THE PERFORM OF K-MEANS BASED ON HYBRID ARTIFICIAL BEE COLONY AND GENETIC ALGO RITHM IN OPTIMIZING THE CLUSTER SIZE OF K-MEANS

HARUNUR ROSYID

SULTAN IDRIS EDUCATION UNIVERSITY

2024

THE PERFORM OF K-MEANS BASED ON HYBRID ARTIFICIAL BEE COLONY AND GENETIC ALGO RITHM IN OPTIMIZING THE CLUSTER SIZE OF K-MEANS

HARUNUR ROSYID

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENT FOR A DOCTOR OF PHILOSOPHY

FACULTY OF COMPUTING AND META-TECHNOLOGY
SULTAN IDRIS EDUCATION UNIVERSITY

2024

05-4506832 pustaka.upsi.edu.my Perpustakaan Tuanku Bainun Kampus Sultan Abdul Jalil Shah PustakaTBainun ptbupsi

UPSI/PS-3/BO 32
Pind : 00 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

Sila tanda (√)
Kertas Projek ☐
Sarjana Penyelidikan ☐
Sarjana Penyelidikan dan Kerja Kursus ☐
Doktor Falsafah ☑

INSTITUT PENGAJIAN SISWAZAH

PERAKUAN KEASLIAN PENULISAN

Perakuan ini telah dibuat pada 15 (hari bulan) Oktober (bulan) 2024

i. Perakuan pelajar :

Saya, Harunur Rosyid, P20161001084,Komputeran dan Meta - Teknologi (SILA NYATAKAN NAMA PELAJAR, NO. MATRIK DAN FAKULTI) dengan ini mengaku bahawa disertasi/tesis yang bertajuk The Perform Of K Means Based On Hybrid Artificial Bee Colony And Genetic Algorithm In Optimizing The Cluster Size Of K Means

adalah hasil kerja saya sendiri. Saya tidak memplagiat dan apa-apa penggunaan mana-mana hasil kerja yang mengandungi hak cipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hak cipta telah dinyatakan dengan sejelasnya dan secukupnya

Tandatangan pelajar

ii. Perakuan Penyelia:

Saya, Prof. Madya Dr. Muhammad Modi bin Lakulu (NAMA PENYELIA) dengan ini mengesahkan bahawa hasil kerja pelajar yang bertajuk The Perform Of K Means Based On Hybrid Artificial Bee Colony And Genetic Algorithm In Optimizing The Cluster Size Of K Means (TAJUK) dihasilkan oleh pelajar seperti nama di atas, dan telah diserahkan kepada Institut Pengajian Siswazah bagi memenuhi sebahagian/sepenuhnya syarat untuk memperoleh Ijazah Doktor Falsafah (SILA NYATAKAN NAMA IJAZAH).

15 Oktober 2024
Tarikh

Prof. Madya Ts Dr. Muhammad Modi Lakulu
Penyelia
Jabatan Komputeran dan Meta-Teknologi
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak
Tandatangan Penyelia

UPSI/IPS-3/BO 31
Pind.: 01 m/s: 1/1

UNIVERSITI PENDIDIKAN SULTAN IDRIS
SULTAN IDRIS EDUCATION UNIVERSITY

INSTITUT PENGAJIAN SISWAZAH /
INSTITUTE OF GRADUATE STUDIES

BORANG PENGESAHAN PENYERAHAN TESIS/DISERTASI/LAPORAN KERTAS PROJEK
DECLARATION OF THESIS/DISSERTATION/PROJECT PAPER FORM

Tajuk / *Title*: The Perform Of K Means Based On Hybrid Artificial Bee Colony And Genetic Algorithm In Optimizing The Cluster Size Of K Means

No. Matrik /*Matric's No.*: P20161001084

Saya / *I* : Harunur Rosyid

(Nama pelajar / *Student's Name*)

mengaku membenarkan Tesis/Disertasi/Laporan Kertas Projek (Kedoktoran/Sarjana)* ini disimpan di Universiti Pendidikan Sultan Idris (Perpustakaan Tuanku Bainun) dengan syarat-syarat kegunaan seperti berikut:-
acknowledged that Universiti Pendidikan Sultan Idris (Tuanku Bainun Library) reserves the right as follows:-

1. Tesis/Disertasi/Laporan Kertas Projek ini adalah hak milik UPSI.
   The thesis is the property of Universiti Pendidikan Sultan Idris
2. Perpustakaan Tuanku Bainun dibenarkan membuat salinan untuk tujuan rujukan dan penyelidikan.
   Tuanku Bainun Library has the right to make copies for the purpose of reference and research.
3. Perpustakaan dibenarkan membuat salinan Tesis/Disertasi ini sebagai bahan pertukaran antara Institusi Pengajian Tinggi.
   The Library has the right to make copies of the thesis for academic exchange.
4. Sila tandakan (√) bagi pilihan kategori di bawah / *Please tick (√) for category below:-*

☐ **SULIT/*CONFIDENTIAL*** — Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub dalam Akta Rahsia Rasmi 1972. / *Contains confidential information under the Official Secret Act 1972*

☐ **TERHAD/*RESTRICTED*** — Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan ini dijalankan. / *Contains restricted information as specified by the organization where research was done.*

☑ **TIDAK TERHAD / *OPEN ACCESS***

(Tandatangan Pelajar/ Signature)

Tarikh: 18 Oktober 2024

Prof. Madya Ts Dr. Muhammad Modi Lakulu
Penyelaras
Fakulti Komputeran dan Meta-Teknologi
Universiti Pendidikan Sultan Idris
35900 Tanjong Malim, Perak

(Tandatangan Penyelia / *Signature of Supervisor*)
& (Nama & Cop Rasmi / *Name & Official Stamp*)

Catatan: Jika Tesis/Disertasi ini **SULIT @ TERHAD**, sila lampirkan surat daripada pihak berkuasa/organisasi berkenaan dengan menyatakan sekali sebab dan tempoh laporan ini perlu dikelaskan sebagai **SULIT** dan **TERHAD**.

Notes: If the thesis is CONFIDENTAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction.

AKNOWLEDGEMENT

I express my utmost admiration and appreciation to Allah Almighty, who has been my primary provider of knowledge, patience, and resilience during this study endeavor. Pursuing this Ph.D., has been a profoundly transformative experience. This work would not have been feasible without the assistance and direction provided by numerous individuals.

First and foremost, I would like to extend my sincere appreciation to my primary supervisor, Professor Madya Dr. Muhammad Modi Bin Lakulu, and my secondary supervisor, Professor Madya Dr. Ramlah Binti Mailok, for granting me the chance to conduct research and offering invaluable guidance throughout this study. Working and studying under their guidance was a tremendous privilege and honor. I am profoundly appreciative of the support and encouragement I received from both my mentors and the Universiti Pendidikan Sultan Idris Malaysia during the challenging periods I encountered throughout my Ph.D., pursuit.

I hereby dedicate this research project to my late parents, Bapak Abd. Salam, and Ibu Siti Zulaichah, Allahuyarham. Their prayers and words of encouragement have consistently resonated in my mind, urging me to persist with unwavering dedication and courage until the very end. May Allah bestow upon them the Paradise, Greetings, my parents-in-law, Bapak Wachid and Ibu Asmah, have consistently supported and assisted me in every situation.

To my esteemed spouse, Dr. Laila Rochmawati, I express my gratitude for the affection and assistance that provided me with stability, enabling me to successfully complete the project I had initiated. She consistently offered her presence for conversation, attentive listening, effective problem-solving, and unwavering encouragement. My exceptional

offspring, Nasyad Harun and Keisha Shakila Harun, persist in displaying benevolence and inquisitiveness. I have strong affection for you!

I would like to express my sincere gratitude to the Rector of University Muhammadiyah Gresik. Thank you for the support that extends beyond this significant achievement. Lastly, I would like to express my gratitude to my friends for their invaluable support, time, expertise, and encouragement. Without the patience and understanding of others, this work would have been difficult to complete.

ABSTRACT

This study tackles the growing complexity of data by presenting a novel approach to K-Means clustering: the Artificial Bee Colony (ABC) and Genetic Algorithms (GA) K-Means algorithm. The traditional K-Means method has inherent weaknesses, such as arbitrary cluster selection and random initialization of cluster centers. This research addresses these issues by focusing on determining the optimal number of clusters in unlabeled data, a key requirement for effective clustering. The proposed ABC–GA–K-Means algorithm overcomes these challenges through autonomous optimization of data collection and cluster centers. It achieves this by integrating ABC optimization and GA to solve the binary optimization problem often encountered in K-Means clustering. Additionally, the inclusion of Genetic Neighbourhood Generators (GNG) enhances the algorithm's ability to compare results within the ABC network, contributing to improved robustness and efficiency. The study conducts extensive experiments with both simulated and real-world datasets to evaluate the performance of the ABC–GA–K-Means algorithm against conventional clustering techniques, including traditional K-Means, Fuzzy K-Means, and other approaches. The results demonstrate that the proposed algorithm consistently outperforms these methods in terms of accuracy, precision, recall, and rand index. Notably, the ABC–GA–K-Means algorithm achieved high accuracy on datasets like Zoo: 0.9119, Breast Cancer: 0.9413, and Soybean: 0.9575, underscoring its effectiveness in optimizing data collection and cluster centers. These results not only validate the robustness of the proposed algorithm but also highlight its versatility, making it suitable for a wide range of applications. Implication of this innovative approach to clustering points to the potential for further research into alternative heuristics within the ABC–GA–K-Means framework, offering new avenues for advancing the field of clustering.

Keyword: Artificial Bee Colony, Clustering, Genetic Algorithm, Gendata, Initial number of Cluster, K-Means

PELAKSANAAN K-MEANS BERDASARKAN HIBRID ARTIFICIAL BEE COLONY DAN ALGORITMA GENETIK DALAM MENGOPTIMALKAN SAIZ KLUSTER K-MEANS

ABSTRAK

Kajian ini menangani kerumitan data yang semakin meningkat dengan membentangkan pendekatan baru kepada pengelompokan K-Means: algoritma K-Means Koloni Lebah Buatan (ABC) dan Algoritma Genetik (GA). Kaedah K-Means tradisional mempunyai kelemahan yang wujud, seperti pemilihan kluster sewenang-wenangnya dan permulaan rawak pusat kluster. Penyelidikan ini menangani isu-isu ini dengan menumpukan pada menentukan bilangan gugusan optimum dalam data tidak berlabel, keperluan utama untuk pengelompokan yang berkesan. Algoritma ABC–GA–K-Means yang dicadangkan mengatasi cabaran ini melalui pengoptimuman autonomi pusat pengumpulan data dan kluster. Ia mencapai ini dengan menyepadukan pengoptimuman ABC dan GA untuk menyelesaikan masalah pengoptimuman binari yang sering dihadapi dalam pengelompokan K-Means. Selain itu, kemasukan *Genetic Neighborhood Generators* (GNG) meningkatkan keupayaan algoritma untuk membandingkan keputusan dalam rangkaian ABC, menyumbang kepada peningkatan keteguhan dan kecekapan. Kajian ini menjalankan eksperimen yang meluas dengan *datasets* simulasi dan dunia nyata untuk menilai prestasi algoritma ABC–GA–K-Means terhadap teknik pengelompokan konvensional, termasuk K-Means tradisional, Fuzzy K-Means dan pendekatan lain. Keputusan menunjukkan bahawa algoritma yang dicadangkan secara konsisten mengatasi kaedah ini dari segi *accuracy, precision, recall,* dan *rand index*. Terutama, algoritma ABC–GA–K-Means mencapai ketepatan tinggi pada *datasets* seperti *Zoo*: 0.9119, *Breast Cancer*: 0.9413 dan *Soybean*: 0.9575, menekankan keberkesanannya dalam mengoptimumkan pengumpulan data dan pusat kluster. Keputusan ini bukan sahaja mengesahkan keteguhan algoritma yang dicadangkan tetapi juga menyerlahkan kepelbagaiannya, menjadikannya sesuai untuk pelbagai aplikasi. Implikasi pendekatan inovatif untuk pengelompokan ini berpotensi untuk penyelidikan lanjutan dalam heuristik alternatif dalam rangka kerja ABC–GA–K-Means, menawarkan jalan baharu untuk memajukan bidang pengelompokan.

Kata Kunci: Artificial Bee Colony, Kluster, Algoritma Genetik, Gendata, Bilanganawal kelompok, K-Means

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABC	Artificial Bee Colony
GA	Genetic Algorithms
Dataset	A Collection of Data
UCI	University of California Irvine
NA	Nectar Amount
(k)	Number Of Clusters
N	Size of the Bee Colony
MCN	Maximum Cycle Number
CN	Cycles Number
AC	Accuracy
PR	Precision
RE	Recall
RI	Rand Index

CHAPTER 1

INTRODUCTION

1.1 Introduction

The primary challenges encountered in this study of computer science are delineated. A comprehensive examination ensues subsequent to the preliminary stage. The thesis commences with an exploration of relevant research. Based on previous assessments, it is imperative that evaluations of study scope be conducted with a high level of rigor in order to guarantee comprehensive understanding. Furthermore, an assessment will be conducted on the research strategy and objectives employed in problem-solving. The forthcoming investigation will ascertain the aforementioned procedural milestone and administer the aforementioned assessments concurrently in order to mitigate potential misinterpretation.

Prior research evaluated the extent and techniques of the investigation. Firstly, it is essential to do an analysis before proceeding to discuss the subsequent two points. Comprehensive evaluations yield the most significant scholarly findings. In order to

mitigate potential confusion, it has been determined that both exams will be conducted concurrently.

This section provides an overview of the pertinent literature and outlines the research problem. The selection of topics and objectives significantly influence the structure and methodology of a literature review. Frameworks will serve as the central point of emphasis for the research. The determination of these frameworks is contingent upon the objective and complexity of the research. The investigation will be guided by the prevailing conditions and research objectives. It serves as the foundation for both perspectives. Numerous studies have demonstrated statistical significance. Review articles often advocate for further research in order to gain a deeper understanding of the subject matter. The evaluation of study approaches and aims is undertaken. In order to prevent any potential confusion, it is advisable to address both phases of the examination. The initial step involves defining the scope of the study, followed by determining the appropriate research methodologies to be employed. Two scholarly articles explore unconventional research inquiries.

The formulation of the issue statement and the comprehensive examination of existing literature provide the conceptual framework for conducting the investigation. The examination of contemporary concerns and the impetus behind research endeavors will guide the process of conducting a comprehensive evaluation of existing literature and formulating an effective research strategy. There are variations in study frameworks based on the specific topic and underlying purpose. The study is driven by a certain objective.

1.2 Research Background

Clustering in the field of computer science facilitates several computational tasks such as data mining, statistical analysis, dimensionality reduction, and vector quantization. Unsupervised clustering is a process in which data, feature vectors, observations, and patterns are subjected to clustering. The objective of clustering is to enhance the homogeneity within clusters while reducing the heterogeneity across clusters. The establishment of hierarchical clustering is necessary.

The arrangement of nodes in hierarchical dendrograms is organized into distinct layers. A cluster or subgroup has been found to improve mathematical or statistical performance. Clustering is a commonly employed technique by data scientists to achieve diverse objectives. The process of clustering aids in the understanding and decision-making process by effectively uncovering fundamental attributes and relationships within extensive datasets. The subject of computer science is expected to play a significant role in promoting the widespread utilization of clustering methods, hence enabling innovative research in several academic domains. (Chakraborty & Das, 2017). This study demonstrates the modifications in compatibility of the K-Means Partitional Algorithm. Clustering of samples improves the analysis of datasets. Businesses have the ability to participate in any category, regardless of the outcome being determined by luck. Clustering is the most challenging *(k)* selection. The letter *(k)* is crucial as it influences the formation of clusters. Theme involving multiple groups.

Clustering is required in this investigation to obtain samples that can be compared. The K-Means Partitional Algorithm has the potential to alter research

methodologies. In 1967, Mac Queen introduced the concept of K-Means clustering, which is still widely utilised today. Outliers exhibit strong clustering patterns within large datasets. Clustering involves the creation of *(k)* centroids in a random manner, followed by the grouping of data points based on their similarity. Utilise similarity as a criterion to partition observations in a clustering dataset n important challenge in K-Means clustering is the selection of the optimal value for *(k)*, which might have a wide range of possible possibilities.

K-Means clustering is effective for datasets with globular clusters, despite its limitations. K-Means clustering, a method devised by Mac Queen in 1967, is highly adaptable and resilient, making it suitable for a wide range of datasets that may contain outliers. Centroids optimise clustering and separation. K-Means, an adaptable and productive method, is widely utilised in numerous academic disciplines. In order to achieve success, it is necessary to have a careful and determined approach to ensure that the findings of clustering meet the goals of the study (D.Sharmilarani N. G., 2014). K-means clustering was chosen in the research due to its ability to partition the data into distinct clusters based on similarity between data objects. This algorithm is known for its simplicity and time efficiency, making it suitable for online or dynamic data. Additionally, the k-means algorithm has a complexity of $O^{(ntk)}$, where n is the size of the data set, t is the number of iterations, and k is the number of clusters, making it a practical choice for dynamic data analysis. The main problem in k-means clustering is the need to specify the number of clusters *(k)* prior to the execution of the algorithm. This can be a challenge as the optimal number of clusters may not be known beforehand and can vary depending on the characteristics of the dataset. Additionally, as the k value

increases, the number of iterations also increases, which can impact the efficiency of the algorithm.

The constraint of K results in clustering that is not global in nature. Due to its approximation of the centre point of scattered data, it is unable to cluster objects of varying sizes, densities, and non-globular geometries. K-Means solely approximates centroids for scattered data, rendering it inappropriate for intricate non-spherical datasets. K-Means clustering is not effective for complex datasets when it comes to straightforward clustering. The presence of empirical cluster number and stochastic K-Means centre initialization poses a challenge in effectively grouping inefficient datasets.

Considering these limitations, computer scientists should use discretion while employing K-Means clustering for intricate data patterns. Alternative clustering techniques designed for non-global patterns and intricate data structures might offer improved accuracy and resilience. The year 2015 was authored by Xiaolong Wang. The majority of clustering algorithms necessitate the determination of the optimal number of clusters. Static databases require data; hence these parameters should remain unchanged. According to Chatti Subbalakshmia (2015), dynamic databases need to regularly update cluster numbers in order to accurately represent data trends and clustering.

The number of clusters is a crucial factor in data modelling within the field of computer science. Data clustering mostly relies on the utilisation of elbow and silhouette coefficients. Methods enhance the efficiency of clustering for data representation. The utilization of the Elbow technique is crucial as inaccurate cluster

predictions have the potential to compromise the integrity of data models. The success of a project relies on the value of *(k)* in data clusters. Cluster number data may hinder activities. Projects might be facilitated or hindered by various data key clusters.

The elbow method and silhouette coefficient are used to determine the number of clusters. Their forecasts encompass both the size and growth of clusters. Nevertheless, each option has advantages and disadvantages. While the Elbow approach has limited estimation capabilities, there are clustering approaches available that do not rely on truth. The elbow method is facing data modelling issues as a result of inadequate clustering. The Silhouette Coefficient does not provide any information. Analysing the elbow method provides a clear understanding of the variation among clusters. Sparse cluster models are advantageous when there are several clusters. Accurate data modelling necessitates a specific number of clusters. Silhouette coefficients are used to identify clusters, while Elbow techniques are used to measure the amount of variance.

Clustering in data modelling necessitates the selection of numerical values. The silhouette coefficient evaluates the level of separation between clusters in the field of computer science. Evaluate the discriminability of clusters during the clustering process. K-Means clustering encounters difficulties when dealing with extensive datasets containing numerous clusters. K-Means membership is determined by hierarchical clusters. Managing clusters excessively is unattainable due of their inherent complexity.

The aforementioned alterations make K-Means clustering unsuitable for datasets including a large number of clusters, reaching into the thousands. In certain cases, the

efficacy of K-Means clustering is limited, hence requiring the utilization of additional approaches to effectively analyze many clusters. The year 2015 marks a significant point in time. The subject matter at hand pertains to my current financial circumstances. In order to effectively implement the K-Means algorithm, it is important to possess prior knowledge of data or make predictions based on clusters (Zalik, 2008). The process of identifying the most suitable number of clusters in the K-Means algorithm is commonly employed in practice. According to Chiang (2009), the occurrence of cluster intermix has a substantial influence on the results of clustering. The unpredictability of the convergence of the technique towards the local optimum arises from the stochastic nature of cluster centroid formation and the variability in the number of clusters present in the datasets. Centroid sensitivity is a concept that pertains to the measure of how sensitive a system is to changes in its centroid.

The conventional K-means clustering algorithm is a clustering technique that necessitates a pre-established number of clusters. According to Jing Xiao (2010), the algorithm's sensitivity could potentially lead to convergence towards a local optimum. The cluster centroids may exhibit imperfections when the iterative process approaches a solution that is locally optimal. In their study, Erisoglu (2011) presents a proposed optimal solution for the identification of the most suitable neighbor. The research investigation centers on two distinct methodologies for quantifying clusters. The initial stage involves the utilization of heuristics. The clustering process is ongoing and yields an infinite number of clusters. The second part involves the development of a methodology for cluster selection in a finite mixture model, utilizing its components (Z˘alik, 2008). The demonstration of population observation probability is accomplished through the utilization of mixed models. Probabilistic mixed models are

employed to characterize subpopulations within a larger population in instances where there is an absence of observed data. A heuristic algorithm is employed for the purpose of optimizing pre-determined data clustering. In his research, Karaboga (2011) utilized the ABC approach for the purpose of classifying clusters. Celal (2014) conducted a study on the discrete implementation of artificial bee colonies (ABC) for the purpose of clustering datasets.

The study specifically examined the impact of the number and starting points of clusters. In the year 201, Rana Fosanti made improvements to the data clustering process by implementing BCO cloning and emphasizing the need of fairness. In the year 2016, a novel approach known as the non-dominated sorting-based multi-objective Artificial Bee Colony algorithm was proposed alongside the concept of data clustering. Certain individuals employed the techniques of stochastic optimization and evolutionary grouping

.

In 2006, Glaszlo and Mukherjee proposed the utilization of a genetic algorithm as a means to enhance the expansion of cluster centers within the K-Means clustering technique. The AGCUK algorithm, as proposed by Yangguo in 2011, is utilized to ascertain the optimal number of clusters and the suitable partitioning for an unknown value of *(k)* in the context of automatic genetic clustering. This study aims to improve the effectiveness of K-Means clustering through the utilization of Artificial Bee Colony (ABC) algorithm and Genetic algorithm (GA). The prior methodology, which required the user to submit the number of clusters as a parameter, is deemed ineffective due to the unavailability of many real-world applications to provide this information. In light of these observations, scholars have conducted inquiries into heuristic methodologies,

such as the Artificial Bee Colony (ABC) algorithm, in order to ascertain the most suitable quantity of clusters inside a certain dataset. The convergence of ABC is impeded by complex issues. In 2007, a study was undertaken by D. Karaboga. Bees utilize the ABC method to forage for food by transferring unidimensional data to a randomly selected adjacent bee. As the magnitude of an entity expands, there is an enhancement in the capacity to locate sustenance, albeit accompanied by a decline in the dissemination of knowledge. Hierarchical and partitional clustering techniques are designed to decompose entities into smaller structures and a pre-defined number of clusters.

The objective is to optimize the hierarchical and partitioned object clustering technique. Hierarchical and partitional clustering utilize different methodologies to attain comparable results. The implementation of the genetic algorithm crossover operator facilitates improved inter-colony communication within bee populations. The optimization of the efficiency of genetic algorithms. In an effort to optimize the efficiency of Genetic Algorithms (GA), Yan (2012) employed a technique to improve their effectiveness. Nevertheless, the assessment of benchmark functions and data clustering revealed that certain functions continued to be confined inside local minimum points. The user's text is void of any content. The user's text is too short to be rewritten academically.

1.3 Problem Statement

From those literature review the main point case for this research problem shown in table 1,1., Randomly choosing K-Means clusters has been shown in computer science literature can cause inefficient globular clusters to develop and to lose efficiency as *(k)* increases. In order to optimize clustering performance and mitigate associated difficulties, it is advisable to decrease the number of clusters to a value below the threshold denoted as *(k)*.

Table 1.1.

Literature a of Problem Statement

Author	Problem	Scope
D.Sharmilarani, N. G 2014	Determination of the number of clusters in unlabeled data, which is a basic input for most clustering algorithms.	Number of Cluster
Xiaolong Wang, 2015	Selection of cluster number k and initial K-Means center has certain influence on the result. It would generate very different aggregation result when confronting with some certain types of data set.	
Chatti Subbalakshmia, 2015	User has to specify the optimum number of clusters prior to execution, for static databases this value remains constant whereas, in the case of dynamic databases the value should be changed.	

The optimal number of clusters is smaller than the value of *(k).* The user's text is a single letter, *(k)*. The adjustment of the sensitivity of the initial cluster centroids introduces an additional layer of intricacy to the K-Means clustering procedure. Due to its inherent sensitivity, the algorithm possesses the capacity to attain a local solution, which may not necessarily represent the most optimal clustering of the dataset. Various

optimization methods, including heuristic, stochastic, and deterministic approaches, are employed to determine the most optimal clustering algorithm and solution.

The research aims to identify the most optimal K-Means clusters, albeit it does not accomplish this task automatically. Prior to the incorporation of centroids, clustering problems were tackled through the utilization of genetic algorithms (GA) and artificial bee colony (ABC) methods. The integration of the crossover operator from genetic algorithms (GA) with the intelligent exploration mechanism of artificial bee colony (ABC) enhances the clustering performance of the ABC-GA-K-Means algorithm.

Evolutionary algorithms such as Genetic Algorithm (GA) and Artificial Bee Colony (ABC) can be used to overcome the K-Means problem by providing alternative optimization techniques for clustering. The K-Means problem involves partitioning a set of data points into a predefined number of clusters, with the objective of minimizing the within-cluster variance.

GA and ABC can be applied to the K-Means problem by formulating the clustering task as an optimization problem, where the objective function is to minimize the within-cluster variance or a similar clustering quality measure. These evolutionary algorithms can then be used to search for the optimal cluster centroids or cluster assignments by iteratively updating the solutions based on the fitness of the clusters.
For example, GA can be used to evolve a population of potential cluster centroids, where the crossover and mutation operations are applied to generate new centroid configurations, and the fitness of each configuration is evaluated based on the within-

cluster variance. Similarly, ABC can be employed to explore the search space of cluster centroids by simulating the foraging behavior of bees, with employed bees representing potential solutions and onlooker bees evaluating the quality of the solutions based on the clustering objective.

These evolutionary algorithms provide a stochastic and population-based approach to optimizing the clustering process, offering an alternative to the deterministic and iterative nature of the traditional K-Means algorithm.

The combination of Genetic Algorithm (GA) and Artificial Bee Colony (ABC) was motivated by the drawbacks of each algorithm and the potential benefits of their integration. The drawbacks of GA and ABC that led to their combination include:

1. GA's drawback: GA can suffer from premature convergence and struggles with maintaining diversity in the population, especially in complex search spaces.

2. ABC's drawback: ABC, in its basic form, is not directly adaptable to binary optimization problems, and modifications are required for its application to binary problems.

The rationale behind the combination of GA and ABC was to leverage the strengths of both algorithms and mitigate their individual drawbacks. By integrating GA's crossover operator into ABC, the combined algorithm aimed to enhance exploration and exploitation capabilities, improve population diversity, and address the challenges associated with binary optimization problems. The integration of genetic operators, such as crossover and swap, into the ABC algorithm aimed to provide a more

effective search mechanism for binary optimization problems, thereby overcoming the limitations of the basic ABC algorithm in handling binary domains.

The drawback of ABC that requires the integration of GA's crossover operator is the need for a more effective mechanism for exploring the search space and generating diverse solutions, particularly in the context of binary optimization problems. The integration of GA's crossover operator into ABC aimed to address this limitation and enhance the algorithm's performance in handling binary optimization problems.

In summary, the combination of GA and ABC aimed to capitalize on their respective strengths and mitigate their individual drawbacks, ultimately leading to a more effective and versatile optimization algorithm, particularly in the context of binary optimization problems. The utilization of crossover operators enhances the efficacy of ABC-GA-K-Means algorithms. Algorithmic adjustments have the capacity to effectively remove performance gaps, hence leading to an improvement in convergence. The K-Means technique enhances hybrid clustering by doing an exhaustive exploration at both the global and local scales.

The usefulness of ABC-GA-K-Means is demonstrated by comparing it to ordinary K-Means clustering on both simulated and actual datasets. The ABC-GA-K-Means algorithm produces superior clustering results compared to the average. The Hybrid K-Means clustering algorithm has superior speed and accuracy.

1. Does minimum local optima can be handled by optimizing number of cluster before initial centroid in K-Means Clustering?

2. How to Develop Genetic Algorithm and Artificial Bee Colony to optimize initial number of cluster in K-Means Clustering problem?

3. What optimal parameter are used to optimize initial number of cluster (k) before initial centroid in K-Means Clustering based heuristic approach?

4. What methods are suitable for testing and validating proposed algorithm performance?

1.4 Research Objective

1. To investigate how to minimize local optima problem based optimizing number of cluster before initial centroid in K-Means Clustering

2. To Develop Hybrid Genetic Algorithm and Artificial Bee Colony to optimize initial number of cluster in K-Means Clustering problem.

3. To determine optimum parameter to optimize initial number of cluster (k) before initial centroid in K-Means Clustering based heuristic approach.

4. To validate and measure performance of proposed algorithm.

1.5 The Limitation of this Research

K-Means clustering is widely recognized and widely used in traditional applications due to its efficiency and popularity. As the complexity of the dataset increases, there is a potential for a decline in its effectiveness. The primary cause of this issue can be attributed to challenges related to empirical cluster selection and the random initialization of K-Means centroids. Dynamic databases exhibit frequent changes in order to adapt to evolving data patterns, whereas static databases possess a predetermined quantity of clusters that remains constant.

A technique was developed by researchers to optimize cluster centroids automatically by considering data homogeneity and cluster heterogeneity, with the aim of minimizing uncertainty in the process of data clustering. This approach effectively arranges data by taking into account the coherence and dissimilarity within clusters. This enhanced clustering technique offers significant advantages for segmentation and massive data processing.

The present methodology improves the efficiency and accuracy of data clustering by specifically addressing the limitations associated with K-Means clustering. The enhancement of cluster analysis and data insights can be achieved through the optimization of cluster numbers and the sorting of data based on measures of homogeneity and heterogeneity.

The proposed approach for optimizing data clustering clusters effectively addresses intricate challenges associated with complicated datasets. This methodology

enhances the field of cluster analysis by including the principles of data homogeneity and cluster heterogeneity, hence enhancing the outcomes and potential applications across several academic domains.

1.6 Research Scope

In this computer science article, the suggested strategy for optimizing the initial number of clusters and cluster centroids is rigorously validated using a range of real-world categorical datasets, namely zoo, breast cancer, soybean, lung cancer, mushroom, and dermatological datasets from the UCI Machine Learning Repository. The clustering findings were evaluated using Yang's accuracy measure and the Rand Index in four simulations, showcasing the efficacy of the k'-means algorithm.

To assess the performance of the clustering strategy, a series of experiments were conducted. Experiment 1 focused on evaluating accuracy, Experiment 2 examined precision, and Experiment 3 analyzed recall. Furthermore, Experiment 4 investigated the effectiveness of the Rand Index in evaluating the clustering findings. These experiments utilized synthetic datasets from the UCI repository.

The results of the rigorous testing approach, encompassing both real-world and synthetic datasets, confirmed the effectiveness of the suggested strategy. Yang's accuracy measure and the Rand Index consistently demonstrated that the clustering strategy produced promising results. The experiments collectively demonstrated the ability of the k'-means algorithm to effectively cluster data.

In conclusion, the comprehensive validation process, involving various real-world and synthetic datasets, lends strong support to the proposed strategy for optimizing the initial number of clusters and cluster centroids. The utilization of established evaluation metrics reaffirms the efficacy of the k'-means algorithm in producing reliable and accurate clustering outcomes.

1.7 Dissertation Organization

The dissertation comprises six chapters, each contributing to the comprehensive investigation. Chapter I serves as an introduction, outlining the research problem, research goals, and pertinent issues that will be addressed throughout the study. In Chapter 2, a concise overview of prior research on the K-Means Cluster issue is provided, with a particular focus on cluster number optimization before initializing centroids. The chapter delves into the examination of various K-Means clustering number cluster optimization heuristic techniques, including Artificial Bee Colony, Genetic Algorithm, and hybridization. Technical analysis of the research challenge and proposed solutions culminate this chapter.

Chapter 3 is dedicated to describing the research methodology, which is structured into four distinct parts. The preliminary study phase encompasses the initial exploration of one or more research objectives, setting the foundation for subsequent analyses. In Chapter 4, the method and benchmark model chosen for evaluating the research are elaborated upon in detail, ensuring a robust and systematic evaluation process. Chapter 5 centers on the assessment of the Representation Model, meticulously

analyzing its effectiveness and performance within the context of the research objectives.

Lastly, Chapter 6 provides an insightful glimpse into future research directions and potential avenues for further exploration. This concluding chapter also offers a comprehensive summary of the dissertation, reiterating its key findings and contributions. Overall, the dissertation aims to advance our understanding of the K-Means Cluster issue, while exploring innovative approaches to cluster number optimization and paving the way for future research endeavors in this field.

1.8 Terminology or Definition of Operation

1.8.1 K-Means

The K-Means clustering approach holds a prominent position among the most widely utilized and established clustering algorithms in computer science. Originally proposed by Mac Queen in 1967, the K-Means algorithm has withstood the test of time and remains a favored method for data clustering. The fundamental concept of K-Means involves the random selection of *(k)* initial cluster centroids, where *(k)* denotes the pre-determined number of clusters before the algorithm is applied. Each selected centroid serves as the starting point for a separate cluster, and the clustering process commences for all *(k)* clusters concurrently (Mac Queen, 1967).

During the execution of the algorithm, a fixed number of iterations take place, which allows for the precise grouping of data into their respective categories. This is achieved by employing a criterion function that aims to minimize the distance between data points within the same cluster while maximizing the separation between data points in different clusters. As a result of this process, data points sharing similarities are placed in proximity to one another within the same cluster, while data points exhibiting dissimilarities are assigned to different clusters. The output of the K-Means algorithm thus presents a clear and organized categorization of the input data into distinct clusters based on their intrinsic characteristics.

1.8.2 Genetic Algorithm

The Genetic Algorithm falls within the category of "evolutionary algorithms," inspired by principles of natural selection (Goldberg, 1989). Widely employed for resolving optimization and search problems, genetic algorithms utilize operators like mutation, crossover, and selection, mirroring physiological processes (Holland, 1975). Genetic programming leverages these operators to pursue optimal outcomes via the fitness function, aiming to optimize available opportunities (Koza et al., 1992). The "objective function" constitutes a vital element in traditional optimization, providing a framework for understanding surrounding circumstances.

The algorithm minimizes the fitness function value to achieve its objective, emphasizing the importance of recording the fitness function, either as a file or an anonymous function, for documentation purposes. Integrating the fitness function as an

argument into the core genetic algorithm function handle enhances its functionality (Mitchell, 1998). Every point that can be applied to a fitness function represents an individual. The significance of individual fitness function scores cannot be understated in the context of physical well-being. Assume that the fitness function is.

$$f(x_1, x_{2,}, x_3) = (2x_1 + 1)^2 + (3x_2 + 4)^2 + (x_3 - 2)^2 \quad (1.1)$$

The vector (2,-3.1) whose length is the number of variables in the problem, is an individual. The score of the individual (2,-3.1) is f(2,-3.1) = 51. An individual is sometimes referred to as a genome and the vector entries of an individual as genes.

1.8.3 Artificial Bee Colony

The Artificial Bee Colony Algorithm (ABC), originating from the University of California, Berkeley, is a nature-inspired metaheuristic mirroring the foraging behavior of bee colonies (Karaboga, 2005). The colony comprises various types of bees, including hired bees, observers, and scouts, each fulfilling distinct roles to meet collective needs. Within the ABC algorithm, the colony segregates into employees and observers, organizing labor based on specific tasks. Efficiency is ensured by assigning each food source to a single worker bee, responsible for exploring and exploiting nearby sources. Upon depleting nectar from its assigned source, a worker bee transitions into a scout to search for new sources, risking cessation of nectar consumption by other bees from the original source.

In the ABC algorithm, food sources symbolize potential optimization solutions, with fitness determined by nectar production (Karaboga & Akay, 2009). Bees in the ABC algorithm denote food sources, and the population's active bees or observers dictate the number of optimization problems addressed concurrently. With finite resources and abilities, each bee can only tackle a few problems simultaneously, prompting the orchestration of collective bee efforts by the ABC algorithm to effectively navigate problem spaces and seek high-quality solutions.