al. 2017). However, the present review focuses on the most common algorithms which attain the best results.

a. Genetic algorithm based clustering

The primary objective of the Genetic Algorithm (GA) is to solve an optimisation problem by producing a collection of candidate solutions. The GA repeatedly improve the candidate solutions by the modification process of superior solutions during subsequent iterations. The GA chooses the candidate solutions based on their objective function value that verifies the quality of the solution. In GAs, the modification process consists of the mutation of current solutions to its local neighbourhood and the crossover that recombines between some selected solutions.

The GA-based clustering algorithm studies started by (Raghavan & Birchard 1979). The authors adopted a direct encoding of the object-cluster relationship. The approach utilises the genetic encoding that shortly indicates *n* objects to *k* clusters, so

that whole candidate solutions contain n genes with integer values in the interval [1, *k*]. The GA intends to find the optimal partitioning according to the objective function value that estimates the quality of the clusters.

Another approach of GA-based partitioning clustering that encodes a prototype variable, such as the cluster centroid, to extend and shape of the variance for each cluster. Several studies have suggested cluster centroids, or medoids to represent the prototype of each data point over a particular cluster. The primary purpose is to guarantee the representation points of every cluster and to attach every data point to cluster with nearest prototype point (Bandyopadhyay & Maulik 2002; Maulik & Bandyopadhyay 2000).

An enhanced GA-based partitioning clustering is offered by (Chen et al. 2010), which focuses on the population diversity issue by investigating the similarity between individuals before the selection. A genetic-based algorithm for determining the appropriate number of clusters in a particulate dataset is introduced in (Liu et al. 2011). An application of a hybrid method inspired from particle swarm optimisation (PSO) and genetic algorithm (GA) for data clustering is offered by (Kuo & Lin 2010). The experiments revealed that the offered algorithm is more accurate than the GA-based and PSO-based clustering algorithms (Kuo et al. 2012). Mustafi et al. (2017) introduced a GA-based clustering algorithm to overcome difficulties that usually affects the k-means clustering algorithm, which enhances the clustering performance suggested by the k-means algorithm and also guarantees the resolution of the necessary number of clusters.

b. Particle Swarm Optimization Based Clustering Algorithms

The particle swarm optimisation (PSO) inspired by the stochastic optimisation method based on Swarm Intelligence (SI) (Kennedy & Eberhart 1995). The initial concept of each particle expresses a possible result that is enhances based on its neighbour's experience. The PSO algorithm accordingly explores within an individuals group. The velocity of the individuals and the swarm particles will be utilised to optimise them. Thus, the optimisation of these individuals and particles will require early knowledge

and the experience of the neighbours. The paths of mobile points in a multidimensional space play a primary role in the search process in the problem space of PSO.

PSO recognised as an efficient optimisation algorithm (Tsai & Chiu 2008). However, it has some limitation like when the multi-objective PSO approach is applied it may become stuck into local optima, and also it converges quickly in mid optimum solutions. Many modifications have been proposed in PSO to address these problems. Ahmadyfard & Modares (2008) offered a hybrid clustering algorithm (PSO-KM). The algorithm initialised the solutions using PSO, which provides a full exploration of search space for global solutions. Next, the K-means clustering employed for faster convergence to perform the data clustering. Alam et al. (2008) presented an evolutionary particle swarm optimization for clustering problems. In this algorithm the particle that does not meet the fitness criteria will be removed by stronger swarm after a fixed number of generations, resulting in a potential optimal swarm.

Jiang et al. (2013) enhanced the PSO performance using a novel searching approach that is based on the particles ageing. Kumar and Sahoo (2015) introduced a novel hybrid metaheuristic algorithm that combines the PSO and the magnetic charge system search (MCSS) for the partitioning clustering problem. The MCSS–PSO incorporates the neighbourhood search approach to produce better clustering solutions. Nayak et al. (2016) offered an enhanced hybrid PSO evolutionary K-means clustering approach to achieve the optimal solutions for the cluster centres. The combination of improved PSO, GA, and K-means algorithm enhances the convergence speed and to obtain the optimal global clustering solutions.

Lashkari and Moattar (2016) introduced a PSO algorithm that is integrated with k-means. Comparative experiments on real-life and synthetic datasets reveal that the proposed algorithm can achieve better and stable clustering solutions. Bouyer & Hatamlou (2018) combined the K-Harmonic Means (KHM) algorithm with PSO and an improved Cuckoo Search (ICS). They used ICS and PSO to avoid the problem of falling into the local optima. A literature review of data clustering algorithms based on PSO can be found in (Alam et al. 2015; Esmin et al. 2015; Inkaya et al. 2016; Rana et al. 2011; Sarkar et al. 2013).

c. Differential Evolution Based Clustering Algorithms

Differential evolution (DE) is a kind of standard EA which assesses the original population by utilising observation models and probability movement, where the evolution of the population is achieved by employing evolution operators (Storn & Price 1997). The primary concept of the DE is to produce a different solution for each solution by employing two random members and one fixed member (usually the best solution) to provide better solutions.

In the recent decade, various improvements have been proposed in DE to produce better clustering solutions. Kwedlo (2011) introduced a new clustering technique (DE-KM), which combines the DE with the K-means algorithm. The experimental results DE-KM obtains lower SSE values for the produced solutions than the other algorithms. In (Tvrdík & Křivý 2011), authors suggested a novel hybrid DE, combining the k-means algorithm as local search.

In (Chen et al. 2014) authors employed the evolutionary clustering DE (deEC). Comparing with the k-means, deEC could achieve a global search in the solution space. Xiang et al. (2015) proposed a dynamic shuffled DE (DSDE) for data clustering problem. In DSDE, mutation strategy DE/best/1 is applied, which can take advantage of the guidance knowledge from the best individual to accelerate the convergence of the DE algorithm. In (Babrdel Bonab et al. 2015), the authors introduced an efficient combination approach for finding optimal clusters centres with proper initialisation (CCIA). The algorithm incorporates the bees algorithm (BA) and DE to solve the clustering problem. Tvrdík and Křivý (2015) introduced a new clustering method by combining DE and k-means. Ramadas et al. (2016) offered a forced strategy improvement to DE named (FSDE), by performing a novel mutation strategy.

d. Ant Colony Optimization Based Clustering Algorithms

The ant colony optimisation (ACO) is a stochastic metaheuristic algorithm applied for solving optimisation problems. ACO is inspired by the ants' natural behaviour to determine the best route for food source nest by using the pheromones. In ACO, agents

individually build solutions in parallel by regularly improving the partial solutions (Dorigo et al. 1996; Dorigo & Blum 2005).

An improved ACO that employs a modified pheromone update strategy to enhance the clustering solutions is offered by (Tsai et al. 2011). The authors have applied two pheromone tables for the foraging knowledge to maintain the convergence and diversity of the population concurrently. Huang et al. (2013) introduced four kinds of a combination are used; sequence approach, global best exchange, parallel approach, and an enlarged pheromone-particle table.

Menéndez et al. (2014) introduced an ACO-based clustering algorithm (ACOC). The suggested method restructures the centroid-based technique into a medoid-based technique. Later, Tao et al. (2016) offered a novel ACO clustering algorithm based on data combination mechanism to enhance the computational complexity and accuracy of the ACO. A comprehensive review of the modifications and results on the qualitative performance obtained by ACO are discussed in (Zhe et al. 2011).

e. Honey Bee Mating Optimization Based Clustering Algorithms

Honey Bee Mating Optimization (HBMO) is inspired by the concept of the mating of real honey bees in nature. HBMO has been employed for data clustering in (Fathian et al. 2007). The authors analysed the HBMO performance with different stochastic algorithms such as ACO, GA, SA and TS algorithms. They showed that the HBMO algorithm shown better solution quality. Chiu and Kuo (2009) hybridised PSO with HBMO to tackle data clustering problem. This hybrid method has experimented on various internal validity function. Experimental on the hybrid approach reveals that it has better performance regarding finding the global optimum.

Teimoury et al. (2011) introduced a combination between HBMO and K-means to solve the data clustering problem. The hybrid approach adopted the silhouette coefficient measure to identify the number of clusters. Shafia et al. (2011) offered an improved data clustering approach based on the combination between HBMO and K-

means (GBTKC). GBTKC algorithm intends to obtain the diversity control of GA to find the global optimum. A comprehensive review of clustering algorithms based on HBMO can be found in (Chakaravarthy & Kalyani 2015).

f. Artificial bee colony algorithm Based Clustering Algorithms

The Artificial Bee Colony algorithm (ABC) (Karaboga & Basturk 2007) consists of three kinds of bees: employed bees, onlooker bees and scouts. The bee that is randomly exploring is recognised as a scout. The bee seeking for the food source and sharing its knowledge kinds is recognised as the employed bee, and the bee is serving on the working region is recognised as onlooker bee. In (Karaboga & Ozturk 2011; Zhang et al. 2010), The ABC algorithm is adopted and applied to solve data clustering problems.

In (Wang & Wang 2014), authors introduced ABC clustering algorithm based on K-means. Wang et al. (2015) introduced an improved combination approach of ABC with K- means (EABCK). The EABCK utilised a new mutation operation that led by the best solution. Gong et al. (2016) offered an improved ABC algorithm for data clustering by enhancing the initial procedure of clustering centres. The proposed approach adopts a new dynamic local strategy. A comprehensive review of clustering algorithms based on ABC Algorithm can be found in (Kumar et al. 2017; Kumar 2015; Gupta & Kumar 2014).

g. Firefly algorithm Based Clustering Algorithms

Firefly Algorithm (FA) is a modern nature-inspired optimisation algorithm, which simulates the fireflies flash behaviour. The candidate solution in the FA moves to other better fitness candidate solution. Each firefly moves by a specific distance depending on the distance separating two firefly particles.

Senthilnath et al. (2011) employed FA for data clustering problem. The FA performance utilising the percentage of the classification error criterion is tested against ABC and PSO. Hassanzadeh and Meybodi (2012) hybridised FA with the K-means algorithm for avoiding trap into the local optimum. The proposed approach two stages

for clustering, which it used the FA initialise the centroids of the clusters. Then, the K-means algorithm is applied to determine the best the centroids for the clusters.

In (George & Parthiban 2015), authors introduced a combination between FA and Group Search Optimizer. The hybridisation is handled by rearranging the worst solution at every iteration GSO by updated solution from FA. Maheshwar et al. (2015) introduced a combination between FA and GA algorithms (FAG), where FA is utilised to initialise the population.

Sadeghzadeh (2016) offered an EA based on FA algorithm solving data clustering problem. Nayak et al. (2017) authors introduced a new FA-based K-means algorithm (FA-KM) for efficient data clustering. A comprehensive review of FA over different optimisation domains is provided by (Fister et al. 2013, 2014).

h. Bacteria Evolutionary Algorithm Based Clustering Algorithms

Bacteria Evolutionary Algorithm (BEA) inspired by the microbial evolution phenomenon. BEA combines two special activities namely, bacterial gene transfer and mutation operations, to improve the population. In (Das et al. 2009), these operations are modified to handle variable-length chromosomes that encode different cluster grouping. Several real-life and synthetic datasets are utilised to assess the performance BEA algorithm. The experiments show the BEA performance is superior regarding final clustering accuracy.

Lei et al. (2011) offered a BEA clustering algorithm for a protein-protein interaction network. The cluster centre is represented as the initial position of the bacterium, and the adjacent nodes of the cluster centre are considered as the positions where the bacterium moves. In this approach, the nodes chosen in the chemotactic process are categorised as clusters when performing the elimination-dispersal and reproduction operations. The algorithm proceeds to generate different clusters until entire nodes are arranged into clusters.

Wan et al. (2012) introduced a novel clustering approach inspired by the bacterial foraging algorithm (BFA), in which the combination of bacteria forage to concentrate to some positions as final clusters. Niu et al. (2013) combined BFA and K-means for data clustering problem to utilise the excellent local search capacities of K-means and BFA global search capability.

i. Black Hole Algorithm Based Clustering Algorithms

Black Hole Algorithm (BH) is based on the natural phenomenon of the black hole. The initial concept of the black hole is inspired by the space regions that are concentrated mass areas producing a high gravitational pull for the nearby objects (Kumar et al. 2015). Firouzi et al. (2010) introduced a combination between the EA based on the BH and the k-means algorithms, named BH–BKmeans. Hatamlou (2013) proposed a heuristic algorithm based on the black hole phenomenon, where it has a simple structure, easy and free from parameter tuning implementation. Chandrasekar and Krishnamoorthi (2014) introduced a combination algorithm between the BH and a heuristic search algorithm to generate high-quality solutions.

j. Grey Wolf Optimizer Based Clustering Algorithms

The grey wolf optimiser (GWO) is inspired by the natural behaviour of hunting and social leadership of grey wolves. In GWO, the search starts by initialising the population with randomly generated wolves (Mirjalili et al. 2014). Zhang and Zhou (2015) combined GWO with Powell local optimisation (PGWO). Jadhav and Gomathi (2017) offered a method of data clustering using the WGC algorithm that utilises the Whale Optimization Algorithm (WOA) computational steps (WEGWO) with an extended fitness function. Kumar et al. (2017) introduced a new clustering technique based on GWA (GWAC). The GWA search capability is utilised to explore the search space for optimal clusters centres.

k. Harmony Search Algorithm Based Clustering Algorithms

Harmony Search (HS) mimics the musical manner of exploring for a perfect harmony state defined by an aesthetic pattern, which has been utilised in problems optimisation (Geem et al. 2001). Kumar et al. (2014b) proposed a modified HS inspired by the process of musical improvisation. Senthilnath et al. (2016) proposed an HS clustering algorithm applied to obtain the clusters centres.

l. Cat Swarm Optimization Based Clustering Algorithms

The cat swarm optimisation (CSO) is one of the recent metaheuristic algorithms that is inspired by the cats' behaviour and utilised to solve optimisation problems (Chu et al. 2006). Kumar & Sahoo (2015) introduced an enhanced CSO method based on Cauchy mutation operator. Razzaq et al. (2016) offered a modified clustering algorithm based on CSO (MCSO). The MCSO is utilised to initialise clusters centre.

m. Gravitational Search Algorithm Based Clustering Algorithms

The gravitational search algorithm (GSA) is one of the recent population-based metaheuristics that is inspired by mass interactions and the law of gravity. Hatamlou et al. (2011) proposed a gravitational search algorithm (GSA) for solving data clustering algorithm. The candidate solutions are created randomly and interact with one solution via Newton’s gravity law to find optimal solutions in the problem space. Dowlatshahi and Nezamabadi-Pour (2014) proposed a grouping-based GSA (GGSA). The GGSA algorithm is similar to the GSA computation steps, and also it adopted a grouping coefficient. In the research of (Huang et al. 2015), authors offered a memetic GSA algorithm (MGSA). The MGSA is joined with the multi-start operator and the pattern reduction operator. Nikbakht and Mirvaziri (2015a) offered a clustering algorithm that combines the GSA with the genetic operators. Han et al. (2017) introduced an enhanced GSA, which is called bird flock GSA (BFGSA). The BFGSA offers a novel approach to GSA by combining diversity. This approach is based on the birds' collective response behaviour.

n. Cuckoo Search Algorithm Based Clustering Algorithms

The Cuckoo search Algorithm (CSA) is inspired by the natural behaviour of some kinds of cuckoo among with the behaviour of levy flight (Shehab et al. 2017). Manikandan & Selvarajan (2014) introduced a new data clustering algorithm based on CSA. Ameryan et al. (2014) offered CSA approaches based on random COA, Chaotic COA, and K-means. The COA Clustering initialises the population randomly, while chaotic COA covers the entire search space to search for better solutions.

In (Zhao et al. 2014), authors introduced an improved CSA (ICS) that modifies the randomisation and movement of the CSA. Zhao et al. (2016) offered an improved the K-Means based on CSA algorithm. The population initialisation was carried out the proposed CSA. Pandey et al. (2017) introduced a new clustering algorithm based on CSA and k-means algorithms. The proposed algorithm extended the abilities of K-means clustering method to produce a better quality of the clustering solutions.

o. Krill Herd algorithm Based Clustering Algorithms

The Krill herd algorithm (KH) is a novel metaheuristic optimisation algorithm that is inspired by the krill swarm behaviour while exploring for food and the method they communicate with each other (Abualigah et al. 2017). Nikbakht & Mirvaziri (2015b) offered a novel clustering algorithm that combines the KH and K-means algorithms. Agrawal & Pandit (2016) combines the neighbourhood distance and genetic reproduction approaches with the KH Algorithm (KHAMC).

In the research of (Jensi & Jiji 2016), the authors offered an improved KH(IKH) by enhancing the krill global search operator for the exploration nearby the determined search space. Li and Liu (2017) introduced an improved KH algorithm (IKHA) that modifies the mutation operators and enhances the global optimisation method. In (Abualigah et al. 2017) authors combined the KH with HS algorithms by adding the HS global search operator with the KH algorithm. The modification aims to improve exploration capability of the KH algorithm.

p. Imperialist Competitive Algorithm Based Clustering Algorithms

Atashpaz-Gargari and Lucas (2007) have proposed an imperialist competitive algorithm (ICA), which presents the imperialism social policy to affect nations and control their sources. The ICA algorithm starts with initialising the population randomly. Then, the imperialists choose a few of the best countries, and the remaining will help the imperialists to build the colonies, which altogether will form an empire. Niknama et al. (2011) employed the ICA as hybridisation with K-means and a local search. Abdeyazdan (2014) presented an enhanced data clustering approach for that adopts the combination of the K-harmonic means algorithm (KHM) and a modified version of Imperialist Competitive Algorithm (ICA) algorithm.

q. Other metaheuristic algorithms for data clustering

Kushwaha et al. (2017) proposed a new clustering algorithm inspired by the Magnetic optimisation algorithm. The experiments reveal that the introduced Magnetic optimisation data clustering algorithm improved the results concerning the accuracy, efficiency and the robustness. In (Yadav & Nanda 2016), the authors introduced the League Championship Algorithm (LCA) clustering algorithm. Later, Wangchamhan et al. (2017) introduced a combined k-means with chaotic LCA (KSC-LCA). In (Jensi & Jiji 2015), the authors proposed a modified bat algorithm for data clustering method. Agarwal & Mehta (2016) offered an improved flower pollination algorithm for data clustering.

r. A Brief summary

Figure 2.8 presents as the summary metaheuristics approach applied for data clustering problem and the relevant literature review that has been conducted in this section.

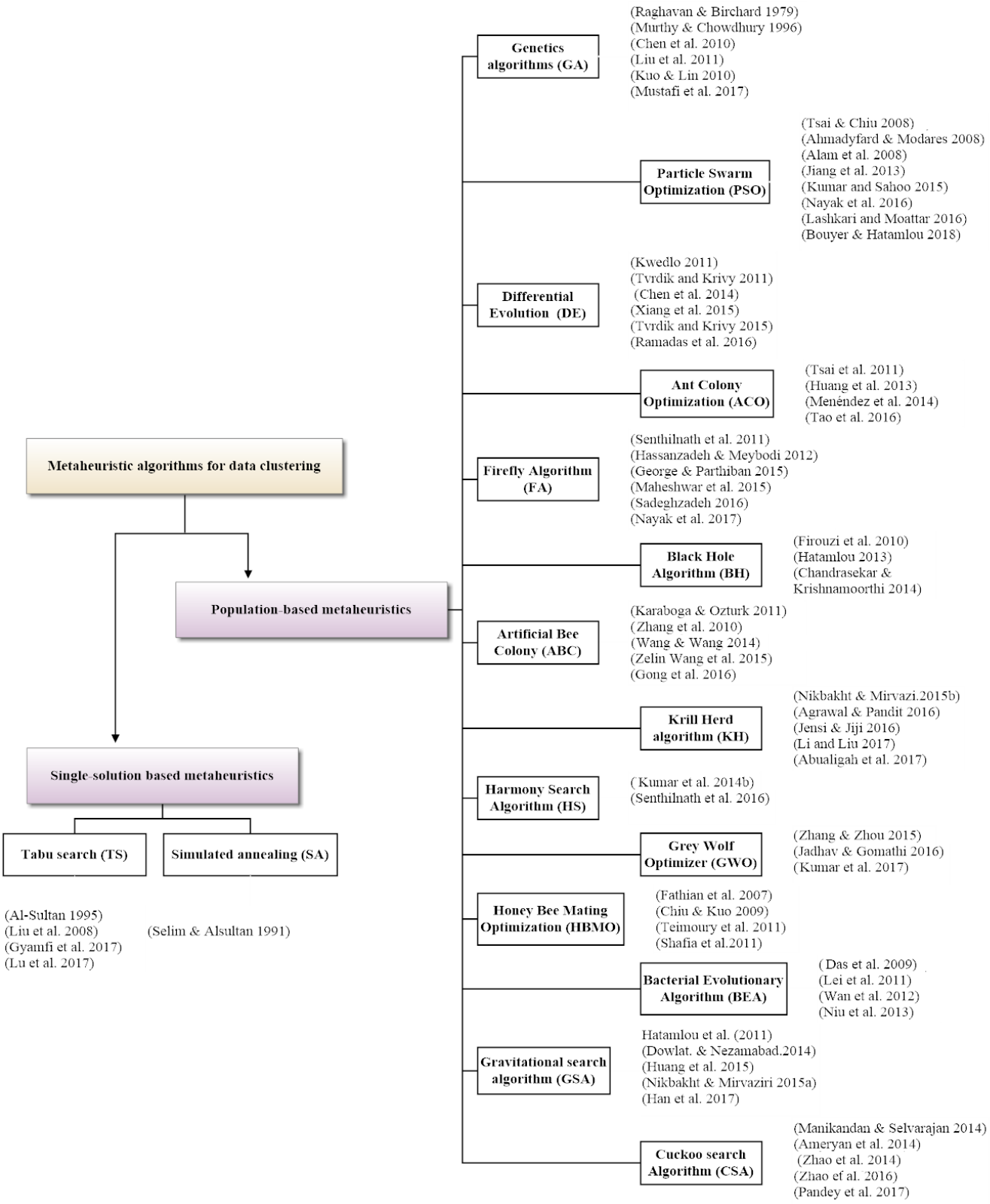


Figure 2.8 Summary of metaheuristics approaches applied for data clustering problem

2.3.3 Main findings from literature review of metaheuristics algorithms for data clustering

In the recent decade, numerous data clustering algorithms are introduced in the literature. Generally, there is no such algorithm can be appropriate to deal will various data types, applications, and requirements. Every algorithm may have its limitations, shortcomings, and advantages. Accordingly, offering new methods for data clustering problems is an active research area. Metaheuristic optimisation algorithms play a significant role in against other traditional clustering algorithms, which are widely applicable, easily to be implemented, and capable of dealing with complex and high-dimensional problems (Das et al. 2009; Talbi 2009, 2012). Nevertheless, some metaheuristic algorithms shortcomings include falling into local optima, premature convergence, uncertain and slow convergence, memory and time complexity problems, and the tuning of the parameters. These shortcomings may hinder the performance of the data clustering concerning the quality of the solutions (Das et al. 2009; Talbi 2009).

The premature convergence, as one of the main drawbacks related to the metaheuristic algorithms, can lead the clustering algorithm to be trapped in local optima. Usually, the premature convergence occurs whenever the search process is trapped inside a limited search space region (Bouyer & Hatamlou 2018; Han et al. 2017), in which the search process cannot explore new search space regions. Consequently, this will lead the search process to a local optima problem. Introducing new practical strategies and approaches to avoid falling in such convergence problems may improve the algorithm robustness to find better clustering solutions. Commonly, in early stages of the algorithm search process, few metaheuristic algorithms converge quite quickly, although the convergence turn to slow throughout the remaining number of iterations. Therefore, the quality of the solutions may not be enhanced toward the global optima. The strategy that can be used to avoid such problems is to hybridise a proper local search algorithm (Abul Hasan & Ramakrishnan 2011; Bouyer & Hatamlou 2018; Talbi 2009, 2012).

Additionally, the trade-off between exploration and exploitation can affect the ability of the clustering algorithm in finding good clusters among the datasets being

used (Dowlatshahi & Nezamabadi-Pour 2014; Kumar et al. 2015; Liu et al. 2012). Some of the earlier proposed clustering algorithms, based on metaheuristics, managed to find good clustering solutions for specific datasets. However, across all datasets, it was unable to find good results reliably, or the results were not robust (Aggarwal & Reddy 2013; Celebi 2015). This issue might be due to the inappropriate balance between exploration and exploitation of the metaheuristic algorithm that may lead to premature convergence (Bouyer & Hatamlou 2018; Dowlatshahi & Nezamabadi-Pour 2014). Similar to convergence problems, some researchers have suggested the hybridisation approaches by hybridising a global search with a local search to obtain a better balance. Table 2.2 presents the summary of the essential strength and limitation of the metaheuristic algorithms applied for data clustering that are available in the literature (Abul Hasan & Ramakrishnan 2011; Saxena et al. 2017; Xu & Tian 2015).

Table 2.2 Summary of the main strength and limitation of metaheuristic algorithms applied for data clustering

Strength	**Limitation**
• Quickly reach high-quality solutions and produce a practical approach to deal with complex and high dimensional problems. • Helpful in situations when traditional clustering approaches fall in local optima problem.	• Do not ensure achieving the best clustering solution. • Several metaheuristic approaches require a specific parameter setting that is only dedicated to a particular approach or even one specific dataset, which leads to lack of approach generality. • The population initialisation can influence the final solution quality. • Some metaheuristic approaches experience slow convergence speed or premature convergence. • Some metaheuristic approaches suffer from an insufficient balance between exploitation and exploration.

It is evident from the conducted literature review that the optimisation algorithm approached applies several approaches to solve the data clustering problem. These approaches may require more improvement to obtain an appropriate and robust performance for the clustering solutions. The most preferred approaches found in the literature to resolve the drawbacks of the metaheuristic approaches in data cluster problems are described as follows:

1) The balance between exploration and exploitation: to improve the algorithm convergence strategy by offering a right balance between exploration and