HYBRID OPTIMIZATION FOR K-MEANS CLUSTERING LEARNING ENHANCEMENT

YOUSEF FARHANG

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

January 2016

Dedicated to my beloved family

ACKNOWLEDGEMENT

I heartily express my gratefulness to Allah s.w.t for His blessing and strength that He blessed to me during the completion of this research.

My sincere thanks go to my supervisor Prof. Dr. Siti Mariyam Shamsuddin for his continuous motivation, constant advice, encouragement and support from start to the completion of my studies.

I am ever grateful to my family, especially my wife, for their continuous support in term of encouragement and motivation.

Furthermore, very genuine appreciation goes to my father (1921-2001) whom I owe my very existence to the world, who always gave me the motivation and courage to look on the bright side every time I felt unmotivated, whom that never let me down and whom I respect the most in my heart.

This research work has been financially supported by UTM's International Doctoral Fellowship (IDF). I would like to thank the members of Universiti Teknologi Malaysia (UTM) for providing the research facilities and thanks to all Soft Computing Research Group (SCRG) lab for their constructive suggestions.

ABSTRACT

In recent years, combinational optimization issues are introduced as critical problems in clustering algorithms to partition data in a way that optimizes the performance of clustering. K-means algorithm is one of the famous and more popular clustering algorithms which can be simply implemented and it can easily solve the optimization issue with less extra information. But the problems associated with K-means algorithm are high error rate, high intra cluster distance and low accuracy. In this regard, researchers have worked to improve the problems computationally, creating efficient solutions that lead to better data analysis through the K-means clustering algorithm. The aim of this study is to improve the accuracy of the K-means algorithm using hybrid and meta-heuristic methods. To this end, a meta-heuristic approach was proposed for the hybridization of K-means algorithm scheme. It obtained better results by developing a hybrid Genetic Algorithm-K-means (GA-K-means) and a hybrid Partial Swarm Optimization-K-means (PSO-K-means) method. Finally, the meta-heuristic of Genetic Algorithm-Partial Swarm Optimization (GAPSO) and Partial Swarm Optimization-Genetic Algorithm (PSOGA) through the K-means algorithm were proposed. The study adopted a methodological approach to achieve the goal in three phases. First, it developed a hybrid GA-based K-means algorithm through a new crossover algorithm based on the range of attributes in order to decrease the number of errors and increase the accuracy rate. Then, a hybrid PSO-based K-means algorithm was mooted by a new calculation function based on the range of domain for decreasing intra-cluster distance and increasing the accuracy rate. Eventually, two meta-heuristic algorithms namely GAPSO-K-means and PSOGA-K-means algorithms were introduced by combining the proposed algorithms to increase the number of correct answers and improve the accuracy rate. The approach was evaluated using six integer standard data sets provided by the University of California Irvine (UCI). Findings confirmed that the hybrid optimization approach enhanced the performance of K-means clustering algorithm. Although both GA-K-means and PSO-K-means improved the result of K-means algorithm, GAPSO-K-means and PSOGA-K-means meta-heuristic algorithms outperformed the hybrid approaches. PSOGA-K-means resulted in 5%-10% more accuracy for all data sets in comparison with other methods. The approach adopted in this study successfully increased the accuracy rate of the clustering analysis and decreased its error rate and intra-cluster distance.

ABSTRAK

Dalam beberapa tahun kebelakangan ini, isu-isu pengoptimuman gabungan telah dikenal pasti sebagai masalah kritikal dalam pengelompokan algoritma bagi pembahagian data dengan cara yang mengoptimumkan prestasi pengelompokan. Algoritma K-min merupakah salah satu algoritma pengelompokan yang terkenal dan popular. Algoritma ini mudah dilaksanakan dan boleh menyelesaikan isu-isu pengoptimuman walau dengan menggunakan maklumat yang sedikit. Namun, masalah yang timbul dengan pelaksanaan algoritma K-min adalah kadar ralat yang tinggi, jarak antara kluster yang tinggi, dan juga kadar ketepatan yang rendah. Para penyelidik telah berusaha keras dalam memperbaiki masalah-masalah ini secara berkomputer, mewujudkan penyelesaian yang berkesan yang membawa kepada analisis data yang lebih baik melalui pengelompokan algoritma K-min. Tujuan kajian ini adalah untuk meningkatkan ketepatan K-min menggunakan kaedah algoritma hibrid dan meta-heuristik. Bagi tujuan ini, pendekatan meta-heuristik dicadangkan untuk penghibridan skim algoritma K-min. Ia menghasilkan keputusan yang lebih baik dengan membangunkan kaedah hibrid Algoritma Genetik-K-min (GA-KM) dan Pengoptimuman Separa Kelompok-K-min (PSO-KM). Akhirnya, meta-heuristik daripada Algoritma Genetik-Pengoptimuman Separa Kelompok (GAPSO) dan Pengoptimuman Separa Kelompok-Algoritma Genetik (PSOGA) melalui algoritma K-min yang telah dicadangkan. Kajian ini mengaplikasikan pendekatan metodologi untuk mencapai matlamat dalam tiga fasa. Pertama, ia membangunkan GA hibrid berasaskan algoritma K-min melalui algoritma lintasan baru berdasarkan pelbagai sifat untuk mengurangkan bilangan kesilapan dan meningkatkan kadar ketepatan. Kemudian, PSO hibrid berasaskan algoritma K-min yang diilhamkan oleh fungsi pengiraan baru berdasarkan pelbagai domain untuk mengurangkan jarak antara kelompok dan meningkatkan kadar ketepatan. Akhirnya, dua algoritma meta-heuristik iaitu algoritma GAPSO dan PSOGA diperkenalkan melalui kombinasi algoritma yang dicadangkan untuk meningkatkan bilangan jawapan yang betul dan meningkatkan kadar ketepatan. Pendekatan ini telah dinilai menggunakan enam set data integer piawai yang disediakan oleh University of California Irvine (UCI). Dapatan kajian ini mengesahkan bahawa pendekatan pengoptimuman hibrid meningkatkan prestasi pengelompokan algoritma K-min. Walaupun kedua-dua GA-KM dan PSO-KM memberi hasil lebih baik daripada algoritma K-min, algoritma GAPSO dan PSOGA meta-heuristik mengatasi pendekatan hibrid. PSOGA-K-min telah menghasilkan kadar ketepatan sehingga 5%-10% untuk semua set data berbanding dengan kaedah-kaedah yang lain. Pendekatan yang diambil dalam kajian ini berjaya meningkatkan kadar ketepatan analisis pengelompokan dan menurunkan kadar kesilapan dan jarak di antara kelompok.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	-	Artificial Intelligence
ANN	-	Artificial Neural Network
CGA	-	Clustering Genetic Algorithm
CS	-	Computer Science
DBSCAN	-	Density-Based Spatial Clustering of Applications Noise
DCPSO	-	Dynamic Clustering approach based on PSO
DE	-	Differential Evolution
EA	-	Evolutionary Algorithm
FGKA	-	Fast Genetic K-means Algorithm
GA	-	Genetic Algorithm
GA-KM	-	Genetic Algorithm-K-Means
GAPSO-KM	-	Genetic Algorithm-Particle Swarm Optimization-K-Means
GDM	-	Genetic Distance Measure
GGA	-	Genetically Guided Algorithm
GKA	-	Genetic K-means Algorithm
GWKMA	-	Genetic Weighted K-means Algorithm
HGACLUS	-	Hybrid GA-based Clustering Schema
I-GA-KM	-	Improved Genetic Algorithm-K-Means
IGKA	-	Incremental Genetic K-means Algorithm
I-PSO-KM	-	Improved Particle Swarm Optimization-K-Means
KFLANN	-	K-means Fast Learning Artificial Neural Network
KM	-	K-Means
K-NN	-	K-Nearest Neighbors
MCBIPSO	-	Mountain Clustering Based on PSO
MCL	-	Markov Cluster Algorithm
MEPSO	-	Multi-Elitist PSO

ML	-	Machine Learning
OPTICS	-	Ordering Points to Identify the Clustering Structure
PSC	-	Particle Swarm Clustering
PSO	-	Particle Swarm Optimization
PSO-KM	-	Particle Swarm Optimization-K-Means
PSOGA-KM	-	Particle Swarm Optimization-Genetic Algorithm-K-Means
SGA	-	Simple Genetic Algorithm
SOM	-	Self Organizing Map
SPMD	-	Single Program Multiple Data algorithm
SSE	-	Sum of the Square Error
Std. Dev.	-	Standard Deviation
TWCV	-	Total Within Cluster Variation
UCI	-	University of California Irvine
UPGMA	-	Un-weighted Pair Group Method with Arithmetic
WCSS	-	Within Cluster Sum of Squares
WKMA	-	Weighted K-Means Algorithm

CHAPTER 1

INTRODUCTION

1.1 Overview

One of the important and constantly developing issues in the world of science is Computer Science (CS), which is the practical and scientific approach used for computation and its related applications. CS studies systematizes the mechanization, feasibility, expression, and structure of methodical algorithms that underlie the acquisition, processing, representation, storage, access to, and communication of information. A significant part of CS is Artificial Intelligence (AI), which includes several subdivisions, e.g., Machine Learning (ML), data mining, and pattern recognition. Among these, ML is of great importance in the field of AI.

ML is a subfield of AI that addresses the construction of systems that are capable of learning from data rather than simply following programmed instructions (Ackerman, 2000; Ayodele, 2010). Additionally, this field is strongly tied with optimization and statistics, delivering both theory and method to the field. ML is applied to various computing tasks in which it is not feasible to design and program algorithms that are explicit and rule based (Gullapalli & Brungi, 2015). There is a conflation among the concepts of ML, pattern recognition, and data mining (Chakrabarti, 2003; Ayodele, 2010).

ML applied to tasks falls into three different types: supervised learning, semi-supervised learning, and unsupervised learning. The supervised learning refers to situations in which a computer is provided with both example inputs and desired outputs, presented by a "teacher"; it aims at learning a general rule that maps inputs to outputs. The supervised learning is a machine learning task through which a function is inferred from labeled training data (Chapelle *et al.*, 2006; Huang *et al.*, 2006; Settles, 2010). A example of supervised learning is classification that is applied to solving problems. On the other hand, in semi-supervised learning, labeled and unlabelled examples are combined for the purpose of generating an appropriate function or classifier (Zhu & Goldberg, 2009). The unsupervised learning-based algorithms are applied to unlabelled inputs in cases where there is not a known desired output. This aims at discovering structures in data through, for example, cluster analysis, rather than generalizing a mapping from input to output (Jain *et al.*, 2000; Ayodele, 2010; Peuquet *et al.*, 2015).

In unsupervised learning, the learning algorithm is not given any label; rather, it is left on its own to cluster similar inputs, perform density estimates, or do the projection of high-dimensional data, the latter of which can be effectively visualized (Berkhin, 2006). Unsupervised learning can be considered as a goal in itself or a means that can be used to achieve a particular end (Tuytelaars et al., 2010). An instance of unsupervised learning is topic modeling through which a list of human language documents is given to a program, and the problem is to explore which document covers similar topics. Clustering is another good example of unsupervised learning in which the clustering is used to solve problems.

In clustering or cluster analysis, a set of objects are grouped. In this method, objects belonging to one group are more similar to each other compared to objects belonging to other groups (Lavanya *et al.*, 2015). This is considered as the most important task of exploratory data mining; additionally, it is known as a general technique for statistical data analysis that is employed in several fields of study such as image analysis, machine learning, bioinformatics, information retrieval, and pattern recognition (Jain, 2010). Cluster analysis is not considered as a specific algorithm per se; rather, it is a general task to be solved. This can be obtained by algorithms that are significantly different in their definition of what forms a cluster and how to find them in an efficient way. Popular conceptions of clusters define them as groups whose members have small distances, intervals, or particular statistical distributions, and dense areas of data space (Jain & Maheswari, 2012). Clustering is a technique of a great importance, which is applied to several fields such as information retrieval and knowledge discovery. Using this technique, scholars are capable of finding related information faster. Therefore, researchers date with new findings in their own field of study (Fayyad et al., 1996). Clustering is a process through which objects are grouped or divided into clusters; the purpose of this process is to place objects that are similar to one another in one cluster and place dissimilar ones within other clusters (Jiang et al., 2004). Grouping is carried out in terms of predefined distance or similarity measure (Berkhin, 2006). At present many studies apply clustering to several areas of investigation, such as classification, decision making, information extraction, and pattern analysis (Xu & Wunsch, 2005). Clustering is divided into a number of models, including connectivity models, distribution models, and centroid models (Xu & Wunsch, 2005; Berkhin, 2006).

K-means clustering, originating from signal processing is a method of vector quantization (Al-Jarrah *et al.*, 2015). This is commonly applied to cluster analysis in data mining. The aim of *K*-means clustering is partitioning n observations into *K* clusters; in this case, each observation belongs to the cluster that has the nearest mean, which serves as a cluster's prototype (Xu & Wunsch, 2005; Dix, 2009; Jain, 2010). The problem has been proved to an NP-hard problem, though a number of efficient heuristic algorithms that have been proposed, which quickly converge to a local optimum. Generally, such algorithms are similar to the expectation-maximization algorithm for mixtures of Gaussian distributions through an iterative refinement approach that is adopted by both algorithms. In addition, both algorithms employ cluster centers for modeling the data. Nevertheless, in the expectation-maximization mechanism, clusters are allowed to have various shapes, whereas *K*-means clustering usually finds clusters of similar spatial extent (Xu & Wunsch, 2005; Celebi *et al.*, 2013) . In the *K*-means clustering algorithms, there are a number of shortages and defects that should be improved.

There are different methods to enhance and improve *K*-means clustering algorithm. One of these methods is to use the optimization method, in which a best element is selected from some of the set of available alternatives. Two important areas pertaining to optimization methods are the hybrid approach and the meta-heuristic approach.

In a hybrid algorithm, two or more algorithms are combined to solve a particular problem. A hybrid algorithm is an algorithm that combines two or more other algorithms that solve the same problem, either choosing one, or switching between them over the course of the algorithm. This is generally done to combine desired features of each, so that the overall algorithm is better than the individual components. Over the course of the hybrid algorithm, one of the algorithms is chosen, which depends on the data, or it is switched between them (Maringer & Kellerer, 2003; Chiarandini *et al.*, 2006). The purpose of this procedure is to combine the desired features of each algorithm, so that the hybrid algorithm could perform better compared to the individual components (Coello *et al.*, 2002; Van den Bergh & Engelbrecht, 2004). Because of the shortcomings that exist in the K-means clustering algorithm, it can be optimized when using in a hybrid algorithm. Two algorithms that are mostly applied to hybrid algorithms are Particle Swarm Optimization (PSO) and the Genetic Algorithm (GA). Given that these algorithms have no label to solve the problem and they do not have additional guides, they can be applied to improvement of *K*-means clustering algorithm performance.

GA is a search heuristic that mimics the natural selection process. This is generally employed for generating practical solutions to search and optimization problems (Mitchell, 1998; Hao *et al.*, 2015). GAs are subsets of the Evolutionary Algorithms (EA) that apply solutions to optimization problems by means of techniques that are inspired by natural evolution, such as selection, mutation, inheritance, and crossover (Whitley, 1994; Kumar *et al.*, 2010). Due to the good performance of the GAs in optimization problems, it can form a hybrid algorithm with *K*-means clustering algorithms. This strategy can remove some of the drawbacks of *K*-means clustering algorithms.

On the other hand, PSO is an optimization algorithm that is globally used to address problems wherein a best solution can be denoted as a surface or point in a space with n dimensions. In this space, hypotheses are plotted and seeded with a communication channel between the particles as well as an initial velocity (Kennedy, 1997; Shi & Eberhart, 1998; Urade & Patel, 2012). Next, particles move all the way through the solution space. Then, after each time step they are assessed based on some fitness criterion (Robinson & Rahmat-Samii, 2004). Over time, particles are speeded up toward the particles positioned in their own communication grouping, which are with better fitness values. The most important advantage of such an approach over other global minimization strategies is that the huge number of members making up the particle swarm make this technique impressively flexible to the local minima problem (Sadeghierad *et al.*, 2010; Shakerian *et al.*, 2011). Due to the good performance of the PSO algorithm in the optimization, it can form a hybrid algorithm together with the *K*-means clustering algorithm. This can eliminate some disadvantages of the *K*-means clustering algorithm.

The heuristic technique has been designed to solve problems better in artificial intelligence, computer science, and mathematical optimization in cases where traditional methods work too slowly, or to find an approximate solution in cases in which the traditional methods cannot find any appropriate solution. This can be obtained through trading optimality, accuracy, completeness, or precision for speed (Renner & Ekárt, 2003).

Additionally, a meta-heuristic is a higher-level procedure that has been proposed to find generate, or choose a lower-level procedure or heuristic that can provide an appropriate solution to an optimization problem, in particular one with incomplete information or a limited capacity of computation in mathematical optimization and computer science (Blum & Roli, 2003; Bianchi *et al.*, 2009; Blum *et al.*, 2011). Meta-heuristics are able to make few assumptions in regard to the optimization problem that is being solved; therefore, they can be practically employed as a solution to various problems. In comparison with the iterative methods and optimization algorithms, meta-heuristics cannot guarantee a globally optimal solution to some classes of problems (Blum & Roli, 2003). Several meta-heuristics put into practice some forms of stochastic optimization in such a way that the solution is dependent on the set of generated random variables (Bianchi et al., 2009). Through searching among several feasible solutions, meta-heuristics is capable of finding appropriate solutions with less computational effort compared to simple heuristics, iterative methods, or algorithms (Blum et al., 2011). Accordingly, meta-heuristics can be considered as a practical approach to optimization problems (Bianchi *et al.*, 2009; Blum *et al.*, 2011). Generally, if two different algorithms are combined for solving a particular problem, the method is called a hybrid approach. However, if more than two algorithms are combined for solving a problem or several heuristic algorithms are combined for solving problem, the method is called a meta-heuristic approach. A hybrid of the GA algorithm and *K*-means clustering algorithm has advantages for good clustering, and a hybrid of the PSO algorithm and the *K*-means clustering algorithm has other advantages. It can be combined with the methods mentioned above in order to obtain an algorithm that combines the advantages of both algorithms. The result will be a meta-heuristic approach, the result of which is better than the previous method for clustering data.

In this research, the proposed algorithms of a hybrid of the Improved Genetic Algorithm in *K*-means (I-GA-*K*M), a hybrid of the Improved Particle Swarm Optimization in *K*-means (I-PSO-*K*M), a meta-heuristic of the Genetic Algorithm and Particle Swarm Optimization in *K*-means (GAPSO-*K*M), and a meta-heuristic of Particle Swarm Optimization and Genetic Algorithm in *K*-means (PSOGA-*K*M) are proposed for real and binary data. The proposed algorithms are evaluated using the standard data sets and used for developing *K*-means algorithm. In this thesis the data was collected from University of California Irvine (UCI) standard data set in all experiments and all proposed algorithms. It used six integer data sets including Balance, Blood, Breast, Iris, Pima and Wine.

1.2 Problem Background

For the first time, the term "*k*-means" was introduced by James MacQueen in 1967 (MacQueen, 1967; Gayathri *et al.*, 2015); however, the idea originally belonged to Hugo Steinhaus (Steinhaus, 1956). Stuart Lloyd was pioneer in proposing the standard algorithm in 1957. It was applied as a technique to pulse-code modulation; however, it was not published until 1982 (Lloyd, 1982). In 1965, the same method was published by Forgy, which is sometimes named Lloyd-Forgy (Forgy, 1965). A more efficient version was published by (Hartigan & Wong, 1979).

A set of observations (x1, x2, …, xn) is given to the *K*-means clustering algorithm, in which each observation is a d-dimensional real vector. The aim of the *K*-means clustering is partitioning the n observations into *K* (≤ n) sets S = { S1, S2, …, Sk } in order to reduce the Within-Cluster Sum of Squares (WCSS) as far as possible (Patel & Sinha, 2010; Ramamurthy & Chandran, 2011; Singh *et al.*, 2011). The standard *K*-means algorithm makes use of an iterative refinement technique. Because of its ubiquity, it often is known as a *K*-means algorithm; it is also named Lloyd's algorithm, particularly in the computer science community. When an initial set of *K* means $m_1,\ldots,m_k$ is given to the algorithm, it proceeds through alternating between two steps: the assignment step and the update step (Patel & Sinha, 2010). In the former, each observation is assigned to the cluster to which its mean yields the least WCSS. As the sum of squares is squared Euclidean distance, this mean intuitively the "nearest" one (Utro, 2011). In the latter, the new means are calculated to be centroids of observations in new clusters. Initialization methods for the *K*-means algorithm fall into two methods, namely Forgy and Random Partition (Faber, 1994; Redmond & Heneghan, 2007).

In the Forgy method, *K* observations are randomly selected from among the data set and used as the initial means (Forgy, 1965; Hamerly & Elkan, 2002). Hamerly et al., (Hamerly & Elkan, 2002) state that, in general, the Random Partition method is preferable for algorithms such as fuzzy *K*-means and the *K*-harmonic means. However, in case of standard *K*-means algorithms and expectation maximization, the Forgy method of initialization is considered preferable (Forgy, 1965; Shirwaikar & Bhandari, 2013). Since this is a heuristic algorithm, there is not any guarantee that it will be converged to global optimum, and the results may be dependent on the initial clusters. Since this is typically a very fast algorithm, it is commonly run for multiple times with various starting conditions(Gariel *et al.*, 2011).

Krovi is the pioneer in investigating the potential applicability of GAs to clustering (Krovi, 1992; Sheikh *et al.*, 2008). A new hybrid GA introduced by K. Krishna and M. N. Murty attempted to find a globally optimal partition of a certain data into a defined number of clusters. The idea behind Fast Genetic *K*-means Algorithm (FG*K*A) (Lu *et al.*, 2004a; Sheikh *et al.*, 2008) came from GKA; however, FGKA had a number of improvements compared to GKA. The experiments conducted in this area indicated that, when *K*-means algorithm are converted to a local optimum, both GKA and FGKA always finally converge to the global optimum, even though FGKA runs with a much higher speed compared to GKA. Incremental Genetic *K*-means Algorithm (IG*K*A) (Lu et al., 2004b) was actually an extension to FG*K*A. Jie *et al.* (Jie *et al.*, 2004) proposed a new clustering algorithm for the mixed data sets through the modification of the common cost function and trace of the within cluster dispersion matrix. Liu *et al.* (Liu *et al.*, 2004) introduced HGA-clustering that was a hybrid genetic-based clustering algorithm in order to find the appropriate clustering of data sets.

To design the dissimilarity measure, Genetic Distance Measure (GDM), which was a genetic algorithm, was proposed in a way to improve the K-modes algorithm performance (Chiang et al., 2006). Demiriz *et al.* (Demiriz *et al.*, 1999) designed a semi-supervised clustering algorithm that was a combination of the benefits of unsupervised and supervised learning methods. The *K*-means Fast Learning Artificial Neural Network (*K*FLANN) that was introduced by Xiang and Phuan (Xiang & Phuan, 2005) was a small neural network with two types of parameters: vigilance, μ, and the tolerance, δ. Single Program Multiple Data algorithm (SPMD) proposed by Du *et al.* (Du *et al.*, 2001) combined GA with uphill that was local searching algorithm. A hybrid of GA and a Weighted *K*-Means Algorithm (WKMA) was proposed by Fang-Xiang et al (Wu, 2008) and termed Genetic Weighted *K*-means Algorithm (GW*K*MA). A hybrid GA-based clustering (HGACLUS) schema introduced by Pan *et al.* (Pan *et al.*, 2003) combined merits of Simulated Annealing. It was presented to find an optimal or near-optimal set of medoids. Katari *et al.* (Katari *et al.*, 2007) introduced data clustering by means of improved IGA to which an efficient method of crossover and mutation was applied (Sheikh et al., 2008).

Omran *et al.* (Omran *et al.*, 2002) designed PSO for clustering through a straightforward implementation. Their algorithm used a fixed clusters number and employed PSO in order to search for these clusters' optimal centroids. Using PSO, Van der Merwe and Engelbrecht (Van der Merwe & Engelbrecht, 2003) introduced two new approaches to cluster data. They demonstrated how PSO could be employed for finding the centroids of a user-specified number of clusters. Fun and Chen (Chen & Ye, 2004) designed PSO-clustering, a technique based on the particle swarm optimization algorithm. They applied the particle swarm optimization to searching automatically for the center of a cluster in the arbitrary dataset. Cui *et al.* (Cui *et al.*, 2005) proposed a PSO document clustering algorithm, which performed a global search within the entire answer space. Cohen and de Castro (Cohen & de Castro, 2006) presented a proposal on data clustering, which was based on the PSO algorithm, which was adapted to place prototypes within regions of the space that denote the natural clusters of input dataset. Abraham *et al.* (Abraham *et al.*, 2007) proposed a method to cluster the complex and linearly non-separable datasets, with no prior knowledge regarding the number of naturally occurring clusters. Their method was based on an improved version of PSO algorithm. Esmin *et al.* (Esmin *et al.*, 2008) introduced two new data clustering approaches by means of the PSO algorithm. This could be employed for finding centroids of a user-specified number of clusters. Sharma and Omlin (Sharma & Omlin, 2009) proposed the use of an adaptive heuristic PSO algorithm to find cluster boundaries directly from code vectors obtained from Self-Organizing Map (SOM). Dong and Qi (Dong & Qi, 2009b) introduced a new clustering algorithm based on PSO (Abul Hasan & Ramakrishnan, 2011; Sethi & Mishra, 2013; Bollmann *et al.*, 2015). Therefore, to enhance the performance of previous hybrid methods, four algorithms are proposed in this study. Among them, two algorithms are designed to improve GA-*K*-means and PSO-*K*-means algorithm; and the other two algorithms are meta-heuristics algorithms obtained from the two previous algorithms with meta-heuristic method. Attempts have been made to overcome with disadvantages. The last two algorithms are named Genetic Algorithm-Particle Swarm Optimization-*K*M (GAPSO-*K*M) and Particle Swarm Optimization-Genetic Algorithm-*K*M (PSOGA-*K*M), which can be applied to clustering dataset.

1.3 Problem Statement

Traditional optimization algorithms cannot provide proper results for clustering problems with high error, high intra cluster distance and low accuracy rate since the result is sensitive to the selection of initial cluster centers and this converges simply to local optima. In recent years, to solve the data clustering problem, several new approaches have been introduced, inspired from biological sciences, including Genetic Algorithm, Particle Swarm Optimization algorithm, and so on. Also, existing hybrid algorithms with K-means clustering suffer from different drawbacks such as lack of providing optimum solution for all problems, getting stuck in local optima, tuning many parameters, slow convergence rate, high number of error and high intra cluster distance. Also, existing meta-heuristic algorithms with *K*-means clustering have low accuracy rate of the clustering and low the number of correct answers, they have good performance only in one of the search spaces. However, the algorithms are robust and have the ability of adapting with changing environment.

Therefore, more works are still required to develop the performance of hybrid and meta-heuristic algorithms in *K*-means clustering algorithm. Hence, new hybrid and meta-heuristic algorithms are introduced in the study to cope with the shortcomings of clustering.

Hence, the hypothesis of the study can be stated as:

The Genetic Algorithm and the Partial Swarm Optimization Algorithm could yield better accuracy for the K-means clustering algorithm.

Therefore, based on the above issues, the main research question is:

Are the proposed hybrid optimized algorithms beneficial for enhancement of the K-means clustering learning?

In order to answer the main issue raised above, the following questions need to be addressed:

i. How to propose an improved hybrid GA-K-means scheme for error reduction?
ii. How to develop a hybrid PSO-K-means scheme to reduce the intra-cluster distance?
iii. How to design and develop the meta-heuristic of GAPSO and PSOGA with K-means algorithm for better accuracy?

1.4 Aim of the Research

The aim of this research is to develop and enhance the *K*-means clustering algorithm using the proposed Improved Genetic Algorithm in *K*-means (I-GA-*K*-means), Improved Particle Swarm Optimization Algorithm in *K*-means (I-PSO-*K*-means), hybrid Genetic Algorithm and Particle Swarm Optimization Algorithm in *K*-means (GAPSO-*K*-means), hybrid Particle Swarm Optimization Algorithm, and Genetic Algorithm in *K*-means (PSOGA-*K*-means) algorithms and reduce the error rate, iteration, related processing time, intra-cluster distance and increase the accuracy rate.

1.5 Research Objectives

In order to reach the answers to the above questions, the objectives of this research have been identified as:

i. To propose an improved hybrid GA-K-means scheme for error reduction.
ii. To develop a hybrid PSO-K-means scheme to reduce the intra-cluster distance.

iii. To design and develop the meta-heuristic of GAPSO and PSOGA with K-means algorithm for better accuracy.

1.6 Scope of Study

To achieve the above objectives, the scope of this study is bounded to the following limitations:

i. This study will identify, analyze and improve the *K*-means clustering algorithm.
ii. In this study, six UCI standard data sets are applied to binary and multi classification problems and clustering: balance, blood, breast, iris, pima and wine.
iii. The focus will be on the improvement of hybrid GA-*K*-means algorithm in the first phase, optimized hybrid methods for improving PSO-*K*-means clustering algorithm in the second phase, and the use of meta-heuristic method in the *K*-means algorithm for developing in third phase.
iv. The comparisons criteria are average, standard deviation, best, and worst. While the comparison factors are intra-cluster distance, number of iterations, number of correct answer, number of errors, error rate, related processing time and accuracy rate.
v. This study concentrates on the minimization of intra-cluster distance.
vi. The programs have been customized, developed, and applied to the problems using MATLAB R2012b software.

1.7 Importance of Study

The study investigates the capabilities of *K*-means in the clustering algorithm. In addition, it develops a clustering algorithm and attempts to eliminate the disadvantages of the clustering algorithm. The significance of this research that is removes shortages of the clustering algorithm and developing it using hybrid methods and optimization algorithms. This research helps to enhance clustering algorithm and develop the clustering. This study uses four algorithms, including I-GA-*K*-means, I-PSO-*K*-means, GAPSO-*K*-means, and PSOGA-*K*-means algorithms for enhancement of the clustering algorithm. The performance of the proposed methods is evaluated to examine whether the proposed algorithms are able to decrease intra-cluster distance, iteration, error rate, related processing time and to increase the accuracy rate.

The potential applications by using proposed methods include: computational finance, adaptive websites, affective computing, bioinformatics, game playing, sequence mining, structural health monitoring, software engineering, search engines, recommender systems, medical diagnosis, brain-machine interfaces, computer vision, optimization and meta-heuristic.

1.8 Thesis Organization

This section presents a brief overview of the contents of this thesis. This study is organized into six chapters. The first is the introductory chapter. The second chapter describes the background as well as the previously-published studies in the field of clustering algorithms. The third chapter describes the research methodology of this study. Chapter Four and Five provide the proposed methods and the analysis of the obtained results in terms of improving the clustering algorithm. Finally, the summary of this study is presented in Chapter Six. The details of each chapter are as follows:

Chapter1, Introduction, the statement of the study is presented. It starts with the introduction of the study followed by the background of the study. The problem statement, objectives, aim, scope, contribution, and limitations are also presented in this chapter. The structure of the study is organized at the end of this chapter.

Chapter 2, Literature Review, a review is done on the literature related to all major areas of our study: data clustering, clustering algorithm, optimization methods, optimization hybrid algorithm, and meta-heuristic algorithm for improving of clustering algorithm. Finally, the discussion and summary of this chapter are given.

Chapter 3, Research Methodology, presents the methodology adopted for this study, including a general framework for three phases of the study and descriptions about the overall tools and standard techniques. Three phases of this research are explained in this chapter.

Chapter 4, Hybridization of *K*-Means Algorithm with GA and PSO, presents the methodology, design, flowchart, coding and the UCI dataset for evaluation of performance for the first and second proposed algorithms. In this chapter, the clustering algorithm is improved by using I-GA-*K*-means and I-PSO-*K*-means algorithms, which they are present in this chapter.

Chapter 5, Design Meta-Heuristic of GAPSO and PSOGA with *K*-Means, presents the methodology, design, coding and the UCI dataset for evaluation of performance for the third and fourth proposed algorithms. This chapter reports the results of experiments conducted on two algorithms, GAPSO-*K*-means and PSOGA-*K*-means. Then, the obtained results are evaluated in regard to various criteria, i.e., intra-cluster distance, accuracy, and number of errors. A comparison shows that the proposed methods have answers with high accuracy.

Finally, in Chapter 6, Conclusion and Future Works, the research is concluded, discussed, along with highlights of the contributions and findings of the research. This chapter also provides suggestions and recommendations for future studies.

government agency data. In addition, other applications of clustering algorithm in medical image such as sonography images, mammography images and radiology images can be studied.

Additionally, the proposed algorithms can be employed in other NP-hard problems and combinatorial optimization problems.

Furthermore, other methods such as fuzzy set and rough set were used for evaluating the performance of the proposed algorithm with new methods.

REFERENCES

Abadi, M. F. H. and H. Rezaei (2015). Data Clustering Using Hybridization Strategies of Continuous Ant Colony Optimization, Particle Swarm Optimization and Genetic Algorithm. *British Journal of Mathematics & Computer Science* 6(4): 336.

Abdel-Kader, R. F. (2010). Genetically improved PSO algorithm for efficient data clustering. *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, IEEE.

Abdeyazdan, M. (2014). Data clustering based on hybrid K-harmonic means and modifier imperialist competitive algorithm. *The Journal of Supercomputing* 68(2): 574-598.

Abraham, A., S. Das and A. Konar (2007). Kernel based automatic clustering using modified particle swarm optimization algorithm. *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, ACM.

Abul Hasan, M. J. and S. Ramakrishnan (2011). A survey: hybrid evolutionary algorithms for cluster analysis. *Artificial Intelligence Review* 36(3): 179-204.

Ackerman, M. S. (2000). The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15(2-3): 179-203.

Ackermann, M. R., M. Märtens, C. Raupach, K. Swierkot, C. Lammersen and C. Sohler (2012). StreamKM++: A clustering algorithm for data streams. *Journal of Experimental Algorithmics (JEA)* 17: 2.4.

Aghdasi, T., J. Vahidi and H. Motameni (2014). K-harmonic means Data Clustering Using Combination of Particle Swarm Optimization and Tabu Search.

Ahmad, A. and L. Dey (2007). A< i> k</i>-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering* 63(2): 503-527.

Ahmadyfard, A. and H. Modares (2008). Combining PSO and k-means to Enhance Data Clustering. *Telecommunications, 2008. IST 2008. International Symposium on*, IEEE.

Al-Jarrah, O. Y., P. D. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha (2015). Efficient Machine Learning for Big Data: A Review. *Big Data Research*.

Al-Shboul, B. and S.-H. Myaeng (2009). Initializing K-Means using genetic algorithms. *World Academy of Science, Engineering and Technology* 54: 114-118.

Alam, S., G. Dobbie and P. Riddle (2008). An evolutionary particle swarm optimization algorithm for data clustering. *Swarm Intelligence Symposium, 2008. SIS 2008. IEEE*, IEEE.

Arthur, D., B. Manthey and H. Röglin (2011). Smoothed analysis of the k-means method. *Journal of the ACM (JACM)* 58(5): 19.

Arthur, D. and S. Vassilvitskii (2007). k-means++: The advantages of careful seeding. *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, Society for Industrial and Applied Mathematics.

Asuncion, A. and D. Newman (2007). UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2007).

Ayodele, T. O. (2010). Types of machine learning algorithms. *Internet: http://www. intechopen. com/articles/show/title/types-of-machinelearning-algorithms*.

Bandyopadhyay, S. and U. Maulik (2001). Nonparametric genetic clustering: comparison of validity indices. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on* 31(1): 120-125.

Bandyopadhyay, S. and U. Maulik (2002). Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern Recognition* 35(6): 1197-1208.

Behera, H., R. B. Lingdoh and D. Kodamasingh (2011). An Improved Hybridized k-Means Clustering Algorithm (ihkmca) for Highdimensional Dataset & It's Performance Analysis. *International journal of Computer science & Engineering* 3: 1183-1190.

Ben-David, S., D. Pál and H. U. Simon (2007). Stability of k-means clustering.*Learning Theory* 20-34, Springer.

Bensaid, A. M., L. O. Hall, J. C. Bezdek and L. P. Clarke (1996). Partially supervised clustering for image segmentation. *Pattern Recognition* 29(5): 859-871.

Berkhin, P. (2006). A survey of clustering data mining techniques.*Grouping multidimensional data* 25-71, Springer.

Beyer, K., J. Goldstein, R. Ramakrishnan and U. Shaft (1999). When is "nearest neighbor" meaningful?*Database Theory—ICDT'99* 217-235, Springer.

Bezdek, J. C. and N. R. Pal (1998). Some new indexes of cluster validity. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 28(3): 301-315.

Bhuvaneswari, K., M. Anusha and J. Sathiaseelan (2015). A Comparative Analysis of Clustering Techniques using Genetic Algorithm.

Bianchi, L., M. Dorigo, L. M. Gambardella and W. J. Gutjahr (2009). A survey on metaheuristics for stochastic combinatorial optimization. *Natural Computing: an international journal* 8(2): 239-287.

Blum, C., J. Puchinger, G. R. Raidl and A. Roli (2011). Hybrid metaheuristics in combinatorial optimization: A survey. *Applied Soft Computing* 11(6): 4135-4151.

Blum, C. and A. Roli (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM computing surveys (CSUR)* 35(3): 268-308.

Bollmann, S., A. Hölzl, M. Heene, H. Küchenhoff and M. Bühner (2015). Evaluation of a new k-means approach for exploratory clustering of items.

Cai, L., X. Yao, Z. He and X. Liang (2010). K-means clustering analysis based on immune genetic algorithm. *World Automation Congress (WAC), 2010*, IEEE.

Celebi, M. E., H. A. Kingravi and P. A. Vela (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications* 40(1): 200-210.

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann.

Chang, D.-X., X.-D. Zhang and C.-W. Zheng (2009). A genetic algorithm with gene rearrangement for K-means clustering. *Pattern Recognition* 42(7): 1210-1222.

Chapelle, O., B. Schölkopf and A. Zien (2006). *Semi-supervised learning*. MIT press Cambridge.

Chau, M., R. Cheng and B. Kao (2005). Uncertain data mining: a new research direction. *Proceedings of the Workshop on the Sciences of the Artificial, Hualien, Taiwan*.

Chen, C.-Y. and F. Ye (2004). Particle swarm optimization algorithm and its application to clustering analysis. *Networking, Sensing and Control, 2004 IEEE International Conference on*, IEEE.

Cheung, Y.-M. (2003). k∗-Means: A new generalized< i> k</i>-means clustering algorithm. *Pattern Recognition Letters* 24(15): 2883-2893.

Chiang, S., S.-C. Chu, Y.-C. Hsin and M.-H. Wang (2006). Genetic Distance measure for K-modes Algorithm. *International Journal of Innovative Computing, Information and Control* 2(1): 33-40.

Chiarandini, M., M. Birattari, K. Socha and O. Rossi-Doria (2006). An effective hybrid algorithm for university course timetabling. *Journal of Scheduling* 9(5): 403-432.

Chuang, L.-Y., C.-J. Hsiao and C.-H. Yang (2011). Chaotic particle swarm optimization for data clustering. *Expert Systems with Applications* 38(12): 14555-14563.

Coello, C. A. C., D. A. Van Veldhuizen and G. B. Lamont (2002). *Evolutionary algorithms for solving multi-objective problems*. Springer.

Cohen, S. C. and L. N. de Castro (2006). Data clustering with particle swarms. *Evolutionary Computation, 2006. CEC 2006. IEEE Congress on*, IEEE.

Cui, X., T. E. Potok and P. Palathingal (2005). Document clustering using particle swarm optimization. *Swarm Intelligence Symposium, 2005. SIS 2005. Proceedings 2005 IEEE*, IEEE.

Danesh, M., M. Naghibzadeh, M. R. A. Totonchi, M. Danesh, B. Minaei and H. Shirgahi (2011). Data clustering based on an efficient hybrid of K-harmonic means, PSO and GA.*Transactions on computational collective intelligence IV* 125-140, Springer.

Das, S., A. Abraham and A. Konar (2008). Automatic kernel clustering with a multi-elitist particle swarm optimization algorithm. *Pattern Recognition Letters* 29(5): 688-699.

Davies, D. L. and D. W. Bouldin (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* (2): 224-227.

Demiriz, A., K. P. Bennett and M. J. Embrechts (1999). Semi-supervised clustering using genetic algorithms. *Artificial neural networks in engineering (ANNIE-99)*: 809-814.

Dix, A. (2009). *Human-computer interaction*. Springer.

Dong, J. and M. Qi (2009a). K-means Optimization Algorithm for Solving Clustering Problem. *Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on*, IEEE.

Dong, J. and M. Qi (2009b). A new clustering algorithm based on PSO with the jumping mechanism of SA. *Intelligent Information Technology Application, 2009. IITA 2009. Third International Symposium on*, IEEE.

Dorigo, M. and M. Birattari (2010). Ant colony optimization.*Encyclopedia of Machine Learning* 36-39, Springer.

Du, Z., M. Ding, S. Li, S. Li, M.-y. Wu and J. Zhu (2001). Massively Parallel SPMD Algorithm for Cluster Computing--Combining Genetic Algorithm with Uphill. *JOURNAL-SHANGHAI UNIVERSITY* 5(SUPP): 10-14.

Duwairi, R. and M. Abu-Rahmeh (2015). A novel approach for initializing the spherical K-means clustering algorithm. *Simulation Modelling Practice and Theory* 54: 49-63.

Eberhart, R. C. and Y. Shi (1998). Comparison between genetic algorithms and particle swarm optimization. *Evolutionary Programming VII*, Springer.

Elavarasi, S. A., J. Akilandeswari and B. Sathiyabhama (2011). A survey on partition clustering algorithms. *International Journal of Enterprise Computing and Business Systems* 1(1).

Eltibi, M. F. and W. M. Ashour (2011). Initializing K-means Clustering Algorithm using Statistical Information. *International Journal of Computer Applications* 29(7).

Entriken, R. and S. Vössner (1997). Genetic algorithms with cluster analysis for production simulation. *Proceedings of the 29th conference on Winter simulation*, IEEE Computer Society.

Erisoglu, M., N. Calis and S. Sakallioglu (2011). A new algorithm for initial cluster centers in< i> k</i>-means algorithm. *Pattern Recognition Letters* 32(14): 1701-1705.

Esmin, A. A. A., D. L. Pereira and F. De Araujo (2008). Study of different approach to clustering data by using the particle swarm optimization algorithm. *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*, IEEE.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. *ACM SIGKDD Explorations Newsletter* 4(1): 65-75.

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science* 22: 138-144.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM* 39(11): 27-34.

Firouzi, B., T. Niknam and M. Nayeripour (2008). A new evolutionary algorithm for cluster analysis. *World Academy of Science, Engineering, and Technology* 36: 605-609.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21: 768-769.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458): 611-631.

Galluccio, L., O. Michel, P. Comon and A. O. Hero III (2012). Graph based k-means clustering. *Signal Processing* 92(9): 1970-1984.

Ganesh, A. D. S. H., D. P. Cindrella and A. J. Christy (2015). A REVIEW ON CLASSIFICATION TECHNIQUES OVER AGRICULTURAL DATA.

Gao, X., B. Xiao, D. Tao and X. Li (2010). A survey of graph edit distance. *Pattern Analysis and applications* 13(1): 113-129.

Gariel, M., A. N. Srivastava and E. Feron (2011). Trajectory clustering and an application to airspace monitoring. *Intelligent Transportation Systems, IEEE Transactions on* 12(4): 1511-1524.

Gayathri, R., A. Cauveri, R. Kanagapriya, V. Nivetha, P. Tamizhselvi and K. P. Kumar (2015). A Novel Approach for Clustering Based On Bayesian Network. *Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015)*, ACM.

Gheyas, I. A. (2009). Novel computationally intelligent machine learning algorithms for data mining and knowledge discovery.

Ghorpade-Aher, J. and V. A. Metre (2014). Clustering Multidimensional Data with PSO based Algorithm. *arXiv preprint arXiv:1402.6428*.

Glickman, M., J. Balthrop and S. Forrest (2005). A machine learning evaluation of an artificial immune system. *Evolutionary Computation* 13(2): 179-212.

Goldberg, D. E. and J. H. Holland (1988). Genetic algorithms and machine learning. *Machine learning* 3(2): 95-99.

Gómez, N., L. F. Mingo, J. Bobadilla, F. Serradilla and J. A. C. Manzano (2010). Particle Swarm Optimization models applied to Neural Networks using the R language. *WSEAS Transactions on Systems* 9(2): 192-202.

Gorgônio, F. L. and J. A. F. Costa (2010). PartSOM: A Framework for Distributed Data Clustering Using SOM and K-Means.

Gu, C. and Q. Tao (2015). Clustering Algorithm Combining CPSO with K-Means. *2015 International Conference on Advances in Mechanical Engineering and Industrial Informatics*, Atlantis Press.

Guan, Y., A. A. Ghorbani and N. Belacel (2004). K-means+: An autonomous clustering algorithm. *Submitted to Pattern Recognition*.

Gullapalli, V. K. and R. Brungi (2015). A Novel Methodology to Implement Optimization Algorithms in Machine Learning. *International Journal of Computer Applications* 112(4).

Hall, L. O., I. B. Ozyurt and J. C. Bezdek (1999). Clustering with a genetically optimized approach. *Evolutionary Computation, IEEE Transactions on* 3(2): 103-112.

Hamerly, G. and C. Elkan (2002). Alternatives to the k-means algorithm that find better clusterings. *Proceedings of the eleventh international conference on Information and knowledge management*, ACM.

Hao, J.-X., Y. Yu, R. Law and D. K. C. Fong (2015). A genetic algorithm-based learning approach to understand customer satisfaction with OTA websites. *Tourism Management* 48: 231-241.

Hartigan, J. A. and M. A. Wong (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied statistics*: 100-108.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press.

Hu, X., Y. Shi and R. C. Eberhart (2004). Recent advances in particle swarm. *IEEE congress on evolutionary computation*.

Huang, H., Q. Tang and Z. Liu (2013). Adaptive Correction Forecasting Approach for Urban Traffic Flow Based on Fuzzy-Mean Clustering and Advanced Neural Network. *Journal of Applied Mathematics* 2013.

Huang, T.-M., V. Kecman and I. Kopriva (2006). *Kernel based algorithms for mining huge data sets: Supervised, semi-supervised, and unsupervised learning*. Springer.

Huang, X. and W. Su (2014). An Improved K-means Clustering Algorithm. *Journal of Networks* 9(01): 161-167.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31(8): 651-666.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jain, A. K., R. P. W. Duin and J. Mao (2000). Statistical pattern recognition: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22(1): 4-37.

Jain, A. K. and S. Maheswari (2012). Survey of recent clustering techniques in data mining. *Int. J. Comput. Sci. Manage. Res* 1: 72-78.

Jain, A. K., M. N. Murty and P. J. Flynn (1999). Data clustering: a review. *ACM computing surveys (CSUR)* 31(3): 264-323.

Janson, S. and D. Merkle (2005). A new multi-objective particle swarm optimization algorithm using clustering applied to automated docking.*Hybrid Metaheuristics* 128-141, Springer.

Jiang, D., C. Tang and A. Zhang (2004). Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on* 16(11): 1370-1386.

Jiawei, H. and M. Kamber (2001). Data mining: concepts and techniques. *San Francisco, CA, itd: Morgan Kaufmann* 5.

Jie, L., G. Xinbo and J. Li-Cheng (2004). A CSA-based clustering algorithm for large data sets with mixed numeric and categorical values. *Intelligent Control and Automation, 2004. WCICA 2004. Fifth World Congress on*, IEEE.

Jordan, M. I. and D. E. Rumelhart (1992). Forward models: Supervised learning with a distal teacher. *Cognitive science* 16(3): 307-354.

Juang, C.-F. (2004). A hybrid of genetic algorithm and particle swarm optimization for recurrent network design. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 34(2): 997-1006.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

Kanungo, T., D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman and A. Y. Wu (2000). The analysis of a simple k-means clustering algorithm. *Proceedings of the sixteenth annual symposium on Computational geometry*, ACM.

Kanungo, T., D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 24(7): 881-892.

Kao, I., C. Tsai and Y. Wang (2007). An effective particle swarm optimization method for data clustering. *Industrial Engineering and Engineering Management, 2007 IEEE International Conference on*, IEEE.

Kao, Y.-T., E. Zahara and I.-W. Kao (2008). A hybridized approach to data clustering. *Expert Systems with Applications* 34(3): 1754-1762.

Karami, A. and M. Guerrero-Zapata (2015). A fuzzy anomaly detection system based on hybrid pso-kmeans algorithm in content-centric networks. *Neurocomputing* 149: 1253-1269.

Katari, V., S. C. Satapathy, J. Murthy and P. P. Reddy (2007). Hybridized improved genetic algorithm with variable length chromosome for image clustering. *IJCSNS International Journal of COmputer Science and Network Security* 7(11): 121-131.

Kaur, N., J. K. Sahiwal and N. Kaur (2012). Efficient k-means clustering algorithm using ranking method in data mining. *International Journal of Advanced Research in Computer Engineering & Technology* 1(3).

Kennedy, J. (1997). The particle swarm: social adaptation of knowledge. *Evolutionary Computation, 1997., IEEE International Conference on*, IEEE.

Kim, K.-j. and H. Ahn (2008). A recommender system using GA< i> K</i>-means clustering in an online shopping market. *Expert Systems with Applications* 34(2): 1200-1209.

Krishna, K. and M. N. Murty (1999). Genetic K-means algorithm. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 29(3): 433-439.

Krishnasamy, G., A. J. Kulkarni and R. Paramesran (2014). A hybrid approach for data clustering based on modified cohort intelligence and K-means. *Expert Systems with Applications* 41(13): 6009-6016.

Krovi, R. (1992). Genetic algorithms for clustering: a preliminary investigation. *System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference on*, IEEE.

Kudova, P. (2007). Clustering genetic algorithm. *Database and Expert Systems Applications, 2007. DEXA'07. 18th International Workshop on*, IEEE.

Kumar, M., M. Husian, N. Upreti and D. Gupta (2010). Genetic algorithm: Review and application. *International Journal of Information Technology and Knowledge Management* 2(2): 451-454.

Kumar, Y. and G. Sahoo (2014). A charged system search approach for data clustering. *Progress in Artificial Intelligence* 2(2-3): 153-166.

Kuo, R., Y. Syu, Z.-Y. Chen and F.-C. Tien (2012). Integration of particle swarm optimization and genetic algorithm for dynamic clustering. *Information Sciences* 195: 124-140.

Lai, J. Z., T.-J. Huang and Y.-C. Liaw (2009). A fast k-means clustering algorithm using cluster center displacement. *Pattern Recognition* 42(11): 2551-2556.

Lavanya, R., V. Saraswathy and N. Kasthuri (2015). Improving Network Intrusion Detection Based on Multi Objective Criteria.

Lee, K. Y. and M. A. El-Sharkawi (2008). *Modern heuristic optimization techniques: theory and applications to power systems*. John Wiley & Sons.

Li, H., H. Wang and Z. Chen (2015). An Improved K-means Clustering Algorithm for Complex Networks. *First International Conference on Information Science and Electronic Technology (ISET 2015)*, Atlantis Press.

Liang, X., F. Qu, Y. Yang and H. Cai (2015). A Highly Efficient Fast Global K-Means Clustering Algorithm. *2nd International Conference on Civil, Materials and Environmental Sciences*, Atlantis Press.

Likas, A., N. Vlassis and J. J Verbeek (2003). The global< i> k</i>-means clustering algorithm. *Pattern Recognition* 36(2): 451-461.

Lin, H.-J., F.-W. Yang and Y.-T. Kao (2005). An efficient GA-based clustering technique. *Tamkang Journal of Science and Engineering* 8(2): 113.

Liu, H. and H. Motoda (2007). *Computational methods of feature selection*. CRC Press.

Liu, Y. G., K. F. Chen and X. M. Li (2004). A hybrid genetic based clustering algorithm. *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*.

Lloyd, S. (1982). Least squares quantization in PCM. *Information Theory, IEEE Transactions on* 28(2): 129-137.

Lu, Y., S. Lu, F. Fotouhi, Y. Deng and S. J. Brown (2004a). FGKA: A fast genetic k-means clustering algorithm. *Proceedings of the 2004 ACM symposium on Applied computing*, ACM.

Lu, Y., S. Lu, F. Fotouhi, Y. Deng and S. J. Brown (2004b). Incremental genetic K-means algorithm and its application in gene expression data analysis. *BMC bioinformatics* 5(1): 172.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, California, USA.

Maringer, D. and H. Kellerer (2003). Optimization of cardinality constrained portfolios with a hybrid local search algorithm. *Or Spectrum* 25(4): 481-495.

Maulik, U. and S. Bandyopadhyay (2000). Genetic algorithm-based clustering technique. *Pattern Recognition* 33(9): 1455-1465.

Meilă, M. (2006). The uniqueness of a good optimum for k-means. *Proceedings of the 23rd international conference on Machine learning*, ACM.

Mishra, B. K., A. Rath, N. R. Nayak and S. Swain (2012). Far efficient K-means clustering algorithm. *Proceedings of the International Conference on Advances in Computing, Communications and Informatics*, ACM.

Mitchell, M. (1998). *An introduction to genetic algorithms*. MIT press.

Mladenović, N., J. Brimberg, P. Hansen and J. A. Moreno-Pérez (2007). The< i> p</i>-median problem: A survey of metaheuristic approaches. *European Journal of Operational Research* 179(3): 927-939.

Mohri, M., A. Rostamizadeh and A. Talwalkar (2012). *Foundations of machine learning*. MIT press.

Mohtashami, A., M. Tavana, F. J. Santos-Arteaga and A. Fallahian-Najafabadi (2015). A novel multi-objective meta-heuristic model for solving cross-docking scheduling problems. *Applied Soft Computing* 31: 30-47.

Moshizi, M. M., V. K. Bardsiri and E. Heydarabadipour (2015). The Application of Meta-Heuristic based Clustering Techniques in Wireless Sensor Networks. *International Journal of Control and Automation* 8(3): 319-328.

Murthy, C. A. and N. Chowdhury (1996). In search of optimal clusters using genetic algorithms. *Pattern Recognition Letters* 17(8): 825-832.

Mythili, S. and A. S. Kumar (2012). A Proficient Accomplishment of Datamania of Genetic Algorithm by applying K-means Clustering. *International Journal* 1(1).

Na, S., L. Xumin and G. Yong (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm. *Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on*, IEEE.

Nazeer, K. A. and M. Sebastian (2009). Improving the Accuracy and Efficiency of the k-means Clustering Algorithm. *Proceedings of the World Congress on Engineering*.

Omran, M., A. Salman and A. Engelbrecht (2005). Dynamic clustering using particle swarm optimization with application in unsupervised image classification. *Fifth World Enformatika Conference (ICCI 2005), Prague, Czech Republic*.

Omran, M., A. Salman and A. P. Engelbrecht (2002). Image classification using particle swarm optimization. *Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning*, Singapore.

Pan, H., J. Zhu and D. Han (2003). Genetic algorithms applied to multi-class clustering for gene expression data. *Genomics, Proteomics, Bioinformatics* 1(4): 279-287.

Parsa, S. and O. Bushehrian (2007). Genetic clustering with constraints. *Journal of research and practice in information technology* 39(1): 47-60.

Patel, B. C. and D. G. Sinha (2010). An adaptive K-means clustering algorithm for breast image segmentation. *International Journal of Computer Applications* 10(4): 35-38.

Paterlini, S. and T. Krink (2006). Differential evolution and particle swarm optimisation in partitional clustering. *Computational Statistics & Data Analysis* 50(5): 1220-1247.

Peuquet, D. J., A. C. Robinson, S. Stehle, F. A. Hardisty and W. Luo (2015). A method for discovery and analysis of temporal patterns in complex event data. *International Journal of Geographical Information Science* (ahead-of-print): 1-24.

Pham, D. T., S. S. Dimov and C. Nguyen (2005). Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 219(1): 103-119.

Phoungphol, P. and I. Srivrunyoo (2012). Boosting-genetic clustering algorithm. *Machine Learning and Cybernetics (ICMLC), 2012 International Conference on*, IEEE.

Premalatha, K. and A. Natarajan (2008). A new approach for data clustering based on PSO with local search. *Computer and Information Science* 1(4): p139.

Premalatha, K. and A. Natarajan (2010). Hybrid PSO and GA models for Document Clustering. *Int. J. Advance. Soft Comput. Appl* 2(3): 302-320.

Rakhlin, A. and A. Caponnetto (2006). Stability of $ K $-Means Clustering. *Advances in Neural Information Processing Systems*.

Ramamurthy, B. and K. Chandran (2011). CBMIR: shape-based image retrieval using canny edge detection and k-means clustering algorithms for medical images. *International Journal of Engineering Science and Technology* 3(3): 209-212.

Raposo, C., C. H. Antunes and J. P. Barreto (2014). Automatic Clustering Using a Genetic Algorithm with New Solution Encoding and Operators.*Computational Science and Its Applications–ICCSA 2014* 92-103, Springer.

Redmond, S. J. and C. Heneghan (2007). A method for initialising the< i> K</i>-means clustering algorithm using< i> kd</i>-trees. *Pattern Recognition Letters* 28(8): 965-973.

Renner, G. and A. Ekárt (2003). Genetic algorithms in computer aided design. *Computer-Aided Design* 35(8): 709-726.

Robinson, J. and Y. Rahmat-Samii (2004). Particle swarm optimization in electromagnetics. *Antennas and Propagation, IEEE Transactions on* 52(2): 397-407.

Roweis, S. T. and L. K. Saul (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500): 2323-2326.

Runkler, T. A. and C. Katz (2006). Fuzzy clustering by particle swarm optimization. *Fuzzy Systems, 2006 IEEE International Conference on*, IEEE.

Sadeghierad, M., A. Darabi, H. Lesani and H. Monsef (2010). Optimal design of the generator of microturbine using genetic algorithm and PSO. *International Journal of Electrical Power & Energy Systems* 32(7): 804-808.

Santhanam, T. and M. Padmavathi (2015). Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis. *Procedia Computer Science* 47: 76-83.

Senthil Kumar, A. and S. Mythili (2012). Parallel Implementation of Genetic Algorithm using K-Means Clustering. *International Journal of Advanced Networking & Applications* 3(6).

Sethi, C. and G. Mishra (2013). A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset. *International Journal of Scientific & Engineering Research* 4(6): 9.

Settles, B. (2010). Active learning literature survey. *University of Wisconsin, Madison* 52: 55-66.

Shakerian, R., S. Kamali, M. Hedayati and M. Alipour (2011). Comparative Study of Ant Colony Optimization and Particle Swarm Optimization for Grid Scheduling [J]. *The Journal of Mathematics and Computer Science* 2(3): 469-474.

Sharma, A. and C. W. Omlin (2009). Performance comparison of Particle Swarm Optimization with traditional clustering algorithms used in Self Organizing Map. *International Journal of Computational Intelligence* 5(1): 1-12.

Sheikh, R. H., M. Raghuwanshi and A. N. Jaiswal (2008). Genetic algorithm based clustering: a survey. *Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on*, IEEE.

Shen, H.-y., X.-q. Peng, J.-n. Wang and Z.-k. Hu (2005). A mountain clustering based on improved PSO algorithm.*Advances in Natural Computation* 477-481, Springer.

Shi, K. and L. Li (2013). High performance genetic algorithm based text clustering using parts of speech and outlier elimination. *Applied intelligence* 38(4): 511-519.

Shi, Y. and R. Eberhart (1998). A modified particle swarm optimizer. *Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence., The 1998 IEEE International Conference on*, IEEE.

Shirwaikar, R. and C. Bhandari (2013). K-means Clustering Method for the Analysis of Log Data. *Int. Conf. on Advances in Signal Processing and Communication*.

Shyr, W.-J. (2010). Parameters Determination for Optimum Design by Evolutionary Algorithm.

Simon, D. (2008). Biogeography-based optimization. *Evolutionary Computation, IEEE Transactions on* 12(6): 702-713.

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data*. John Wiley & Sons.

Singh, K., D. Malik and N. Sharma (2011). Evolving limitations in K-means algorithm in data mining and their removal. *International Journal of Computational Engineering & Management* 12: 105-109.

Sohn, S. (2007). *A Random Duplication/deletion Model in Genome Arrangement*. ProQuest.

Steinhaus, H. (1956). Sur la division des corp materiels en parties. *Bull. Acad. Polon. Sci* 1: 801-804.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology* 59(1): 1-34.

Sun, H.-j. and L.-h. Xiong (2009). Genetic algorithm-based high-dimensional data clustering technique. *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on*, IEEE.

Sun, J., W. Xu and B. Ye (2006). Quantum-behaved particle swarm optimization clustering algorithm.*Advanced Data Mining and Applications* 340-347, Springer.

Tsai, C.-Y. and I.-W. Kao (2011). Particle swarm optimization with selective particle regeneration for data clustering. *Expert Systems with Applications* 38(6): 6565-6576.

Tucker, A. B. (2004). *Computer science handbook*. CRC press.

Tuytelaars, T., C. H. Lampert, M. B. Blaschko and W. Buntine (2010). Unsupervised object discovery: A comparison. *International journal of computer vision* 88(2): 284-302.

Urade, H. S. and R. Patel (2012). Performance Evaluation of Dynamic Particle Swarm Optimization, IJCSN.

Utro, F. (2011). Algorithms for internal validation clustering measures in the Post Genomic Era. *arXiv preprint arXiv:1102.2915*.

Van den Bergh, F. and A. P. Engelbrecht (2004). A cooperative approach to particle swarm optimization. *Evolutionary Computation, IEEE Transactions on* 8(3): 225-239.

Van der Merwe, D. and A. P. Engelbrecht (2003). Data clustering using particle swarm optimization. *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, IEEE.

Vattani, A. (2011). K-means requires exponentially many iterations even in the plane. *Discrete & Computational Geometry* 45(4): 596-616.

Vincze, V. (2014). Uncertainty Detection in Natural Language Texts.

Wagstaff, K., C. Cardie, S. Rogers and S. Schrödl (2001). Constrained k-means clustering with background knowledge. *ICML*.

Wang, J. and X. Su (2011). An improved K-Means clustering algorithm. *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*, IEEE.

Wang, L., Y. Liu, X. Zhao and Y. Xu (2006). Particle swarm optimization for fuzzy c-means clustering. *Intelligent Control and Automation, 2006. WCICA 2006. The Sixth World Congress on*, IEEE.

Wang, Y., X. Ma, M. Xu, Y. Liu and Y. Wang (2015). Two-echelon logistics distribution region partitioning problem based on a hybrid particle swarm optimization–genetic algorithm. *Expert Systems with Applications* 42(12): 5019-5031.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition* 38(11): 1857-1874.

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing* 4(2): 65-85.

Wu, F.-X. (2008). Genetic weighted k-means algorithm for clustering large-scale gene expression data. *BMC bioinformatics* 9(Suppl 6): S12.

Xiang, Y. and A. T. L. Phuan (2005). Genetic algorithm based k-means fast learning artificial neural network.*AI 2004: Advances in Artificial Intelligence* 828-839, Springer.

Xu, R. and D. Wunsch (2005). Survey of clustering algorithms. *Neural Networks, IEEE Transactions on* 16(3): 645-678.

Yang, C.-S., L.-Y. Chuang and C.-H. Ke (2008). Comparative particle swarm optimization (CPSO) for solving optimization problems. *Research, Innovation and Vision for the Future, 2008. RIVF 2008. IEEE International Conference on*, IEEE.

Yazdani, D., S. Golyari and M. R. Meybodi (2010). A new hybrid approach for data clustering. *Telecommunications (IST), 2010 5th International Symposium on*, IEEE.

Yedla, M., S. R. Pathakota and T. Srinivasa (2010). Enhancing K-means clustering algorithm with improved initial center. *International Journal of computer science and information technologies* 1(2): 121-125.

Yeung, K. Y., M. Medvedovic and R. E. Bumgarner (2003). Clustering gene-expression data with repeated measurements. *Genome Biol* 4(5): R34.

Zha, H., X. He, C. Ding, M. Gu and H. D. Simon (2001). Spectral relaxation for k-means clustering. *Advances in neural information processing systems*.

Zhang, B., M. Hsu and U. Dayal (1999). K-harmonic means-a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*.

Zhang, C. and S. Xia (2009). K-means clustering algorithm with improved initial center. *Knowledge Discovery and Data Mining, 2009. WKDD 2009. Second International Workshop on*, IEEE.

Zhang, Z., J. Zhang and H. Xue (2008). Improved K-means clustering algorithm. *Image and Signal Processing, 2008. CISP'08. Congress on*, IEEE.

Zhou, H., Y. C. Soh and X. Wu (2015). Integrated analysis of CFD data with K-means clustering algorithm and extreme learning machine for localized HVAC control. *Applied Thermal Engineering* 76: 98-104.

Zhu, X. (2005). Semi-supervised learning literature survey.

Zhu, X. (2015). Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education.

Zhu, X. and A. B. Goldberg (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* 3(1): 1-130.