Mathematical optimization or mathematical programming is the selection of the best element from some set of available alternatives. Optimization problems of sorts arise in all quantitative disciplines from computer science and engineering to operations research and economics, and the development of solution methods has been of interest in mathematics for centuries. In the simplest case, an optimization problem consists of maximizing or minimizing a real function by systematically choosing input values from within an allowed set and computing the value of the function. The generalization of optimization theory and techniques to other formulations constitutes a large area of applied mathematics. More generally, optimization includes finding "best available" values of some objective function given a defined domain (or input), including a variety of different types of objective functions and different types of domains.

Dr. Gaurav Dhiman received his Ph.D. in Computer Engineering from Thapar Institute of Engineering & Technology, Patiala. He is currently working as an Assistant Professor with the Department of Computer Science, Government Bikram College of Commerce, Patiala. He has published more than 25 peer-review research papers (indexed in SCI-SCIE).
Gaurav Dhiman
Jaswinder Singh
Amandeep Kaur

Hybrid Metaheuristic Optimization Approach: Theory and its Application
Hybrid Metaheuristic Optimization Approach: Theory and its Application

Gaurav Dhiman
Jaswinder Singh
Amandeep Kaur

LAP LAMBERT Academic Publishing
Imprint
Any brand names and product names mentioned in this book are subject to trademark, brand or patent protection and are trademarks or registered trademarks of their respective holders. The use of brand names, product names, common names, trade names, product descriptions etc. even without a particular marking in this work is in no way to be construed to mean that such names may be regarded as unrestricted in respect of trademark and brand protection legislation and could thus be used by anyone.

Cover image: www.ingimage.com

Publisher:
LAP LAMBERT Academic Publishing
is a trademark of
International Book Market Service Ltd., member of OmniScriptum Publishing Group
17 Meldrum Street, Beau Bassin 71504, Mauritius

Printed at: see last page

Zugl. / Approved by: Desh Bhagat University

Copyright © Gaurav Dhiman, Jaswinder Singh, Amandeep Kaur
Copyright © 2019 International Book Market Service Ltd., member of OmniScriptum Publishing Group
Hybrid Metaheuristic Optimization Approach: Theory and its Application

Authored by:

Gaurav Dhiman, Jaswinder Singh, and Amandeep Kaur
Dedication

“Dedicated to our Father and Mother”
ABSTRACT

Clustering is the process of grouping data into clusters where objects within each cluster are highly similar but differ from objects in other clusters. Similarities are evaluated based on the value(s) of the attribute that best describes the object. For the purpose, distance measurements are often used. Clustering has its roots in many fields, including data mining, statistics, biology, and machine learning. This thesis focuses on the clustering concept, i.e. automatically determining the number of features and the number of clusters. Clustering method is a vital technique in data mining and is implemented using various algorithms such as hierarchical algorithms, partitioning, density based algorithms and grid based algorithms. There is an abundance of subtypes and distinctive algorithms to find the clusters within each of the sorts. The "Hierarchical Clustering" is a grouping strategy in which a group command chain is divided into the comparative dataset. Data items are isolated into non-covering groups in "partitioning clustering" so that each object is inaccurately in one sub-set. The reason for isolating the items into a few sub-sets is because it is not computationally feasible to check all conceivable sub-set frameworks; there are certain covetous heuristics plans that are used as part of the iterative enhancement type.

Clustering algorithms based on density attempt to discover groups in a cluster based on the thickness of information focus.

Grid-based clustering is used to isolate the information space into a limited number of cells that shape the network structure and align (cluster formation) on the grids.

Clustering based on Hierarchy and partitioning is most commonly used in various applications. Clustering based on partitioning makes the groups, as opposed to making a few stages, in one stage. One and only group arrangement is shaped towards the end of grouping, although some bunches arrangements may be made within. The partitional clustering is usually faster than the hierarchical clustering.

Different hierarchical clusters require only a measure of comparability, while partitional grouping requires more grounded suspicions, such as number of groups and introductory focus. Hierarchical clustering is more appropriate for unmitigated information where it is possible to characterize the length of a similarity measure accordingly. A partitional clustering splits information set into a disjoint bunches arrangement. in the related to this partitional clustering algorithm like-Mean Algorithm. K-mean introduce by Mac Queen, 1967 [36]. K-Means clustering aims to divide n objects into k clusters in which each object belongs to the nearest mean cluster.
This method produces different clusters of difference exactly k. The best number of clusters k that lead to the greatest distance is not known as a priori and must be calculated from the data. The aim of the K-means that agglomeration is to cut back the overall intra-cluster variance or the sq. error function [37, 38]. This book focuses on the clustering concept, i.e. the number of features and number of clusters will be determined automatically. Clustering technique is treated as a clustering problem due to unknown number of clusters. Meta-heuristic algorithms are used to solve this problem. Hybrid meta-heuristic optimization algorithms have been proposed in this thesis, namely hybrid optimization algorithms to solve real-life design problems in engineering. On standard benchmark test functions, the proposed approaches are evaluated. In order to ensure the applicability of proposed algorithms, convergence and computational complexity were also analyzed. The performance of the proposed algorithms is analyzed and compared with various algorithms in such a way that the concept of determining the better solution and convergence perform s better than the others and produces high convergence.

This thesis also presents the problems associated with the k-means algorithm, which has been successfully applied to many practical clustering problems, has several disadvantages due to its initialization selection. However, its performance depends on the initial centroid state and can be trapped in the local optima. In this thesis, we propose a hybrid method. A hybrid technique based on the combination of the K-means algorithm, the genetic algorithm, the Nelder–Mead simplex search and the K–GA-NM–PSO particle swarm optimization is proposed. The KM-GA–NM–PSO searches for cluster centres of an arbitrary data set as well as the K-means algorithm, but the global optima can be found effectively and efficiently.

“The new technique K–GA–NM–PSO algorithm is tested on data sets, and its performance is compared with those of k-mean, GA, PSO, K-mean-GA-NM-PSO and K-means clustering. Results show that K-mean–GA–NM-PSO is better than other cluster.”

This approach finds both the characteristics and the number of clusters from the given data set simultaneously. It is suggested that a variable length agent encode both cluster centres and features of various cluster numbers. The performance of the proposed technique is tested on different applications in real-life and compared to existing algorithms. Experimental results show that the approach proposed is superior to the other algorithms of the competitor.

**Keywords:** Optimization, Meta-heuristic, Hybrid Optimization Algorithm Clustering, Feature selection, Benchmark test functions, Convergence, Diversity.
Contents:

Chapter 1 .......................................................... 2
Chapter 2 .......................................................... 12
Chapter 3 .......................................................... 15
Chapter 4 .......................................................... 20
References: ......................................................... 21
Chapter 1: INTRODUCTION

In recent years, data analysis of real-life applications has posed some critical challenges due to advances in internet technology and multimedia tools. These challenges include extracting information from a wide data set, analyzing data from various dimensions, categorizing data, and summing up the data relationship. In order to address these challenges, data mining is a preferable technique. Generally speaking, data mining techniques are classified into two categories, such as direct and indirect data mining. Figure 1 shows the classification of data mining techniques.

Figure 1: Classification of Data Mining Techniques.
Direct data mining techniques are classified in three categories, including classification, estimation and prediction. Classification is a process to divide the set of data into different classes based on the knowledge of predefined classes or the structure of the set of data. It is also referred to as supervised learning. The number of classes in a given set of data is known in advance in supervised learning and assign, the class label of non-classified cases in a given set of data. Indirect data mining techniques are divided into three categories, such as clustering, mining and visualization. Clustering is widely used in indirect data mining. Clustering is a group of objects in a cluster that are very similar to each other or different from other clusters. The discovery of the set of relevant objects without prior knowledge is an unmonitored learning approach. It has been used in many fields of engineering, such as bioinformatics, data collection, and forecasting and image segmentation [5]. Clustering techniques are generally classified in two classes, such as classical and metaheuristic techniques [5]. Classical clustering algorithms are classified into five main classes, such as hierarchical clustering, grid-based clustering, clustering based on density, partitional clustering and clustering based on models. Due to its efficiency and simplicity, K-means is a widely used classic partitional clustering algorithm. However, premature convergence is suffering [6]. Meta-heuristic techniques (see Figure 2) are used in clustering to find the best global partition to overcome this problem. Clustering techniques based on meta-heuristics are divided into two categories. These are clustering techniques based on single and multi-objective.

The main focus of this research is single-objective hybrid method clustering technique based on partitions. It is used to address two major clustering problems, such as determining the number of clusters before the clustering process and finding the best partitions for the cluster. To solve these problems, the partitional single-objective clustering technique is repeatedly used with different number of classes as input and then the partitioning of the data leading to the best cluster validity indices is selected. The validity index of a cluster cannot detect the correct number of clusters. Therefore, a new fitness function must be built that includes more than one cluster validity indices to maintain the inter-cluster and intra-cluster properties. An efficient feature selection technique is required to improve the efficiency of the algorithm and minimize redundancy for uncontrolled data sets. The selection of features in unattended learning is a difficult problem because class labels cannot be used as advisors to search for relevant information. For efficient clustering a new single-objective hybrid approach meta-heuristic technique is therefore required.

1.1 Single-objective:

Minimize/Maximize: $F(\sim z) = f_1(\sim z)$  

(1.21)
Subject to:

\[ g_j(z) \geq 0, \quad j = 1, 2, \ldots, p \] (1.22)

\[ h_j(z) = 0, \quad j = 1, 2, \ldots, q \] (1.23)

\[ Lb_j \leq z_j \leq ub_j, \quad j = 1, 2, \ldots, r \] (1.24)

Many real-life issues need to achieve multiple goals such as minimizing risks, minimizing costs, maximizing reliability, etc.[40]. Single-objective optimization is aimed primarily at finding the best optimal solution and addressing only one goal to be minimized or maximized. The mathematical formulation of single-target optimization is as follows:

Where \( p \) is the number of inequality constraints, \( g_j \) is \( j^{th} \) inequality constraints, \( q \) is the number of equality constraints, \( h_j \) is \( j^{th} \) equality constraints, \( r \) is the number of variables, \( Lb_j \) and \( ub_j \) are lower and upper bounds of \( j^{th} \) variable, respectively.

### 1.2 Multi-objective:

Multi-objective optimization refers to optimizing the function of a given problem with more than one objective (criterion). It can be described as \([41, 42]\):

Minimize: \( F(z) = [f_1(z), f_2(z), \ldots, f_n(z)] \) (1.25)

Subject to:

\[ g_i(z) \geq 0, \quad i = 1, 2, \ldots, m \] (1.26)

\[ h_i(z) = 0, \quad i = 1, 2, \ldots, p \] (1.27)

Where \( z = [z_1, z_2, \ldots, z_d]^T \) is the vector of decision variables, \( m \) is the number of inequality constraints, \( p \) is the number of equality constraints, \( g_i \) is \( i^{th} \) inequality constraints, \( h_i \) is \( i^{th} \) equality constraints, and \( obj \) is the number of objective functions \( f_i: \mathbb{R}^{obj} \rightarrow \mathbb{R} \). Because of multi-criterion comparison metrics [34], the solutions in a search space cannot be compared by relational operators. Edgeworth [44] first proposed the comparison of two solutions and Pareto [45] extended it further. Pareto dominance's mathematical formulation is described as [44].
1.3 Tabu search algorithm:

Glover formalized Tabu Search (TS) in 1986[1]. To manage an embedded local search algorithm, TS has been designed. It uses the search history explicitly, both to escape the local minima and to implement an exploratory strategy. Indeed, its main feature is based on the use of human memory-inspired mechanisms. From this point of view, it takes a path contrary to that of SA that does not use memory and is therefore unable to learn from the past. Different types of memory structures are commonly used through the search space the algorithm has undertaken to remember specific trajectory properties. A Tabu list (from which the name of the meta-heuristic framework derives) records the last encountered solutions (or some of their attributes) and prohibits the re-visitation of these solutions (or solutions containing one of these attributes) as long as they are in the list. This list can be viewed as a short-term memory, recording information about solutions recently visited. Its use prevents the return to the solutions recently visited, thus preventing endless cycling and forcing the search to accept even deteriorating movements. The Tabu list's length controls the search process's memory. The search will focus on small areas of the search space if the length of the list is low. On the contrary, a high length forces the search process to explore larger regions, as it prevents a higher number of solutions from being revisited. During the search, this length may vary, resulting in more robust algorithms, such as the Reactive Tabu Search algorithm [2].

Step 1: Choose an initial solution \( i \) in \( S \). Set \( i^* = i \) and \( k = 0 \).

Step 2: Set \( k = k + 1 \) and generate a subset \( V^* \) of solution in \( N(i, k) \) such that neither one of the Tabu conditions is violated or at least one of the aspiration conditions holds.

Step 3: Choose a best \( j \) in \( V^* \) and set \( i = j \).

Step 4: If \( f(i) < f(i^*) \) then set \( i^* = i \).

Step 5: Update Tabu and aspiration conditions.

Step 6: If a stopping condition is met then stop. Else go to Step 2.

Some immediate stopping conditions could be the following [2]:

- \( N(i, K+1) = 0 \). (No feasible solution in the neighborhood of solution \( i \))
- \( K \) is larger than the maximum number of iterations allowed.
- The number of iterations since the last improvement of \( i^* \) is larger than a specified number.

Evidence can be given that an optimum solution has been obtained.

1.4 Evolutionary Computation algorithm

Evolutionary Computation (EC) is the general term for several optimization algorithms inspired by the Darwinian principles of the ability of nature to evolve well-adapted living beings to their environment. The fields of genetic algorithms [3], evolutionary strategies [4], evolutionary programming [5] and genetic programming [6] are usually found grouped under the term EC algorithms (also called Evolutionary Algorithms (EAs)). Despite the differences
that will be shown later between these techniques, they all share a common underlying idea of simulating the evolution of individual structures through selection processes, Recombination and reproduction of mutations, thus creating better solutions. Each iteration of the algorithm corresponds to a generation in which a population of candidate solutions to a given problem of optimization, called individuals, is capable of reproduction and is subject to genetic variations followed by the environmental pressure which causes natural selection (survival of the most fitting). New solutions are created by applying recombination, which combines two or more selected individuals (so-called parents) to produce one or more new individuals (children or offspring) and mutation, which allows the appearance of new characteristics in the offspring to promote diversity. The fitness of the resulting solutions (how good the solutions are) is evaluated and then an appropriate selection strategy is applied to determine which solutions will be maintained in the next generation. A predefined number of generations (or function assessments) of simulated evolutionary processes are usually used as a termination condition, or some more complex stop criteria can be applied.

**Evolutionary Computation algorithm:**

1. *Initialize* the population with random individuals;
2. *Evaluate* each individual:
3. Repeat
4. *Select* parents;
5. *Recombine* pairs of parents;
6. *Mutate* the resulting offspring;
7. *Evaluate* new individuals;
8. *Select* individuals for the next generation;

Until a termination condition is satisfied;

**1.5 Swarm Intelligence algorithm**

Swarm Intelligence (SI) is an innovative distributed intelligent paradigm for solving optimization issues inspired by a group of social insect colonies and other animal societies ' collective behaviour. Typically, SI systems consist of a population of simple agents (an entity capable of performing / performing certain operations) that interact locally with each other and their environment. These entities with very limited individual capacity can carry out many complex tasks that are necessary for their survival together (cooperatively). While there is normally no centralized control structure dictating how individual agents should behave, local interactions among such agents often lead to global and self-organized behaviour emerging. Several algorithms of optimization inspired by swarming behaviour metaphors are proposed in nature. Examples to this effect are ant colony optimization, particle swarm optimization, bacterial foraging optimization, bee colony optimization, anti-facial immune

1.6 Ant Colony Optimization algorithm:
M. introduced ant colony optimization (ACO). Dorigo and colleagues [11, 12, 13] as a metaheuristic inspired by nature to solve difficult problems of combinatorial optimization. ACO is inspired by real ants' foraging behaviour. These ants initially explore the area around their nest by performing a randomized walk when searching for food. Ants deposit a chemical pheromone trail on the ground along their path between food source and nest to mark a favourable path that should guide other ants to the food source [14]. After some time, there is a higher concentration of pheromone in the shortest path between the nest and the food source and therefore it attracts more ants. Artificial ant colonies used this characteristic of real ant colonies to create solutions to an optimization problem and exchange quality information through a communication scheme that is reminiscent of that adopted by real ants [15].

Algorithm ACO:
1. Initialize
2. while termination condition not met do
3. Construct Ants Solutions;
4. Update Pheromones;
5. Daemon Actions;
6. End.

1.7 Bee colony optimization-based algorithms
Bee colony-based optimization algorithms are a new type of algorithm inspired by honeybee colony behaviour that exhibits many features that can be used as models for smart systems and collective behaviour. These characteristics include waggle dance (communication), food foraging, queen bee, task selection, collective decision making, selection of nest sites, bee colony mating and marriage, floral / pheromone laying and navigation systems[16]. For a specific task, each model defines a particular behaviour. The bee hive colonies structure of Honeybee contains a single queen mated to a large number of males (drones) and thousands of workers. The queen is the only female egg-laying in a bee hive, secreting a pheromone that
keeps all other females sterile in the colony. A fertile queen can lay fertilized or unfertilized eggs selectively. In workers or virgin queens, fertilized eggs hatch, while unfertilized eggs produce drones. Individual worker bees are always female because male drones, apart from matching with queens during marriage flights, do not contribute to social life. Workers perform various tasks as nurses tending, nest-building, hive defence, and as foragers by collecting nectar and pollen to make honey and feed the brood. A handle of algorithms such as Queen-bee Evolution Algorithm (QBE) [17] and Queen Bee-based crossover operator for GA [18] were developed in the literature on the basis of the Queen Bee concept. Bee Dance and Communication Bees exchange information about where food sources are located by performing a series of movements, often called waggle dance. Each hive has a so-called dance floor area where the bees who have discovered nectar sources dance to promote food locations and persuade their estimates to follow them. When a bee decides to leave the hive to collect nectar, one of the bee dancers follows to one of the nectar areas. The communicative procedures of honey bees such as Bee hive algorithm [19] and Discrete Bee Dance Algorithm [20] have inspired some bee colony-based algorithms.

**BCO Algorithm:**

Initialization: Read problem data, parameter values (B and NC), and stopping criterion.

Do

1. Assign a (n) (empty) solution to each bee.
2. For (i = 0; i < NC; i++)
   //forward pass
   For (b = 0; b < B; b++)
   For (s = 0; s < f(NC); s++)
   //count moves
   (i) Evaluate possible moves;
   (ii) Choose one move using the roulette wheel;
   //backward pass
   (b) For (b = 0; b < B; b++)
      Evaluate the (partial/complete) solution of bee b;
   (c) For (b = 0; b < B; b++)
      Loyalty decision for bee b;
   (d) For (b = 0; b < B; b++)
      If (b is uncommitted), choose a recruiter by the roulette wheel
3. Evaluate all solutions and find the best one. Update xbest and f(xbest)

While stopping criterion is not satisfied.

Return (xbest, f(xbest))
1.8 GENETIC ALGORITHM (GA)

This method is a search method which is based on the principle of natural selection and genetic [21, 22]. GA concept was formalized for the first time by Holland [22]. It emulates the natural selection and the evolution mechanism of Darwin. It has been found and proven to be the most efficient, effective and powerful global optimization algorithm which in general forms combinational optimization problems while in particular the problems having discrete optimization parameters. There is no discontinuous or differentiable object function. The main and basic building blocks of the binary GA are chromosomes and genes. The optimization parameters are encoded by the conventional binary into binary code string. [23] To evolve and develop better solutions and ways perform the selection which is natural, a certain parameter is required to discriminate better from worse solutions. Concerning GA, the measure can function objectively and can be a computer simulation model based on mathematics or it can even be functioning subjectively where human may select finer and better ways and solutions other than ones which are worse. Essentially, the measures for the fitness ought to find a fitness which is relative of a candidate solution, which GA will thus use to direct rise and emergence of better solutions [24]. There is one another essential of GA which is population belief. In contrast to the traditional search methods, GA depends on a population of candidate solutions. The population size, which is a parameter, determined by the user, is one of the important elements which influence the GA performance and the scalability. For instance, a small-sized population might result in an immature convergence and offers solutions which are below standard. But huge sized population might result in wasting valuable time in the computing process [25]. After encoding the problem manner of chromosomes and after choosing a parameter of fitness which differentiate the good and bad solutions, the GA becomes ready to find the best solution by means of using the steps bellow [26]:

**Genetic Algorithm:**

1. **Initialization:** the primary or starting solutions of candidates for population is mostly produced through the search space in random way.

2. **Evaluation:** Just when new population is generated or population is initialized, evaluation of the fitness values of the candidate solutions are carried out.

3. **Selection:** more copies solutions are allocated by means of selection with high level fitness and the idea of survival possibilities of being fittest is imposed on the solutions of the candidates. The main notion choice is to choose the solutions which are better and favouring them to other ones. Therefore, many procedures of selection were suggested to accomplish this notion. Amongst these procedures: the roulette-wheel selection, the stochastic universal selection, the ranking selection and tournament selection.
4. **Recombination**: Combining parts of two or more of the main solutions to have new and better solutions (i.e. offspring). There are several ways to achieve this and the efficient performance relies on the recombination mechanism which should be properly designed.

5. **Mutation**: At the time when two or more parental chromosomes are operated by recombining, it results in modification of mutation, a local and random way to solution. Various distinctive types of mutation, but mutation commonly include one change or more that occur in the individual feature(s). In other words, mutation performs a random walk in the candidate solution vicinity.

6. **Replacement**: The offspring population which is caused by selection, recombination, and mutation replaces the original parental population. Many replacements methods like the Steady state replacement, elitist replacement, generation wise replacement are utilized in GA.

7. Repeating all the steps from 2 till 6 till reach termination condition.

### 1.9 Nelder-Mead Algorithm:

This simple search method, first developed by Spendley, Hext and Himsworth (1962) [27] and subsequently refined by Nelder and Mead (1965)[28], is a derivative-free line search method used to find the minimum or maximum objective function. See, for example, Olsson and Nelson (1975) [29]. The fitness function value at (N+1) of the initial simplex is evaluated. In this, the function's value is high and new, and then replaced by a good point. Which can be located in a negative gradient form (direction)? These are considered a direct line search technique as one of the best resources. Four basic practices in this process Processing per-soil is applicable. Reflecting, diversifying, storing and reducing these local surface points can be more intensive and the general can make great progress in itself. Thus, in the example below, the function of two variables is minimized (N=2) the basic NM procedure is shown.

1. Sort the A, B, and C function values. Assume if(C) <f (B) <f (A) is the highest of the three function values and must be replaced. In this case, a reflection is made in point D to point E through the centre of BC.

2. If f (E) < f(C) is expanded to point J. Then we replace E or J with r for A, depending on which function value is lower.

3. If f(E)>f(C), there is a contraction to point G or H as a substitute for A, depending on which of f(A) and f(E) is lower, provided that f(G) or f(H) is lower than f(C). If either f(G) or f(H) is greater than f(C), the contraction failed and a shrinkage operation is carried out. The shrinkage procedure reduces the size of the simplex by moving everything but the best point C halfway to the best point C. We've got new points A and B. Return to step 1.
1.10 PSO algorithm:

PSO stands for particle swarms optimization (PSO) it is most popular evolutionary optimization techniques developed by Kennedy and Eberhart (1995)[30,31] in this algorithm population based and evolutionary in nature. It is inspired by the collective behaviour of birds flying around in the sky - those who are engaged in search of their food and are same as fish schooling [32]. This search space is applied to a fitness function to reach good results. The particles swarm through the fitness function solved to search space to find the maximum value return by the objective function. That is a used a number of particles constitute a swarm moving award in search space locking for the best solution. Each particle in search space adjusts its “swarm” according to own swarm experience as well as the swarm experience of the other particle. PSO is same as a genetic algorithm, but the main difference is that they cannot apply filtering. This means that all the members of the population Survive through the entire search process.

The following steps of the PSO algorithm:

1. Initialization process. Randomly generate 5N potential solutions called" particles', 'N being the number of parameters to be optimized and a randomized velocity is assigned to each particle.

2. Velocity Update the particles then' fly' through hyperspace while updating their own velocity, which is achieved by taking into account their own past flight and that of their companions.' The velocity and position of the particle is dynamically updated with the following equations[33]:

\[
V_{id}^{\text{New}} = (W \cdot V_{id}^{\text{old}}) + c_1 \cdot r_1 \cdot (p_{id} - x_{id}^{\text{old}}) + c_2 \cdot r_2 \cdot (p_{gd} - x_{id}^{\text{old}}) \tag{4}
\]

\[
X_{id}^{\text{New}} = x_{id}^{\text{old}} + V_{id}^{\text{New}} \tag{5}
\]

Where \(c_1\) and \(c_2\) are two positive constants, \(w\) is an inertia weight, and \(r_1\) and \(r_2\) are random number generated [34, 35].
Chapter 2: LITERATURE REVIEW

This chapter provides a review of the major clustering issues. From three different points of view the literature is reviewed. The first section deals with the review for data clustering of multi-objective Meta-heuristic techniques. A brief overview of the research work based on the techniques of multi-objective selection of features is then covered. The third section uses multi-objective Meta-heuristic techniques to describe the brief review of data clustering and feature selection.

Handl and Knowles [57] proposed an Automatic K-determination (MOCK) multi-objective clustering approach. It consists of two main phases, the initial clustering and the selection of models. Using multi-objective evolutionary algorithm (MOEA), it optimizes the two complementary clustering goals in the initial phase. Liu et al. [58] developed an automatic Genetic algorithm based clustering approach, called the Multi-objective Genetic K means Algorithm (MOKGA), which minimizes two objective functions such as cluster number and Partitioning error mistake et al. [59] demonstrated that when the technique is applied to large data sets, MOCK's computational cost is too high. They developed a web mining Data clustering algorithm based on a scalable automatic k –mean determination scheme.

Bandyopadhyay et al. [60] proposed a clustering technique based on Non-dominated Genetic Sorting Algorithm 2 (NSGA II) that simultaneously optimized two measurements of cluster validity. A real centroid-based encoding has been used.

In a multi-objective optimization framework, Suresh et al. [61] applied the Differential Evolution (DE) algorithm on automatic clustering. The study optimized both conflicting fuzzy validity indices simultaneously. A Real-coded search variables representation was used. In selecting the most interesting solutions from the Pareto front, the gap statistics were applied.

Kundu et al. [62] proposed a GADE based Genetic Algorithm (GA) and Differential Evolution (DE) hybrid clustering approach. In the multi-objective framework, GADE incorporated GA operators. The Xie - Beni index and Fuzzy C-means (FCM) measurement were used to optimize objective functions.
Paoli et al. [63] presented a clustering methodology based on multi-objective swarm optimization (MOPSO). This approach was proposed to resolve three different issues related to the segmentation of hyper spectral images: (i) estimation of class statistical parameters, (ii) detection of the best discriminative bands without requiring their number a priori, and (iii) estimation of the number of data classes (or number of clusters) characterizing the image being considered. MOPSO is guided by three criteria for optimization, namely the log-likelihood function (LF), the Bhattacharyya statistical distance (BD) between classes and the minimum length of description (MDL).

Ma et al. [64] developed the multi-objective clustering algorithm for immune dominance (IDCMC) based on immune-dominant clonally selection. Based on three different measurements, this approach divided the single population into three subpopulations. Efficiency and non-linear time complexity were the main advantages of this approach.

Saha and Bandyopadhyay [65] proposed stability induced a multi-objective optimization technique (SSym-AMOSA) based on sym-index and simulated annealing. It simultaneously optimized two objective functions, one reflecting the total symmetry of the cluster and the other reflecting the stability of the partitions obtained over various data set bootstrap samples.

The automatic multiobjective clustering technique with point based symmetry distance (VA MOSA) [66] defines both XB and Sym indices as objective functions for determining the appropriate number of clusters.

Saha et al. [67] proposed another fuzzy, multiobjective differential evolution (MODE) clustering technique. This technique, called Fuzzy Clustering (MOMoDEFC) based on multi-objective modified differential evolution, encodes cluster centers and optimizes XB index and FCM measurement.

Liu et al. [68] proposed a multi-objective Invasive Weed Optimization (MOIWO) approach to resolving the clustering problem where the number of clusters was uncertain. MOIWO optimized both the XB index and the FCM measure at the same time. It adopted a real-coded scheme of variable length in which strings encode centroids in clusters.

Similarly, Nanda and Panda [69] proposed an automatic multi objective clustering algorithm Called MOCLONAL. It used the cloning, hyper-mutation, and immune cell re-selection Operators in relevant dimensions, data objects belonging to the same cluster were similar.
NSGA-II was adopted as the basis for the soft subspace clustering (MOEASSC) multi-objective evolutionary approach proposed by Xia et al. [70]. The validity similarity index of the projection (P SV Index) has been used.

Zhong and Zhang [71] have proposed an automatic fuzzy clustering method based on adaptive multi-objective differential evolution (AFCMDE) to optimize XBindex and FCM measurement using a two-layer fuzzy clustering system.

A hybrid multi-objective algorithm based on multi-objective immunized particle swarm optimization (MOIMPSO) was introduced by Nanda and Panda [72]. This approach combined operators from the principle of clonally selection (CSP) with the optimization of particle swarm (PSO). It evaluated the cluster centers automatically and optimized both objective functions simultaneously to evaluate the effectiveness of multi-objective domain solutions.

Abubaker et al. [73] presented a multi-objective hybrid clustering algorithm based on Multi-objective Particle Swarm Optimization and MOPSOSA. The features of particle swarm optimization (PSO) and simulated annealing (SA) were combined in this algorithm. At the same time, MOPSOSA optimized three objective functions such as I index DB, (ii) index Sym.

In a multi-objective framework (Genetic Cluster MOO), Saha and Bandyopadhyay [74] developed a generalized automatic clustering algorithm. It optimized the three objective functions like I total partitioning compactness based on Euclidean distance, (ii) total cluster symmetry based on point symmetry distance, and (iii) cluster connectivity.
Chapter 3: PROPOSED HYBRID APPROACH

An improvement over the algorithm is a hybrid technique based on combining the K-means algorithm with various other algorithms. The combined approach of different algorithms therefore provides better performance using the goodness of the whole algorithm to overcome the disadvantage of any particular algorithm. Genetic algorithm is one of the most commonly used evolutionary algorithm techniques to solve a clustering problem. Therefore, a hybrid data clustering algorithm based on GA and k-means (GA-KM), which uses the advantages of both algorithms. The GA-KM algorithm helps the k-means algorithm to escape local optimum. GA has been shown to be able to determine the best cluster initialization and to optimize initial parameters. GA defines a randomly generated population of people. These people are involved in the generation of new and better offspring by mutation / crossover. Decision on better offspring / individuals is achieved by fitness. The greatest benefit of genetic algorithms is that the fitness function can be changed to change the algorithm's behaviour. There is a wide variety of representations of individual or chromosomes. The solutions are traditionally represented using fixed length strings, in particular binary strings, but alternative encoding has been developed. The main focus of the GA-based algorithm was to generate high-quality clusters in optimized time. The focus of the current research was to use GA as an initial centroid selection tool and to study the performance of improved clustering of k-means. The applications of GA-based k means have been tested in literature on standard data sets, but educational data set specifically from the problem of school children has not been investigated. Current research has focused on developing an appropriate system to study school children's problems using basic k-means and improved k-means (GA with k-means). Consequently, the approach to the development of a new algorithm was problematic and the selection criteria or initial centroid influenced the nature of the domain. In short, according to the problem area, the fitness function in GA has been defined. Apart from identifying preferable technique for out of school children problem, there is always a need to analyze quality of clusters. There will be good method to measure the quality of the better clusters and performance of clustering. The hybrid (KM-GA-NM-PSO) Algorithms contain all the best features of the existing algorithm that overcome the limitations of the individual algorithm when combined. The improvement of this combined approach will lead to even better results. This will be requiring a minimum number of evaluations of functions to achieve the optimum solution. Compared to other methods, the hybrid approach will be produces high-quality clusters with small standard deviations on
selected data sets. It is proposed to combine with KM, GA, NM, PSO algorithms. This combination of hybrids improves the quality of data clustering and improves the algorithm.

3.1. Experimental results

Step 1: K-mean method applies randomly choose k centroids from dataset for desired clusters assign to each data object to the cluster with the closet centroids. Update the centroids by calculating the mean value of object within clusters. Repeat step 1.2 and 1.3 until termination centroids are met.

Step 2: Generate initial population of size i(\(\{j_1, j_2, j_3, \ldots, j_i\}\)).

- J1= k-mean (dataset)
- J2= min (dataset)
- J3= mean (dataset)
- J4= max (dataset)
- J5= Ji= random value of (dataset)

Step 3: GA algorithm apply

Apply crossover operator on N particle (GA).
Apply mutation operator on update N particle (GA).

Step 4: NM simplex method apply

Initialization: Generate a population of size 3N+1.
Evaluation and Ranking: Evaluate the fitness of each particle rank them on the basis of fitness.
Apply NM operator to the top N+1 particle and replace the (N+1) particle with the update.

Step PSO algorithm apply

Apply PSO operator for updating the remaining 2N particles.
Selection: from the population select the global best particle and the neighbourhood best particles. Velocity Update: apply update to the 2N particle with worst fitness according equations (3) & (4);

Step 5: If the termination conditions are not meet then go to back 4.2.

3.2. Experimental result

Iris Data Set We used the Iris data set to bring our algorithms a pragmatic result. In this case, each data set in the Iris Data Set has the number of their own distributions that these items of clusters and data are important to. Iris is used to set up a good comparison and algorithm for data sets. In this data set (\(n=150, d=4, k=3\)) it has three equal squares of 50 squares. In this data set we have 150 samples. It covers each class type of a class iris Flowers, in which four-
digit properties are also included. These data sets are such that the length of the sepal in cm, width and height of the petals Widths are in cent-meters. There is no missing value in this data set.

### 3.3. Performance measure

The Iris data set has been used in separate different algorithms, a predominantly KM algorithm, GA, NM,PSO Algorithm and K-GA-NM-PSO Algorithm have been developed in a table. In which good results have been found and the individual's best performance has been received. That compares to other clustering algorithms. K-mean algorithm in some cases there are problems. Just as in the beginning, there may be a set of solutions for the K-GA matching solution to the problem of a satellite base and its solutions. So, we are using the PSO algorithm. With the help of algorithms, it helps to maintain the integrity of all algorithms and simultaneously solve their problems. This is how the NM algorithm has been defeated again. NM algorithm helps us to provide a lot of efficient local research process from algorithms. But the NM algorithm is dependent on the starting point and this convergence is sensitive to choose the randomly the starting point and this can also be an algorithms increase percentage in algorithm. Tables 1-2 and Figure 2 show the efficiency comparison.

<table>
<thead>
<tr>
<th>K Value</th>
<th>k-mean</th>
<th>GA</th>
<th>NM</th>
<th>PSO</th>
<th>k-mean+ GA+NM+PSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=1</td>
<td>68.6166</td>
<td>66.0783</td>
<td>60.0123</td>
<td>70.2568</td>
<td>35.1443</td>
</tr>
<tr>
<td>K=2</td>
<td>82.6219</td>
<td>68.635</td>
<td>59.3256</td>
<td>94.2564</td>
<td>50.6701</td>
</tr>
<tr>
<td>K=3</td>
<td>129.5325</td>
<td>92.3585</td>
<td>91.6584</td>
<td>135.2567</td>
<td>20.3042</td>
</tr>
<tr>
<td>K=4</td>
<td>203.5256</td>
<td>150.1065</td>
<td>149.2569</td>
<td>278.2547</td>
<td>48.0000</td>
</tr>
<tr>
<td>K=5</td>
<td>355.2576</td>
<td>121.4141</td>
<td>360.5698</td>
<td>396.4567</td>
<td>91.2340</td>
</tr>
<tr>
<td>K=6</td>
<td>328.016</td>
<td>198.7141</td>
<td>365.1245</td>
<td>421.2584</td>
<td>68.2677</td>
</tr>
<tr>
<td>K=7</td>
<td>432.2051</td>
<td>268.4564</td>
<td>456.2584</td>
<td>547.1234</td>
<td>48.8133</td>
</tr>
<tr>
<td>K=8</td>
<td>516.0121</td>
<td>339.368</td>
<td>591.4568</td>
<td>621.5487</td>
<td>16.2100</td>
</tr>
<tr>
<td>K=9</td>
<td>645.2582</td>
<td>367.8258</td>
<td>679.2465</td>
<td>754.2547</td>
<td>54.5613</td>
</tr>
<tr>
<td>K=10</td>
<td>766.1073</td>
<td>241.8844</td>
<td>790.4658</td>
<td>875.2547</td>
<td>33.0444</td>
</tr>
</tbody>
</table>
The comparison performance shown in the table is making it show as KM, GA, NM, PSO vs. k-mean-GA-NM-PSO people have been reproduced and individual clusters are made in and between them. And calculation of performance details etc. Thus all the sets of KM-GA-NM-PSO algorithms are tested and as well as solutions of high-quality cluster have been developed. Which are designed in the form of distance of the best inter cluster. Also
discovered are the storms standard deviation and the smallest found to near optimal solution of the run other algorithm may trap local optima in some of run. It is found to better results, thus KM-GA-NM-PSO keeps Algorithm a stronger one. This K-MEAN algorithm requires a smaller number compared to other algorithms and it is in relation to the functional visits. In this way we can say that by using the result of K-MEAN in KM-GA-NM-PSO, the GA is in a good way, which is a great way to get access to a great tool from a single GA Is of algorithm produces new generation population from traffic for generation of pig production and the environment is resolved to a new baby environment. In this way a child’s solution has many features of his measurement which can be created from new parents to newborn babies. But still there is not a good start with G.A., a good start with the combination of KM-GA to overcome its shortage can be started and new parents from new parents can be produced, and a suitable population size can also be made. Thus, KM-GA can be better equipped with algorithmic combination than PSO, meaning that the new population can be created at the onset of the cluttering process and can be speeded up in this situation and the health status can be discarded because it Less cluttering needs lesser working people, After we have done all the procedure, we can say that the outcome of PSO and NM-PSO clustering can be revised. With the K-MEAN algorithm, this hybrid algorithm ends with the first K-MEAN algorithm and if there is no change in this cluster's satire rayon vector, in the case of K-PSO, K-MEAN algorithm results in one particle used in the form. The 5N-1 particles start randomly, so this hybrid is used in K-GA-NM-PSO. The 3N-1 angle creates the points continuously and NM-PSO then forms this form to complete the process.
Chapter 4: CONCLUSION AND FUTURE SCOPE

In this thesis, a hybrid method is proposed for efficient data clustering objects. It tries to exploit the merits of four algorithms simultaneously, where the k-means and NM are used for generating the initial solution, GA and PSO are improvement algorithms. The performance of the existing algorithm is compared with other approaches. Moreover, the proposed approach will be combination of the existing algorithm with NM-PSO which can produce high quality clusters with small standard deviation on selected datasets compared to other methods. In future research, the proposed method may be applied to other applications, such as image segmentation, software engineering, data clustering, and various optimization fields. The combination of the KM-GA-NM-PSO with other heuristic approaches and their application to feature selection is also another research contribution.
References:


Buy your books fast and straightforward online - at one of world’s fastest growing online book stores! Environmentally sound due to Print-on-Demand technologies.

Buy your books online at
www.morebooks.shop

Buy your books online at
www.morebooks.shop

Kaufen Sie Ihre Bücher schnell und unkompliziert online – auf einer der am schnellsten wachsenden Buchhandelsplattformen weltweit! Dank Print-On-Demand umwelt- und ressourcenschonend produziert.

Bücher schneller online kaufen
www.morebooks.shop

KS OmniScriptum Publishing
Brivibas gatve 197
LV-1039 Riga, Latvia
Telefax +371 686 20455
info@omniscrptum.com
www.omniscrptum.com