INSIGHT STUDY OF VARIOUS MODEL USED FOR ONLINE ANALYTICAL PROCESS

A Thesis

Submitted towards the Requirement for the Award of Degree of

DOCTOR OF PHILOSOPHY

IN

COMPUTER SCIENCE AND APPLICATION

Under the Faculty of Computer Science and Application

By

ANJANA YADAV
(Enrollment No – 161595506208)

Under the Supervision of

Dr. Balveer Singh
Professor, Department of Computer Science & Engineering
P.K. UNIVERSITY

Year-2023
P.K. University
NH-27, Village. Thanra (P.O. - DINARA),
Shivpurii M.P. India–473665
www.pkuniversity.edu.in

CERTIFICATE OF THE SUPERVISOR

This is to certify that the work entitled **"Inside Study of Various Model used for Online Analytical Process"** is a piece of research Work done by Mrs. Anjana Yadav Under my Guidance and Supervision for the degree of **Doctor of Philosophy of Computer Science and Application,** P.K. University (M.P) India.

I certify that the candidate has put an attendance of more than 240 day with me. To the best of my Knowledge and belief the thesis:

I – Embodies the work of the candidate herself.

II – Has duly been completed.

III – Fulfill the requirement of the ordinance relating to the Ph.D. degree of the University.

Signature of the Supervisor

Dr.Balveer Singh

Professor,P.K.University

Date: 28/07/2023

DECLARATION BY THE CANDIDATE

I declare that the thesis entitled **"Inside Study of Various Model used for Online Analytical Process"** is my own work conducted under the supervision of **Dr. Balveer Singh** Supervisor at P. K. University, Shivpuri Approved by Research Degree Committee. I have put more than 240 days of attendance with Supervisor at the center.

I further declare that to the best of my knowledge the thesis does not contain my part of any work has been submitted for the award of any degree either in this University or in any other University Without proper citation.

Signature of the candidate

Date: 10/ July /2023

Place: Shivpuri

FORWARDING LETTER OF HEAD OF INSTITUTION

The Ph.D. thesis entitled **"Inside Study of Various Model used for Online Analytical Process"** Submitted by **Smt. Anjana Yadav** is forwarded to the university in six copies. The candidate has paid the necessary fees and there are no dues outstanding against her.

Name Dr. Balveer Singh

Date: 28/07/2023.

Place: Shivpuri

Seal

(Signature of Head of Institution where the Candidate was registered for Ph.D degree)

Signature of the Supervisor Date:

Date: 28/07/2023.

Place: Shivpuri

Address: Shivpuri (MP)

ACKNOWLEDGEMENT

I would like to thank all the people who contributed in some way to the work described in this thesis. At this moment of accomplishment, first of all I would like to pay the homage **to Sh. Jagdish Prasad Sharma Hon'ble Chancellor** P.K.University, Shivpuri who made this glorious temple to realize spiritual, technical and scientific knowledge about this vast existing universe. I express my thanks to **Prof.(Dr.) Ranjit Singh, Hon'ble Vice Chancellor**, P. K. University, Shivpuri,M.P. for his cooperation and also for giving me the opportunity to do the research work. I am indeed grateful to **Dr. Jitendra Kumar Mishra,** Director Admin and Academics, P. K. University, Shivpuri, for providing the necessary facilities, infrastructure and encouragement for completion of thesis work.

I would like to express my sincere and whole hearted gratitude to **Dr.Deepesh Namdev**, Registrar, P. K. University, Shivpuri. I take this opportunity to express my gratitude and sincere thanks to **Dr.Bhasker Nalla,** Dean Research, continuous support of my Ph.D study and related research. His guidance helped me in all the time of research and writing of this thesis.

I embrace the opportunity to express my deep sense of gratitude to my mentor and supervisor **Dr. Balveer Singh**, Professor Department of Computer Science and Engineering, P. K. University, Shivpuri, for his able guidance, valuable help enthusiastic attitude and suggestions throughout the period of my research work. I am very fortunate for having the opportunity to work under him. He had been a perennial source of inspiration and his guidance helped me to find the ways through hurdles.

I am indeed grateful to **Mrs. Aiman Fatima, Mrs. Nisha Yadav,** Library In-charge and **Er.Pankaj Sharma ,IT Head** for their cooperation for completion of thesis work.

I would like to express my sincere thanks to late **Dr. Anand Kumar Tripathi** for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge.

I would not complete my duty if I omit to pay my sincere gratitude to all the persons, friends and relatives, who have directly or indirectly helped me in the completion of this work.

ANJANA YADAV

P.K.UNIVERSITY
SHIVPURI (M.P.)

UniversityEstablished Under section 2F of UGC ACT 1956 Vide MP Government Act No 17 of 2015

CENTRAL LIBRARY

Ref. No. C.LIB/PKU/2023/Ph.D Scholar/093 Date: 01.03.2023

CERTIFICATE OF PLAGIARISM REPORT

1. Name of the Research Scholar	:	Anjana Yadav
2. Course of Study	:	Doctor of Philosophy (Ph.D.)
3. Title of the Thesis	:	Insight study of various model used for online analytical process
4. Name of the Supervisor	:	Dr. Balveer Singh
5. Department	:	Computer Science Engineering & IT
6. Subject	:	Computer Science
7. Acceptable Maximum Limit	:	10% (As per UGC Norms)
8. Percentage of Similarity of Contents Identified	:	0%
9. Software Used	:	Ouriginal (Formerly URKUND)
10. Date of Verification	:	08.09.2022

Signature of Ouriginal Coordinator
(Librarian, Central Library)
P.K.University, Shivpuri (M.P.)

LIBRARIAN
P.K. University
Shivpuri (M.P.)

ADD: VIL: THANRA, TEHSIL: KARERA, NH-27, DIST: SHIVPURI (M.P.) -473665
MOB: 7241115902, Email: library.pku@gmail.com

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis:

"**Inside Study of Various Model used for Online Analytical Process** "

Candidate's Name: **Anjana Yadav**

COPYRIGHT TRANSFER TO

The undersigned hereby assigns to the P.K. University, Shivpuri all copyrights that exist in and for the above thesis submitted for the award of the Ph.D. degree.

Date:

Place:P.K.University,Shivpuri

Anjana Yadav

Name of the scholar

Note: However, the author may reproduce/publish or authorize others to reproduce, material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the University's copyright notice are indicated.

LIST OF ABBREVIATION"

ACO	-	Ant Colony Optimization
ALO	-	Ant Lion Optimization
CF	-	Configuration
CS	-	Consistency
DFOC	-	Dragon Fly Optimization based Clustering
ER	-	Entity Relationship
ETL	-	Extraction Transformation Loading
FA	-	Foodstuff sources Appeal
FFO	-	Fruit Fly Optimization
GA	-	Genetic Algorithm
GWO	-	Grey Wolf Optimization
HOLAP	-	Hybrid Online Analytical Process
HTML	-	Hyper Text Mark-up Language
IoMT	-	Internet of Medical Things
IoT	-	Internet of Thing
IT	-	Information Technology
K-SSOC	-	KMeans-Salp Swarm Optimization based Clustering
KPIs	-	Key Performance Indicators
MAC	-	Multidimensional Aggregation Cube
MATLAB	-	Matrix Laboratory
MOLAP	-	Multi-dimensional Online Analytical Process
MS	-	Maintenance Spending
OE	-	Opponent Escaping
OLAP	-	Online Analytical Process
OLTP	-	Online Transactional Processing
PDF	-	Portable Document Format
PSO	-	Particle Swarm Optimization
RDBMS	-	Relational Data Base Management System
ROLAP	-	Relational Online Analytical Process

SBI	-	Social Business Intelligence
SQL	-	Structured Query Language
SR	-	Standard Severance
UML	-	Unified Modeling Language
WMS	-	Whole Maintenance Spending
WQS	-	Whole Query Spending
WSF	-	Whole Maintenance Function
XML	-	Extensible Mark-up Language

"LIST OF TABLES"

"LIST OF FIGURES"

"ABSTRACT"

The Online Analytical Processing (OLAP) based Multidimensional examination hassles for several stockpiling magnificence over huge data. For as much to recognize queries answering time companionable by OLAP framework users and understanding entire business perceive mandatory, OLAP data is structured as a data cube (a multidimensional model). The OLAP queries are responded in speedy and steady time by utilizing the cube materialization for assessments takers. But, this also involves unendurable expenses, regarding to stockpile memory and period, and as a data depot, OLAP has an average dimension and dimensionality which is to be significant on query processing. Consequently, cube assortment has got to be finished motivating to diminish inquiry management expenses, maintaining as a restraint the materializing gap. Several techniques and heuristics like deviationist and insatiable algorithms have been utilized to offer an estimated result. In this work, a Fruit Fly Optimization (FFO) approach is implemented in a lattice structure to obtain an optimal materialized data cube for reducing the query processing expenses. The results illustrate that FFO generates better performance than Particle Swarm Optimization (PSO) in terms of frequency and number of dimensions.

The data cube assessments dependent on Online Analytical Processing (OLAP) trouble for numerous depositing splendors over broad information. In favor of appreciating question answering era pleasant with OLAP skeleton patrons and allowing complete industry organized notice compulsory, OLAP information is organized as a data cube model. The OLAP questions are answered in rapid and sturdy time by exploiting the cube embodiment for appraisals buyers. Until now this moreover insets insupportable charge, concerning to accumulation remembrance and time, yet as a data storage area had a typical length and extent which will be influential on stimulating procedure. Thus, cube classification has visited to be refined fascinating to moderate question managing charge, preserving as a control the materializing breach. Numerous strategies and heuristics like divergence and voracious approaches have been exploited to suggest a vague solution. Here, a Grey Wolf Optimization (GWO) strategy is exploited in a lattice structure for finding the best data cube to decrease the question processing charge. The outputs describe the superior efficiency of GWO against GA, PSO and ALO based on total dimensions and frequency.

Medicine is a fresh way to utilize for curing, analyzing and detecting the diseases through data clustering with OLAP (Online Analytical Processing). The large amount of multidimensional clinical data is reduced the efficiency of OLAP query processing by enhancing the query accessing time. Hence, the performance of OLAP model is improved by using data clustering in which huge data is divided into several groups (clusters) with cluster heads to achieve fast query processing in least time. In this chapter, a Dragon Fly Optimization based Clustering (DFOC) approach is proposed to enhance the efficiency of data clustering by generating optimal clusters from multidimensional clinical data for OLAP. The results are evaluated on MATLAB 2019a tool and shown the better performance of DFOC against other clustering methods ACO, GA and K-Means in terms of intra-cluster distance, purity index, F-measure, and standard deviation.

The performance of query processing over OLAP (Online Analytical Processing) model is decreased due to higher query access time for huge multidimensional data. Therefore, the clustering is introduced to improve the OLAP model efficiency by getting quick query processing because of dividing the large data into various clusters. The K-Means is a famous technique of clustering the data into groups to solve various real life issues. However, K-Means has some drawbacks like sensitivity to primary centroid assortment in cluster and local optimum convergence. Hence, a KMeans-Salp Swarm Optimization based Clustering (K-SSOC) is implemented to improve the performance of K-Means by providing optimal clustering over huge OLAP multidimensional data. The outcomes are obtained on MATLAB 2019a environment based on the parameter purity index, standard deviation, F-measure, intra-cluster distance and running time complexity over 1000 iterations. The results illustrate the superior performance of K-SSOC against K-Means, ACO and PSO over total six multidimensional datasets based on parameters.

"TABLE OF CONTENTS"

CHAPTER-1
INTRODUCTION

1.1. Introduction

Database [95] and data warehouse applications deal with huge amount of data. In real life applications managing the vast quantity of data is an immense dispute. Optimizations techniques [74] are required to reduce the space and time constraint. Even though data warehouse processes are not transactional processing, time and space complexity optimizations are crucial to manage terabytes of data. The construction of warehouse as well as loading the data into it takes place generally in batch mode. However the decision making has to be online, to support business intelligence.

Sometimes, there are situations where the warehouse formation as well as the decision making is online, yielding the concept of virtual data warehouse. However in this thesis we do not incorporate any research work on virtual data warehouse.

The Online Analytical Processing (OLAP) is highly suitable for examination and processing of healthcare data by utilizing the consistent models in IoT devices. The OLAP models are introduced on large medical data warehouses for computation and decision making to reduce the processing overhead of data in IoT devices. The multidimensional, relational and hybrid representation of healthcare data are specifically used for describing the OLAP models [94]. The multidimensional data cube model is examined in sturdy time for processing of OLAP queries over IoT devices. The query processing is highly expensive and time consuming for large medical data warehouses.

This cost and time of query processing are reduced by using several strategies and heuristics to offer efficient OLAP examination by improving the performance materialized medical data cube selection strategy in IoT devices. The query processing is also improved by using the data clustering [77, 78] over huge data warehouses storing the different types of information such as medical, defense, industrial and confidential transmitted through IoT devices. These numerous amounts of medical

information are easily and quickly accessed and processed by IoT users, if the data is arranged into groups using clustering [80] rather than in raw form.

The main motive of this research paper is given an extensive collection of valuable healthcare data about OLAP analysis, models, query processing techniques, problems and optimization methods in various environments using IoT devices. Hence, the researchers can be utilized the important medical data over IoT devices to implement novel and efficient techniques for OLAP data analysis [69]. A brief literature of prior researches is further illustrated in textual and tabular arrangements to give the inspiration to researchers for implementing innovative hypothesis for OLAP data examination.

1.2. OLAP Models

The OLAP [64] is developed for business data examination by performing the tough computations, trend examination, and complicated data modeling. The OLAP is composed of a wider grouping of business intellect, which in addition combines concerned database, details script and data clustering [96]. The OLAP tools facilitate customers to examine multidimensional information collectively from various possibilities [25].

The OLAP lies of 3 common models Relational Online Analytical Processing (ROLAP), Multi-dimensional Online Analytical Processing (MOLAP), and Hybrid Online Analytical Processing (HOLAP).

1.2.1. MOLAP Model

The MOLAP is a typical variety of OLAP [50] which is occasionally indicated as simply OLAP. It saves the information in multi-dimensional array space optimally, not in related tables. The efficiency of query processing strategy is enhanced by utilizing the index and cache in multi-dimensional space. Several compression methods are introduced for reducing the information size in MOLAP model to reduce the query processing time and cost (Figure 1.1).

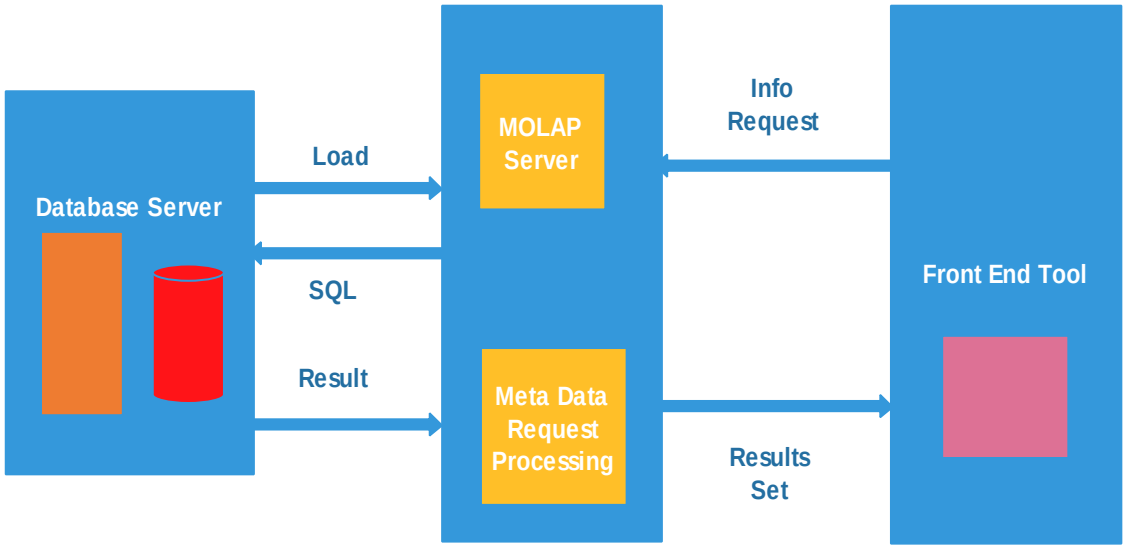

Figure 1.1. MOLAP Architecture

1.2.2. ROLAP Model

The ROLAP functions straight by means of relational information which is not involve any prior evaluations. The relational tables are used for saving the base and dimensional information. The recent tables are formed to seize the clan data stooping on expert schema architecture.

It remains on stacking the saved information in relational tables to provide the emergence of slicing and washing up processes of conventional OLAP [24]. It is measured in the direction of highly flexible in managing huge data warehouses, mainly dimensional models having extremely elevated cardinality (Figure 1.2).

The ROLAP design is provided several benefits: (i) it is simply incorporated into former previous relational data architectures and (ii) the relational information is saved perfectly and precisely as compare to multidimensional information [36].

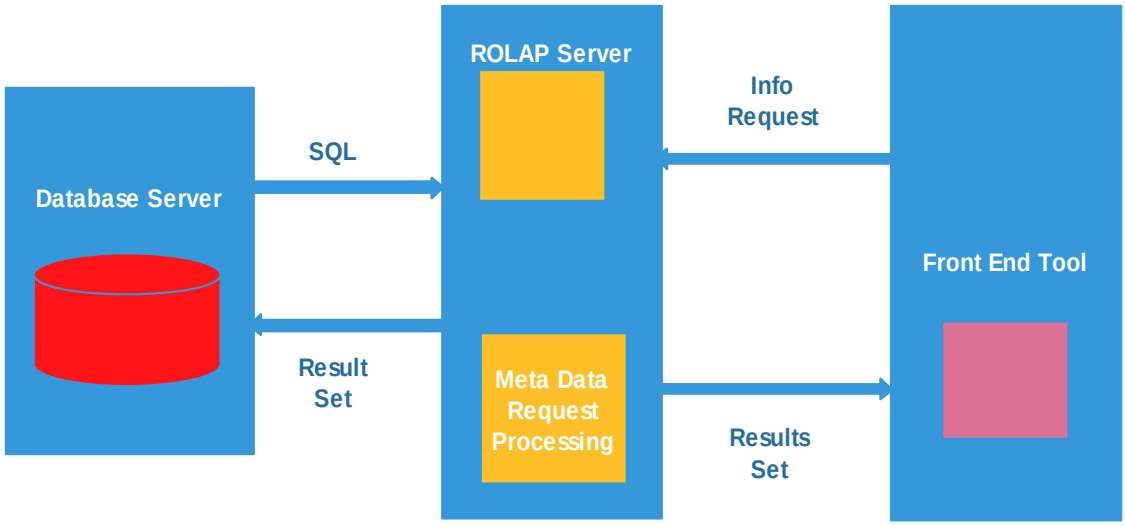

Figure 1.2. ROLAP Architecture

1.2.3. HOLAP Model

The detrimental transaction between extra price and deliberate query processing has checked that largely mercantile OLAP tool currently utilize a HOLAP model, which permits the model architecture to make a decision which fraction of information would be saved in MOLAP and which fraction in ROLAP [48] (Figure 1.3).

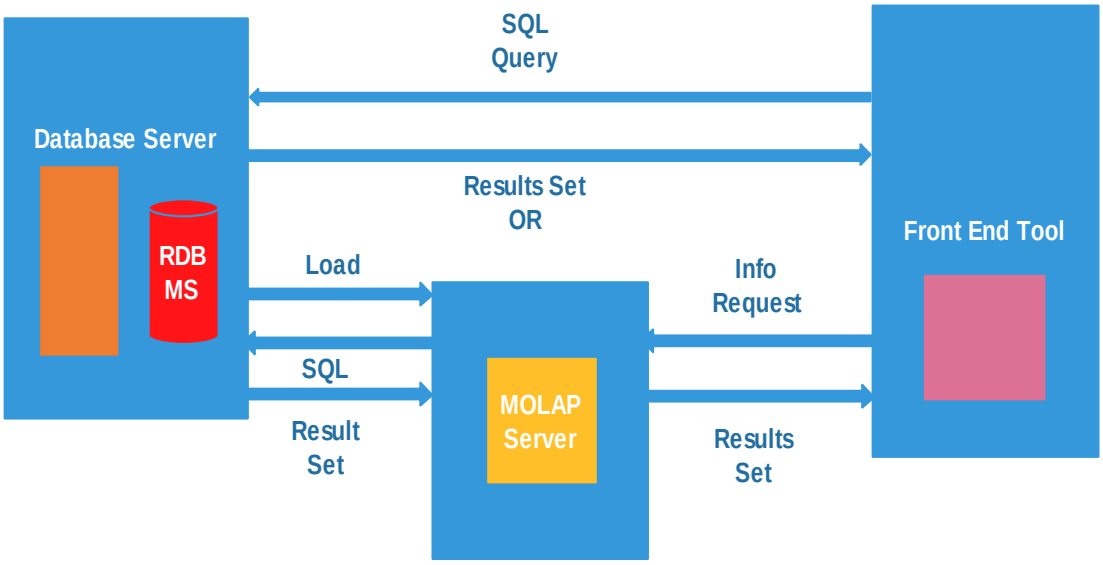

Figure 1.3. HOLAP Architecture

The data warehouse is an accumulation of enormous details of various data regarding numerous societies exploiting for appraisals. These appraisals are occupied through multifarious queries introduced to data warehouse and models for decreasing the reply instant. The OLAP models are exploited to process the query for data warehouse and excerpting the needful information for examination [9]. The query running cost and time is also major concerns for data models to illustrate the efficiency and power of OLAP. The business intelligence [88, 98] utilizes the intellectual properties, rules and laws for examining the personal and social information for solving the problems considering in future aspects. The data information is combined into groups (clusters) for fast accessing and query processing through OLAP models. All the types of OLAP models are achieved different level of query running speed through several optimization strategies and clustering methods [38, 97].

1.3. Data Warehouse

"Data Analytics" is a big buzz across the industries and in the research of computer science and IT domain [84, 86]. Data analytics is inevitable to incorporate business intelligence in the business model of all the organizations. This requirement leads to the notion of OLAP. Data warehouse is the most common practice to introduce the concept of OLAP.

Data warehouse allows the time-variant Online Transactional Processing (OLTP) data to be integrated from heterogeneous sources to define a fact (subject) for business processing [94]. This requirement leads to define suitable data models for efficient storage and access in terms of time and space. However, in analytical processing there are situations when the data models are designed and data are loaded in offline or batch mode. In these cases, even if the time requirement for loading data is high, the models could be acceptable if the user queries are answered from these models at run-time.

In data warehouse "data cube" or cuboids are used to represent the multi-dimensional data model. Each dimension contributes certain aspect of business [55]. The cuboids are controlled in the appearance of a lattice for representing all possible combinations of the associative dimensions of the given business problem. As the numbers of cuboids corresponding to a lattice is high, it is important to travel between different

cuboids efficiently. I will study a traversal mechanism and also observer abstract interpretation framework based on Galois connection. As I studied lattice, I identified a major drawback of this structure. Once a lattice is formed, newer dimensions are not allowed to be inserted [5, 76]. However in reality, business requirements frequently change. Thus to appropriately depict the business requirements it is required to insert new dimension at any place of lattice.

As data cube is important to represent data warehouse as a multi-dimensional model, concept hierarchy is important to represent a dimension in its exact abstraction. A dimension may be represented in more than one form and these representations (abstractions) often correspond to a lattice structure. Again, traversing within the lattice structure of concept hierarchy is important. I will study traversal mechanism among the different abstractions of concept hierarchy and formalized using abstract interpretation framework based on Galois connection [65, 87]. As the dimensions may have multiple abstractions, when it is integrated with lattice of cuboids the numbers of cuboids grow exponentially. If I consider physical memory organization, these huge numbers of cuboids reside at different types of memory elements having different access time or speed.

Hence, from computational aspect it is challenging to manage these high numbers of cuboids in different memory. We trying to propose a framework to consider lattice of cuboids by means of the theory structure of the dimensions in several storage components and the decision is taken to generate or traverse the target cuboid from the source cuboid dynamically based on a new algorithm [98].

Normal SQL query processing is not capable to handle this. Thus we used co-operative query language to frame the queries against the successful traversal on the concept hierarchy together in top-down and bottom-up comportment when necessary. Amalgamation of assorted data is a basic obligation for building any data warehouses [98].

In this context XML plays an important role as it is the most widely used language in web environment. However XML is semi-structured [100]; whereas the data warehouse tools are generally based on relational model which is structured. Therefore this

conversion is always challenging and has drawn researchers' attention for last few years. In a novel approach, we first take a single XML schema and convert the same into ER data model and relational model towards generating the data warehouse schema. However at this stage I could not able to identify star schema and snowflake schema. I will approach reverse engineering to acquire reverse XML representation as of information warehouse representation and validated the methods to proof correctness [101].

1.4. Online Transactional Processing

The Online Transactional Processing (OLTP) takes care of the transactional related issues in the database [94]. The Relational Database Management System (RDBMS) is the most popular way to represent the OLTP. SQL (Structured Query Language) is the language to implement OLTP (Figure 1.4).

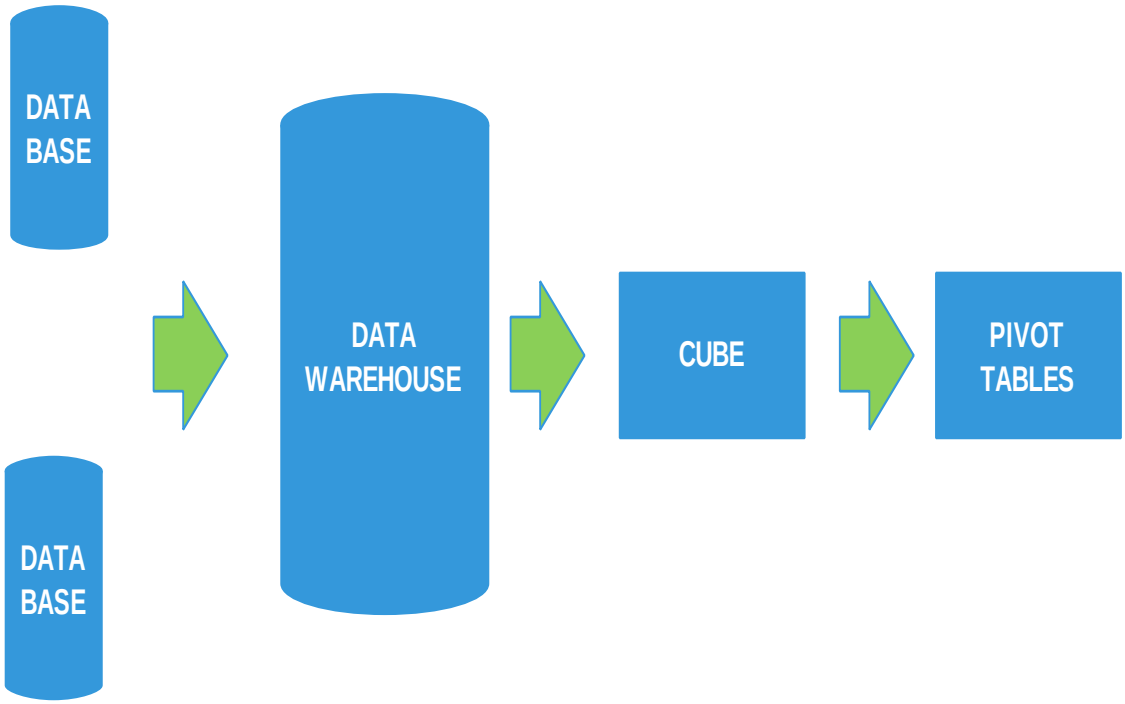

Figure 1.4. OLAP Analysis

RDBMS and OLTP are widely used by business organization, corporate, institutions etc. It is almost a standard to use RDBMS and SQL in database related applications. The data that are stored in OLTP is generally huge in size. However OLTP does not

support knowledge representation, decision making or business intelligence. But in business environment these things are hugely demanding. Hence there is a requirement of intelligent processing over data. Online Analytical Processing (OLAP) is the advancement over OLTP to incorporate the concept of analytical processing over the data to generate business intelligence, to represent knowledge or to allow decision making. Data warehousing technique is the most important OLAP tool [94].

Data warehouses [52] simplify and merge information in multidimensional field. The structure of data warehouses introduces information onslaught, information incorporation, and information alteration. Often the data warehouse tools are termed since Extraction Transformation-Loading (ETL) tools in the IT industry. Data warehousing could be represented like a significant pre-processing stage for information extraction. The data warehouse is elaborated like a category depended, incorporated, instance deviation, and nonvolatile assortment of information in sustain of management's assessment creation procedure.

1.5. Optimization Methods & its Components

The best elements are selected from existing set of elements and this selection procedure is known as optimization; which is introduced with the help of numerous conditions. It represents the optimum value of function in terms of minimum and maximum utilization for multiple objectives. It is strategy of generating the minimum and maximum assessment assortment as compared to obtain solution of problem [1].

An Optimization scheme is a plan, work or practice or of structure marvelous as entirely supreme, realistic or proficiency as expected. An optimization scheme evaluates least or highest values of function by critically taking inputs in prearranged variety and computing the function. The detailed optimization examination and schemes to the entire opposition introduces variety of practical arithmetic. On the whole, optimization contains obtaining "best reachable" standards of function contained by a variety of constraints [5].

The optimization has variety of major factors:

Objective Function: The different types of factors are merged to form an objective function for controlling the minimum and maximum values evaluation. These factors, which are known as objects, are not equivalent with each other; so few weight terms are introduced for equivalency [7].

Variables: the variables are defined as combinations of unknowns. The variables enclose the utility to demonstrate objective necessity and constraints [20]. Invented variables might not pull out erratically confirming distinct operative and other usefulness.

Constraints: few circumstances are called as constraints using for providing the convinced values devoid of others. Once the invented objectives, variables, and constraints are designed and merged to establish objective function [28].

1.6. Optimization Techniques

- **Ant Colony Optimization (ACO)**

The ACO [1] procedure is a casual exploration scheme depending on the perception of ant colonies and sited the beneath elements in datasets. Ants explore food and append to entire ants in approach in the course of pheromone spreading stuff on routes enthused. Ant colony depended data selection ensures entire regulations and situations and provides superior results against existing schemes depending on QoS factors. The exploration and exploitation steps are well defined and used in ACO for datasets.

- **Genetic Algorithm (GA)**

The common mixture strategy mechanism is introduced in GA [99] to suggest lowest or highest values of function having numerous objectives [6]. It balances the weights similarity among dataset elements through operators like mutation, selection and inversion; which is formulized through assortment, genetic practice and replacement evaluations. The genes are combined with chromosomes to generate the best elements weights among existing set of elements [7].

- **Ant Lion Optimization (ALO)**

The ALO is a meta-heuristic scheme of optimization to utilize the ant lion and its nature of hunting. An ant lion young insect obtains a narrowed curved space through disturbing alongside a spherical road in the sandpaper and dismissal the sand by means of its immense jaw. After quarry ensnare, young insect obscure at cone foundation and catches for trapping the ants in the crater. In earlier times, the ant lion knows that a quarry is wedged in ensnare, the ant lion blazes sand left from and shirking its quarry into the shaft. When a quarry is lodged into the jaw, the ant lion pulls the quarry toward itself and consumes. The numerical explanation of this scheme is introduced to explain various optimization problems [81].

- **Particle Swarm Optimization (PSO)**

The PSO [82, 85] is a meta-heuristic and energetic comprehensive examination and utilization practice of optimization. The PSO focused over slightest charge of calculation of a function examination. The global and local examination is competently applied by means of numerous calculating functions merging the several parameters of problem. The positions of particles alter at different time intervals through various functions. In the direction of procedure, the particle speed is altered at separate time intervals through positions. In next step, the positions of particles are yet again recomputed through speed. Therefore, the best particle locations are generated for next iterations.

1.7. Open Research Issues

During categorize to advance memory and reclamation of information cubes, the cuboid patterns are distributed while the commencement of the OLAP arrangement. Numerous researches of evaluation concerns of cube, collection, vast dimensionality and furthermore on the OLAP processing like roll-up, and drill-down [97]. An undemanding scheme is offered in [11] to differently evaluate every cuboid commencing the stand cuboid, utilizing typical set wise strategies.

Consequently the stand cuboid is comprehended and procedure in favor of every cuboid to evaluate. The static scheme forwards to pitiable concert together based on space and time complexity. An additional optimization strategy Amortize shown in [12] scopes at amortizing diskette studies through calculating since numerous set wise while probable, mutually in storage.

Suppose to believe a cuboid WXYZ. Whether the set wise of WXYZ is saved on disk, to diminish disk examine expenses of the cuboids WXY, WYZ, WXZ and XYZ are to be calculated in individual examine of WXYZ. Contribution classifies optimization strategy [12] is precise to the classification depended schemes and focuses at distribution arrangement price transversely several set wise [83]. At what time a hash-table is excessively huge to robust in storage, information is distributed. Consequent collection is finished for every section that locates in storage. This variety of contribution distribution optimization [16] is precise to the hash dependent schemes. Time-series information [23] frequently explains the ladder of numerous dimensional information. Still, the ladder and addictions among the cuboids is not inevitably on instant sequence unaccompanied [11]. The competence of roll-up and drill-down processing the information assortment saved in the relational model is measured [12].

1.8. Key Performance Factors

1.8.1. Intra-cluster distance

It is explained as the mean distance among data entities in identical cluster. It must have least value for optimized clustering [68].

1.8.2. Purity Index

It is illustrated the frequent clustering of data entities. It must have maximum value for optimized clustering [70, 71].

1.8.3. F-Measure

It is obtained from precision (prec) and recall (rcl) for data reclamation. It must have maximum value for optimized clustering [72, 73].

1.8.4. Standard Deviation

It is explained the data clustering strength about the mean standards. It must have least value for optimal clustering [102].

1.9. Motivation

The motivation of this study work is to frame a literature review for competence memory and admittance of data warehouse depending upon OLAP. During categorize to accomplish this target we will introduced data models, integrate existing and next label data models, strategies are considered to competently negotiate the architectures related to data warehouse, statistical analysis are performed to quantify the importance of attribute and also suitable formalism and validation are established towards the correctness of the proposed methods [99].

The problem definitions vary from the structures like lattice of cuboids to concept hierarchy; involve languages like XML, co-operative query language, SQL; models are considered across relational model, ER data model, data warehouse schema, XML schema and also to include the consideration of huge data of the data centric applications materialized views are generated. The methods and techniques used throughout the dissertation involve diversified approaches. In order to resolve the different problem definitions as defined in every phase; mathematical and statistical techniques, data modeling methods, formalism and validation approaches are used to achieve diversity [43].

The major goal of the analytical study is towards managing large data of OLAP applications to ease the storage and access. The approaches obviously try to achieve better space and time management. However in some of the cases of data warehouse applications time complexity is not always an inherent requirement during the construction of the data models and loading of the data [76].

This is due to the fact that construction of data models and loading of data into these models take place in offline mode. But the access to data during query processing should be online. Hence the major performance criteria of these data models are to

fetch or access the data at runtime and to achieve fair time complexity. The space complexity of these problems is also taken care of as less space complexity generally lead to less time complexity. We have also used different mathematical and statistical approaches like Galois connection, Abstract interpretation, Standard deviation, linear regression and nonlinear regression [32].

1.10. Problem Identification

At first I will analysis a mechanism to travel between any two cuboids contained by the lattice architecture to minimize the space and time requirement. In order to make the process faster we innovated a new numeric scheme to uniquely identify each dimension and cuboid. Tag number is used for dimension identification and Tagged value for cuboid. Roll-up and Drilldown operations are the major two operations on lattice of cuboids. A Galois link is discovered for these operations on lattice traversal. Roll-up and drill-down mechanisms are articulated in terms of abstraction and concretization respectively.

We also proved different mathematical properties of lattice of cuboids for the implementation in data warehouse. Thereafter, the traversal on lattice of cuboids is considered in term of existence of cuboids in physical memory. Try to find out the framework in this context computes the different access time of the various memory elements. Finally, in the end we analyzed the problems of static structure of lattice that prevents the insertion of new dimension.

In order to overcome this limitation, a novel algebraic structure will be structured. This allows insertion of any dimension in between level-1 to (N-1) where N is the numbers of dimensions of the initial lattice. A dimension could be added multiple times in different levels and to different cuboids if required. Hyper-lattice is actually a new algebraic model to replace lattice. We have established several lemmas and propositions of Hyper-lattice and also described its different properties.

We also proposed a new data warehouse schema using this model and termed it as Hyper-lattice schema. In order to efficiently travel different cuboids a traversal mechanism is also given to take the advantage of reduced time and space requirement.

Therefore we conclude that the main contribution of this chapter is efficient traversing the different cuboids of lattice and Hyper-lattice under different considerations and constraints and establishing Hyper-lattice as a new algebraic model alternative to lattice.

We find that existing research work on data cube is focusing on how to reduce the lattice structure in terms of cuboid numbers and size of data of each cuboid based on certain criteria. Our proposed algorithms will be traverse the lattice would also be applicable on this reduced structure. Abstract interpretation based framework can be applied on these for formal validation and correctness. Further our analysis on insertion the cuboids in separate storage components are applicable to other models of cuboids as well as in other computations where data blocks, searching of elements in various types of memory elements are taken into consideration. Finally, Hyper-lattice will be proposed as a new algebraic structure is applicable to different branches of science where the computations suffer due to the static nature of lattice.

1.11. Contribution of Research

The OLAP modeling is exploited to control and manage the huge data in shimmering generated distributed resources utilizing vast areas. A distributed system is superfluous suitable for applications to consider the huge heterogeneous data [85]. The variety of researchers have explored and examined the OLAP data modeling and data clustering schemes and obtained numerous fruitful results.

The present learning [45] illustrates a technique to utilize the historical information associating with construction presently and clan from IoT for detracting the future activities of the construction, even as representing the calculated values which are conscientious for negative construction presentation, devoid of guidance [10]. The OLAP and data mining [18, 76] are used healthcare data information taken from various heterogeneous sources.

The platform utilizes several rules and laws to calculate quality and processing speed of medical data by using Hadoop Map/Reduce interface [39, 47]. The real life datasets are used for modeling the sortilege model with business intelligence. A particular

methodical web entrance offering, which proposes concerted efficiency invigilating and assessment creation, is offered. The outputs present that the models confer extremely precise key performance indicators throws and give expensive presence into novel promising chances and problems [65].

A major amount of power is addicting in industrial building zone, ensuing in several undesirable concerns. A data cube model is introduced with relationship rule mining applying over industrial buildings power expenses survey dataset (6700 industrial buildings) to diminish the power expenses and enhance the power efficiency in industrial buildings. The OLAP is applied to power expenses industrial data to analyze the power based on atmospheric conditions, amalgamated equipment, construction types and cooling systems [25].

1.12. Research Objectives

Objective of this study is to analysis large amount of multidimensional data used in data ware house which help in business transactions, report generation to understand etc. The major goal of the analytical study is towards managing large data of OLAP applications to ease the storage and access. To achieve this goal my work plan is as follows:

1. First we have to understand the types of data deals by any data warehouse. How multidimensional data captured and stored in data ware house?
2. After understanding the types of data in data warehouse, we analyze the different type of schema used for it. With help of different secondary source we collect information regarding processing of different schema.
3. Third Importance aspect is how the online data is processing is carried out. We will study how data is efficiently process online?
   What is the different model used to processing task?
4. After all the lecture review, second year I will able to understand the limitation and advantages of different model used in OLAP.

5. In last in my research work we will try to propose a novel algebraic model which is not now an involvement about data depot other than moreover applicable to several branches of science which utilizes lattice.

1.13. Scope of the Work

In this research I will concentrate on the schemes that optimize operations during the OLAP model – in particular, the basic data models of data warehouse [33] such as cuboids lattice representations and theoretical structure. We emphasize on optimizing the access methodologies of these models. We also consider optimizations on query answering in a co-operative query language approach on concept hierarchies. Thereafter we consider the requirement of integrating XML schemas into data warehouse schemas.

Further we define formalism on this method by converting data warehouse schemas to XML schemas again. Finally we conclude with the materialized view construction which is helpful to reduce time complexity on any data centric applications. Statistical methodologies are used to build the materialized views and those methods are compared for performance evaluation.

I will approach reverse engineering to acquire reverse XML representation from data depot representation and validated the methods to proof correctness. Finally, I focus on query optimization by forming the materialized views for any data centric system like data base, data warehouse etc. We will the analysis at attribute level using statistical methods to measure the inter-attribute relationship. At first we propose a numeric scale to define the relationship among attributes, thereafter based on this knowledge we form materialized views. Initially, we use standard deviation to frame the idea of measuring inter-attribute relationship. Thereafter, we use linear regression to perform a better analysis and constructed materialized views. This study thus encapsulates four closely associated areas towards improving storage and retrieval of data for better analytics. These are exploring hierarchy of data using lattice of cuboids, concept hierarchy, building data warehouse from semi-structured XML storage and finally, creating materialized view for efficient query processing.

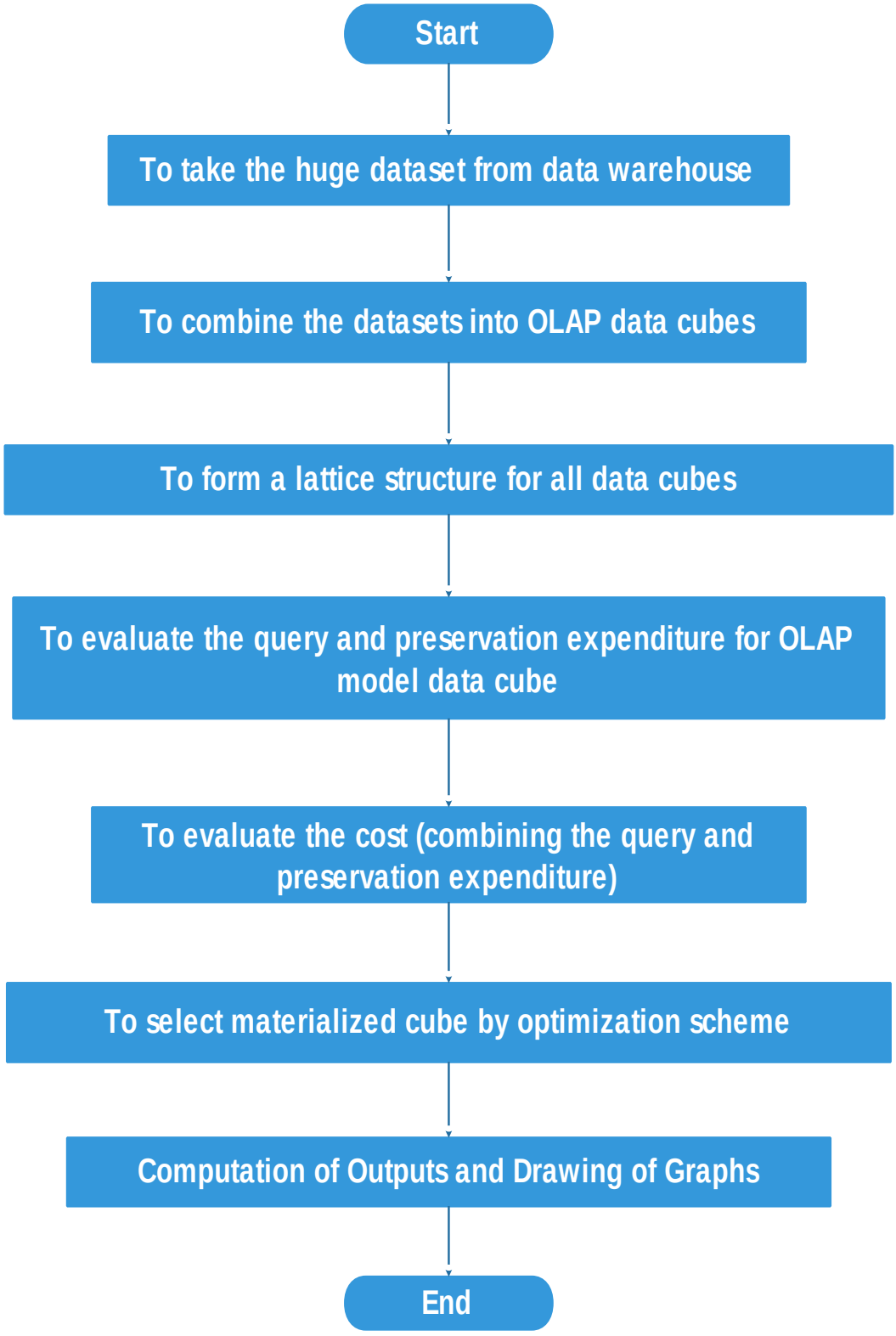

Figure 1.5: Flow Chart of OLAP Multidimensional Model Cube Selection

1.14. The Description of OLAP Multidimensional Model for Cube Selection

An Optimization scheme will be implemented to examine cube selection process for multidimensional OLAP data model displayed in figure 1.5.

Step1. To take the huge datasets from data warehouse.

Step2. To combine the datasets into OLAP data cubes.

Step3. To form a lattice structure for all data cubes.

Step4. To evaluate the query and preservation expenditure for OLAP model data cubes.

Step5. To evaluate the cost with the help of query and preservation expenditure for OLAP model.

Step6. To select materialized cubes with minimum cost by utilizing an Optimization scheme.

Step7. Outputs will be computed and Graphs will be drawn for analyzing the competence of this work against existing researcher`s work.

1.15. The Outcomes of This Research Work

The outcomes of the work will be:

1. We will understand the types of data deals by any data warehouse. The method of multidimensional data capturing and storing will be explored in data ware house?
2. We will analyze the different type of schema used for data warehouse with help of different secondary source.
3. The processing method of online data will be examined. We will study about online processing methods and models.
4. The limitation and advantages of different model used in OLAP will be analyzed.
5. An Optimization Technique for data cube selection will be developed.
6. The competence of this work will be examined against existing works.

1.16. Organization of Thesis

There is nine chapters, which are described the whole research work. Remaining chapters have been illustrated as beneath:

Chapter 2: Review of plentiful OLAP data model cube selection and data clustering are described in this chapter and popular factors such as energy expenses, cost, time complexity, response time and delay have been examined. A lot of characteristics such as implementation tools, applied schemes, pros and cons for OLAP model analysis have been illustrated. The plenty of OLAP model examination schemes, cube selection strategies and data clustering approaches are represented in table to demonstrate the necessity of OLAP data models in numerous application era.

Chapter 3: in this chapter, a Fruit Fly Optimization (FFO) scheme is illustrated for OLAP materialized cube selection in optimal mode. The entire work has been performed into two levels of development of function with several objectives and performing of fruit fly optimization over function. The advanced competence of FFO has been calculated in terms of query processing expenditure. The assessment of competence of FFO has been evaluated in a plenty of situations.

Chapter 4: in this chapter, a Grey Wolf Optimization (GWO) scheme is illustrated for OLAP materialized cube selection in optimal mode. The entire work has been performed into two levels of development of function with several objectives and performing of grey wolf optimization over function. The advanced competence of GWO has been calculated in terms of query dispensation expenses. The assessment of competence of GWO has been evaluated in a plenty of situations.

Chapter 5: Comparative assessment of competence of GWO and FFO are demonstrated. The advanced competence of GWO and FFO has been calculated in terms of query dispensation expenses. The assessment of competence of GWO and FFO has been evaluated in a plenty of situations.

Chapter 6: in this chapter, a Dragon Fly Optimization based Clustering (DFOC) scheme is illustrated for clustering of huge datasets in optimal mode. The entire work

has been performed through dragon fly optimization which is applied over a function with several objectives. The sophisticated ability of DFOC has been calculated in terms of intra-cluster distance, purity index, F-measure and standard deviation. The assessment of ability of DFOC has been evaluated in a plenty of situations.

Chapter 7: in this chapter, a KMeans-Salp Swarm Optimization based Clustering (K-SSOC) scheme is illustrated for clustering of huge datasets in optimal mode. The entire work has been performed through K-Means and salp swarm optimization which are applied over a function with several objectives. The sophisticated ability of K-SSOC has been calculated in terms of intra-cluster distance, purity index, F-measure, standard deviation and running time complexity. The assessment of ability of K-SSOC has been evaluated in a plenty of situations.

Chapter 8: Comparative estimation of ability of DFOC and K-SSOC are demonstrated. The sophisticated ability of DFOC and K-SSOC has been calculated in terms of intra-cluster distance, purity index, F-measure, standard deviation and running time complexity. The assessment of ability of DFOC and K-SSOC has been evaluated in a plenty of situations.

Chapter 9: the conclusion of complete work is demonstrated in this chapter and explained future concerns of this research work.

1.17. Summary and Discussion

In this chapter, OLAP model and OLAP processing schemes have been demonstrated in brief. OLAP modelling and variety of optimization schemes have been discussed. This chapter introduces the ability of a plenty of OLAP processing mechanisms. In a while the computation of optimal OLAP processing mechanisms, the estimation of schemes motivates us to be recognizable by means of the limitations, pros and cons in OLAP model processing and optimization schemes.

This chapter demonstrated vast active OLAP mechanisms like fact model, PSO, cube algebra and cost model etc. and coalition in the course of them. Widespread estimation of numerous OLAP data modelling, data processing and optimization schemes reliant

on used ability terms is notated in table. Subsequently processing of queries over OLAP data model becomes graceful. The OLAP model processing mechanisms are oppressed lonely for data dispensation and optimization schemes are oppressed for mixture of optimal selectors in existing works.

Afterwards the inspection of optimal OLAP model processing mechanisms, a variety of open research issues demonstrates the fervour of the research ability in OLAP data model selection mechanisms. The delineation of this chapter moreover demonstrates a plenty of objectives of this work which is inferior illustrated reliant on block diagram elucidation the OLAP model and optimization scheme. The planned research work and outcomes are demonstrated in brief.

CHAPTER-2
STATE OF THE ART: REVIEW

2.1. Introduction

At the moment, an examination of vast materialized cube selection strategies and data clustering schemes [92, 93] for OLAP data processing methods has been performed, united issues and utilized parameters such as query processing expenditure, purity index, throughput, F-measure, space complexity, time complexity, intra-cluster distance and standard deviation have explained and vast optimization based data clustering and cube selection strategies have been illustrated. The predictability of flexibility in data clustering, cube selection and optimization schemes have explained in the table architecture.

2.2. Literature Review

The trajectory information permits the revise of nature of affecting elements, from person to creatures by utilizing the wireless transmission, portable machines and techniques. The key concern of the spatial database management system is saving and processing the huge volume of informative data through OLAP models. The indexing, saving and extracting the spatial information from big data [66] warehouses is performed by using various data management techniques. The problems arising in trajectory database and the solutions of problems in big data analysis are combined in tabular format for explaining the use of several OLAP data management techniques to improve the quality of decision making [22, 91].

The contemporary day progression is gradually more digitizing the real life scenarios with quick enhancement of information. The novel and important knowledgeable information are extracted from data warehouses with the help of multidimensional data models. Although the Hadoop architecture is well defined policy to deal with big datasets and it has several computing architectures for multiple application fields. This creates a strategy for dividing the big data into several groups over cloud computing

platform. The key concerns and challenges of big data analysis are identified and solutions are proposed for data analysis over cloud environment [89].

The organization of COVID-19 epidemic shows various extraordinary issues in numerous areas from drug to biology, which might advantage from examination techniques capable to amalgamate the enhancing existing COVID-19 and associated information such as effluence, demographics, and weather. On the basis of this information, a COVID-WAREHOUSE is developed for saving and processing the COVID-19 data of effluence and weather. The time and environment position are two major factors of dimensional fact model of OLAP examination [27].

A console application is anticipated and urbanized to proceed as an identical twin which can be notated the precise significance of sharing responsibility for few future crashes. The present learning [51] illustrates a technique to utilize the historical information associating with construction presently and clan from IoT [34] for detracting the future activities of the construction, even as representing the calculated values which are conscientious for negative construction presentation, devoid of guidance. This console application is implemented in java language and key performance indicators are used to visualize the OLAP model for verification and revelation purpose [10].

The precise feeling is the major concern to anatomical fitness sensing of subversive petroleum mines, although utilizing nerve Bragg strident sensors. On the other hand, the earlier urbanized machines for architectural invigilating of subversive mines contain restricted to invigilating devoid of some strength of harm exposure [35]. Consequently, this paper incorporates an extremely precise invigilating machine on an Internet of things (IoT) framework over Web 2.0 server [61]. The prime element examination, besides hierarchical clustering, is utilized to locate the scratch pointer of the mine. The scratch manifestation is verified, representing the least value for rigidity diminution. Therefore, incorporation of this method with Internet would be efficiently introduced for premature security measurement of subversive petroleum mines and data contribution in real time [15].

The insidious medical services are recognized several components like internet, ad-hoc environment, and transmission techniques to provide better solutions of challenges in medical system. The health information is taken from the IoT devices and utilized with the help of machines and OLAP models. The OLAP and data mining [52, 53] are used healthcare data information taken from various heterogeneous sources. The platform utilizes several rules and laws to calculate quality and processing speed of medical data by using Hadoop Map/Reduce interface [39].

The solar energy data are generated from sensors introduced in various geographic positions and systems of weather organizations. Still, the Portable Document Format (PDF) and Hyper Text Mark-up Language (HTML) formats of files are not given meaningful solar energy data extracting from various resources. So, a query platform is developed with solar data processing, where the data is taken and combined from various heterogeneous resources [40].

At present era the industry environment demands contribute sequences to be as compared to theoretical, which need a novel analysis method for data mining sortilege. A sortilege model is provided the combination of progression, routine and data mining models. This model is also calibrated with the business intelligence to evaluate the efficiency and performance through rules and key performance indicators. The real life datasets are used for modelling the sortilege model with business intelligence. A particular methodical web entrance offering [31], which proposes concerted efficiency invigilating and assessment creation, is offered. The outputs present that the models confer extremely precise key performance indicators throws and give expensive presence into novel promising chances and problems [65].

The model having relational information is possibly the maximum habitually utilized database structure [32]; still, complex queries for huge dataset are not preferably run and examined by relational model. The OLAP concept is introduced to evaluate the multidimensional data for online processing and examination. The business intelligence has grown with improving the OLAP model facilitating the efficient data cube selection. The ROLAP and MOLAP have improved their performance against query processing time and cost to store the reporting time of working hours over economical

data. The outcomes represent the better quality performance of ROLAP as compared to OLAP for performing data cubing [75].

The telecom organizations have to inflate the services with cheapest price on the basis of customer requirement information along with call detail record and behaviour of purchase. This model is developed and designed abstractly, reasonably and physically to solve the problems of data mart in sufficient time. The OLAP model is developed to provide superior performance nearer to customers purchase nature and enhance the marketing of goods [23].

The conceptual models [11] are major concerns in OLAP models using the data warehouse, which improves the use of logical models for better performance. The conceptual methods have several limitations like maximum learning arc, not easily understandable and flexible in user friendly environment. These techniques are more complex to study as well as analysis for knowledge workers. Then the cube algebra is introduced as a conceptual structural model providing the maximum level of knowledge about the database and OLAP to workers. The undesired information is hidden from the unknown users for security purpose [16].

The OLAP is utilized for multidimensional data representation and processing for various application areas. The cube presentation model is presented and examined over unified modular language to show the data cube more precisely and accurately. Hence needful data are extracted for different organizational systems and represented through data cubes using extensible mark-up language and unified modular language to enhance the visualization strength of information of multidimensional data cubes [12].

A major amount of power is addicting in industrial building zone, ensuing in several undesirable concerns. A data cube model is introduced with relationship rule mining applying over industrial buildings power expenses survey dataset (6700 industrial buildings) to diminish the power expenses and enhance the power efficiency in industrial buildings. The OLAP is applied to power expenses industrial data to analyze the power based on atmospheric conditions, amalgamated equipment, construction types and cooling systems [25].

The query running costs and time is decreased by using an optimal group of materialized data cubes in the data warehouse. The Particle Swarm Optimization (PSO) [41] is well suitable algorithm utilizing for optimal selection of the data cubes. The speed of PSO is higher as compare to other greedy and heuristic techniques. The global optimal results are also achieved by the PSO in terms of materialized data cube selection. The PSO is applied to a collection of data cubes to find out the best cubes to reduce the query running time cost. The exploration and exploitation power of PSO is superior for searching the local and global optimum values of data cubes enhancing the accuracy and performance of the system. The results represent the better quality efficiency of PSO [55] against the optimization technique like Genetic Algorithm (GA) based on multiple performance factors [8].

The data clustering [56, 62] is introduced to examine the numerous data, where the data are divided into multiple partitions for further processing. Therefore, the data is easily accessible in least time for users to save the extra cost of query processing in data warehouse. The K-Means are a famous method to partition the data into clusters for data analysis. Several optimization algorithms [42, 54] are also introduced for data clustering to generate optimal clusters of data for reducing the computational cost and time over huge data information of data warehouses. The PSO is one of the best utilized approaches to improve the strength of data clustering with least error rate and highest convergence speed as compared to other clustering techniques [37, 43].

Table 2.1 represents the various works of researchers in comparative way.

Table 2.1: Comparative Analysis of OLAP based Research Works

Authors	**Approach/Method**	**Application Area**	**Platform**	**Advantages**	**Limitations**
G. Agapito et. al. [27] (2020)	Dimensional Fact Model & OLAP	Medical (COVID-19)	Python	Used automatic extraction, transformation and loading	Not used graphical user interface

A. Papacharalampopoulos et. al. [10] (2020)	Key performance indicator & OLAP	Production System	Java	Used the IoT to indicate the future failures	Not used differential order and knowledge based libraries
B. W. Jo et. al. [15] (2018)	Fiber Bragg Grating	Coal Mines	Web 2.0	Hierarchical clustering utilized for harm pointer of mine	Subversive mines unsympathetic situation
J. N. S. Rubi et. al. [39] (2019)	Internet of medical things	Healthcare	e-Health Sensor Kit	Automatic preparation of data and knowledge extraction method	Throughput, and average time is not evaluated
J. L. S. Carvantes et. al. [40] (2016)	Solar radiation extraction and query platform	Weather Stations	Sensor Web Enablement	Reuse the data and develop the web application	Sensors not used like Thermometer, hydrometer etc.
N. Stefanovic [65] (2014)	Supply Chain Model & Key performance indicator	Business Intelligence	Web Portal	Global, Collaborative, predictive analysis	Not support visual intelligence.
P.	ROLAP &	Business	Structure	Server side	Time

Westerlund [75] (2008)	MOLAP	Intelligence	d Query Language (SQL)	data analysis	Consuming and costly
D. Camilovic et. al. [23] (2009)	OLAP	Data Mart	SQL	Dynamic data processing	Not provide any time and cost model
C. Ciferri et. al. [16] (2012)	Cube algebra model & OLAP	Pollution Control	SQL	User friendly model with flexible design	Not include spatial and multimedia data
A. S. Maniatis [12] (2005)	Cube presentation model and OLAP	Rational Rose	XML and UML	Stereotype extension of data	Not provide visualization and automatic generation
B. Noh et. al. [25] (2019)	Data cube model with relationship rule mining	Industrial building data	R Tools	Evaluated the power strength	The time and harmful effects not considered
A. Gosain et. al. [8] (2016)	PSO	Sales data	MATLAB Tool	Several frequencies and dimensions are taken	Processing time is not calculated

Data Warehouse [9] is a collection of huge information of assorted data about several organizations utilizing for assessments. These assessments are taken by complex

queries applied to data warehouse [27] to reduce the answer time. The materialization in data cubes is well utilized for query processing in an efficient manner with lesser time consuming. Entire views of data cubes can be materialized to access the data quickly with minimizing the answering time of queries. The Online Analytical Processing (OLAP) [64] is used for query processing over data warehouse and extracting the useful information for analysis work. It is also helpful for assessment support to provide the users overviews. OLAP is combined with Key Performance Indicators (KPIs) [10, 65] to provide a dashboard application over historical data. The production efficiency is explained with numerical examples and developed in the Java programming language.

OLAP is also utilized for energy cost analysis in commercial areas to reduce the expenditure with improving the power efficiency [14]. A multidimensional cube model is utilized for evaluating the power consumption at several stages of generalization. This time OLAP [22] is used with association rules to generate feasible solutions for huge data about buildings in commercial sectors. Unified Modeling Language (UML) [16] and Structured Query Language (SQL) are also introduced for query processing in OLAP. The web based software is implemented for analyzing the students' results based on the object oriented methodology. The analytical comparison of huge student data is easily performed by OLAP to minimize the stress and workload of schools. Lectures are easily delivered to students through this object oriented platform [25]. Another application of OLAP is a police intellect, a decision scheme to catch the criminals and take an efficient decision about crime [48].

The analytical process is also utilized for intellectual study over multidimensional data for business purposes [88] in Social Business Intelligence (SBI). The social, private and public data can be analyzed easily using OLAP semantic analysis to model the huge information of companies for business point of views [36, 98]. The health care data [97] is also formulated and structured by mining in OLAP. The Internet of Medical Things (IoMT) systems are well suited for providing health information of patients and also predict the treatment of diseases by using medical data analysis [39]. The extraction of solar radiation data from huge amount of information is performed by query processing in several geographical positions and structures. The sensor data of

this system accesses by user to initiating the query on solar data and analyzing the data at several stages [40].

The numerous amounts of data and information of different industries are combined to form a data warehouse [9]; which is further exploited for processing and managing the information. These evaluations are occupied through multifarious queries functioned on data warehouse [27] to decrease the resolving period. The embodiment in data cubes is specially developed for query answering in a competent mode with smaller time intense. The whole data cubes sight could be embodied to penetrate the information hastily with decreasing the answering period of queries or questions [37].

The OLAP [64, 89] is exploited for query dealing through data depot and fetching the needful data to examine effort. It is furthermore cooperative for appraisal carry for offering the client aspects. The historical information is well recognized with a dashboard scheme to improve the key performance indicators [23, 65] of OLAP database. This scheme is implemented in Java and evaluated with mathematically exploited to enhance the production effectiveness.

The power consumption in industrial fields is very interesting concept to perform better examination of energy for OLAP query initialization [10, 11]. The energy using for OLAP query evaluation is exploited through data cube at numerous steps of simplification. Here, OLAP [14, 22] is developed with associated regulations to achieve practicable results for vast information concerning buildings in industrial fields. The web related applications are using the structured query language for modifying the database and also performing the efficient query processing with the help of the unified modelling language [75]. The object oriented paradigm based database is well suitable for preparing the students performance based results. This mathematical examination of vast student database is performed by OLAP model; which helps to examiners and institutes for reducing the overhead and tension to manage this large data. The object based architecture is simple to implement for classes delivering to students in flexible and efficient manner [16]. The cyber security related data is also utilized for taking the decisions against suspicious activities to reduce the crime rate and catch the suspicious persons [25].

The social business intelligence is a specific strategy to fulfil the industrial requirements for analyzing the multidimensional information intelligently [36, 48]. The communal, confidential and secret information could be examined simply exploiting OLAP semantic study to form the vast data of organizations for industrial behaviour [40]. The medical information [98] is furthermore processed and maintained by information mining in OLAP; which is collected from various internets of things devices. These devices are specifically appropriated for obtaining medical data of patients and furthermore expect the remedy of disease by utilizing health information exploitation [97]. The mining of solar radiation information from vast data warehouses is initiated by query dispensation in numerous environmental locations and architectures. The sensor information of solar data executes by persons to apply the query for requesting and examining the information at various levels [39, 88].

The huge amount of data is collected in the form of data warehouse [9, 27, 64] to combine all the information about organisations. This data information is very difficult to access in minimum time due the big data [57] for OLAP. to improve the performance of OLAP, the data is organised in several groups to save the accessing time and query processing cost. This organisation of data into groups is known as data clustering. KPI (Key Performance Indicator) [10] is also merged with OLAP [65] to perform fast query processing.

OLAP is also used for power cost examination in marketable areas to diminish the expenses with increasing the influence performance [14, 22]. A multidimensional data is used for calculating the influence expenses at various stages of simplification. Hence the rule association is developed with OLAP to obtain efficient results on numerous building data using UML (Unified Modeling Language) [16] and SQL (Structured Query Language) [25, 48]. The decision support system is also developed with data clustering for fast accessing the huge data [36, 88] with maximum accuracy of information with respect to future aspects [97, 98].

The data warehouse [9, 27, 64] is stored the numerous amount of structured and unstructured data [46, 49] and information on various application areas over network. This huge collection of data is very complicated to access and use for different OLAP

applications because of its size and its higher data accessing and processing time. The OLAP [23, 65] model is combined with KPI [10] (Key Performance Indicator) to achieve a quick processing time of query. The Multidimensional Aggregation Cube (MAC) scenario is also used by OLAP for selection of cubes to reduce the complexity of huge data processing [11].

The organizing of data is well performed by using clustering, in which data are maintained in various clusters (groups), so suitable data information is easily accessed and processed in minimum time duration and least cost. The analysis of energy cost is performed by using OLAP in the industry and market fields to moderate the expenditure with escalating the influence concert [14, 22]. The influence expenditure is evaluated by utilizing multidimensional data [75] at several generalization steps. The UML (Unified Modelling Language) [16] and SQL (Structured Query Language) [25, 48] are applied over huge construction data to generate effective outputs with the OLAP rule association. The data clustering is also combined with decision support structure to provide speedy data accessing [36, 88] with higher precision of data in terms of future aspects [37, 97].

2.3. Summary and Discussion

The reliable OLAP models are employed to analyze and deal with important informative data of data warehouse. The overhead of handling the huge data and making the decision for data processing is diminished through OLAP models. The ROLAP, MOLAP and HOLAP are three popular models for data warehouse to explain and organize the data into multidimensional, relational and hybrid structures. The query processing expenses and time are abridged by introducing various methods and optimization techniques with multidimensional data model to enhance the OLAP performance. The efficiency of OLAP model is also enhanced by utilizing the data clustering methodologies [58, 59], in which the important huge data information such as medical, industrial and secret data is combined into clusters. So, data clustering is improving the accuracy and accessing capability of OLAP models over data information. The key concern of this research work is specified a widespread compilation of precious data about OLAP examination, models, query performing,

problems and optimization techniques in multiple eras. Therefore, the researchers can be used the necessary information to develop novel and proficient strategies for OLAP data examination. The text and tabular arrangements of previous works are further utilized for innovative OLAP strategy selection.

Consequently, it concludes that data clustering, cube selection and optimization strategies are developed for optimal OLAP data processing but they enclose nevertheless numerous restrictions, then the data cube selection and data clustering with optimization strategies have been projected in subsequent chapter to broaden the usefulness and to moderate the restrictions of prior strategies.

CHAPTER-3
Selection of OLAP Materialized Cube by using a Fruit Fly Optimization (FFO) Approach: a Multidimensional Data Model

3.1. Introduction

The Online Analytical Processing (OLAP) based Multidimensional examination hassles for several stockpiling magnificence over huge data. For as much to recognize queries answering time companionable by OLAP framework users and understanding entire business perceive mandatory, OLAP data is structured as a data cube (a multidimensional model). The OLAP queries are responded in speedy and steady time by utilizing the cube materialization for assessments takers. But, this also involves unendurable expenses, regarding to stockpile memory and period, and as a data depot, OLAP has an average dimension and dimensionality which is to be significant on query processing.

Consequently, cube assortment has got to be finished motivating to diminish inquiry management expenses, maintaining as a restraint the materializing gap. Several techniques and heuristics like deviationist and insatiable algorithms have been utilized to offer an estimated result. In this work, a Fruit Fly Optimization (FFO) approach is implemented in a lattice structure [13] to obtain an optimal materialized data cube for reducing the query processing expenses. The results illustrate that FFO generates better performance than Particle Swarm Optimization (PSO) [44, 60] in terms of frequency and number of dimensions.

In the above analysis of works, materialized data cubes are generated from important data extraction from the huge amount of data. These data cubes are selected by using optimization approaches in large space to provide efficient query processing in an OLAP multidimensional data model. The PSO is one of the nature inspired optimization approach initiating on data cubes to generate optimal materialized data cubes. Here, we proposed another optimization technique FFO, which generates better results than PSO

over OLAP multidimensional model in terms of frequency and number of dimensions to reduce the query processing expenditure in several constraint spaces.

3.2. The Fruit Fly Optimization (FFO) approach for selection of OLAP Materialized Cube (a Multidimensional Data Model)

3.2.1. FFO Approach

A bio-inspired Fruit Fly Optimization (FFO) approach is utilized for generating the global optimal solutions on the basis of foraging inspired by fruit flies. Various realistic explanations to optimization quandary are illustrated by foraging of fruit flies in FFO. The fruit fly gaits by plunging to the food, exploits its vigilant spirit to realize food and where it correlates flock and then it gaits by plunging into a route.

Start

Step1. Put primary standards of position, fitness function, smell, generation and population randomly of entire fruit flies. (eq. (1))

$$X_{\emptyset} = X_{axis\_value} + random_{value} \quad \& \quad Y_{\emptyset} = Y_{axis\_value} + random_{value} \tag{1}$$

Step2. Compute the fitness standards on the basis of distance (***Dist***) and smell (***Sml***) for entire fruit flies. Hence, optimized solutions are obtained with fitness of individual and population. (eq. (2), (3) & (4))

$$Dist = \sqrt{X_{\emptyset}{}^2 + Y_{\emptyset}{}^2} \tag{2}$$

$$Sml = \frac{1}{Dist} \tag{3}$$

$$Fitness_{Function} = Function(sml) \tag{4}$$

Step3. Change the standards of best index and position for entire fruit flies (eq. (5) & (6)). Hence, update the fruit fly`s position (eq. (7) & (8)).

$$BestIndex(t+1) = \mu \times BestIndex(t) + \eta_1 \times Rand \times \left(Dist - X_{\emptyset}(t)\right) + \eta_2 \times Rand \times \left(Sml - X_{\emptyset}(t)\right) \quad (5)$$

Where $\boldsymbol{\mu}$ = indolence coefficient

$\boldsymbol{\eta_1}$ and $\boldsymbol{\eta_2}$ = Constant.

$$[BestSmell \;\; BestIndex] = minimum(Sml) \quad (6)$$

$$X_{axis\_value} = X(BestIndex) \quad (7)$$

$$X_{\emptyset} = Round\Big(X_{\emptyset} \times (EntireNodes - EntireSourceNodes - EntireDestinationNodes)\Big) + EntireSourceNodes + 1 \quad (8)$$

Step4. Establish optimized result, if not, go over 2.

If established the optimized result, found best node.

End

3.2.2. Lattice Structure

The lattice structure combines entire probable data cubes at dissimilar stages of concentricity by describing cubes on their reliance. Two cubes C_x and C_y are connected by a route for generating reliance association ($C_x \leq C_y$). This association illustrates that an answer to the OLAP query given by C_x, can also be provided by C_y. In lattice structure, 2^Ndata cubes are possible for N dimensions for an information association. Here, 2^3 = 8 data cubes are obtained from 3 dimensional data “Sales” (Customer (C), Supplier (S), Part (P)) (Figure 3.1).

The minimum concentricity represents by bottom cube (eg-CSP) and maximum concentricity represented by the top cube. The structure illustrates that an answer to the OLAP query given by child cube (*,S,*) can also be provided by any parent cubes (*,S,P), (C,S,*) or (C,S,P) by concluding data beside few dimensions.

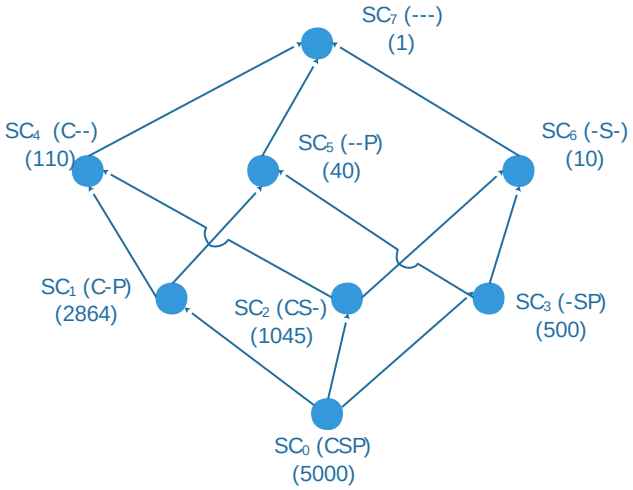

Figure 3.1. Lattice Structure

3.2.3. Cube Selection using FFO approach

The objective of an OLAP model is to reduce expenditure of query and preservation with fulfilling a constriction like materialized space. The Entire Query Expenditure (EQE) for an OLAP query is evaluated by utilizing answering cost and frequency of query using eq. (9).

$$EQE = \sum_{x=1}^{N} fq_x \times E(q_x, C_M) \qquad (9)$$

Here,

fq_x= query q_x frequency.

$E(q_x, C_M)$ = query q_x answering expenditure (cost) with C_M (Materialized Cube) which is calculated by using eq. (10).

$$E(q_x, C_M) = Minimum\left(|SC_x|, Lpre(q_x, C_M)\right) \qquad (10)$$

Here,

$|SC_x|$ = Number of sub-cubes of query q_x.

$Lpre(q_x, C_M)$ =Least Predecessor (Lpre) conception of query q_x with C_M.

The Preservation Expenditure (PE) of a materialized cube $C_x \in C_M$ is evaluated in terms of least predecessor (Lpre) of C_x using eq. (11).

$$PE(C_x, C_M) = |Lpre(C_x, C_M)| \tag{11}$$

After that Entire Preservation Expenditure (EPE) is evaluated by eq. (12).

$$EPE = f_{PE} \sum_{C \in C_M} PE(C, C_M) \tag{12}$$

Here,

f_{PE}= frequency of inclusion in support association.

Hence, the Entire Cost Function (ECF) is calculated by combining the EQE (Entire Query Expenditure) and EPE (Entire Preservation Expenditure) using eq. (13).

$$Minimum \qquad ECF = \sum_{x=1}^{N} fq_x \times E(q_x, C_M) + f_{PE} \sum_{C \in C_M} PE(C, C_M) \tag{13}$$

The optimal value of ECF is obtained by applying FFO approach on this objective function ECF. The initial values of expenditure can be evaluated from pedestal association and the value of C_M **Error! Bookmark not defined.** is empty initially. After that optimal values are evaluated using FFO. The ECF function is utilized as variation function, as it discovers the solution fitness throughout the following of the several targets. The ECF function is minimized to obtain least expenditure of query with highest fitness.

3.3. Result and Analysis

The FFO and PSO approaches have implemented in MATLAB 2019a environment (windows 8, 8 GB RAM, Core i3 processor) and analyzed in terms of frequency and number of dimensions (Figure 3.2 (a) (b) (c)). The results illustrate the best quality performance of FFO over PSO with minimum query processing expenditure. Here, we

measured several belongings of space constrictions as 10%, 20%, 30%, 40%, 50% and 60% to analyze the FFO and PSO approaches.

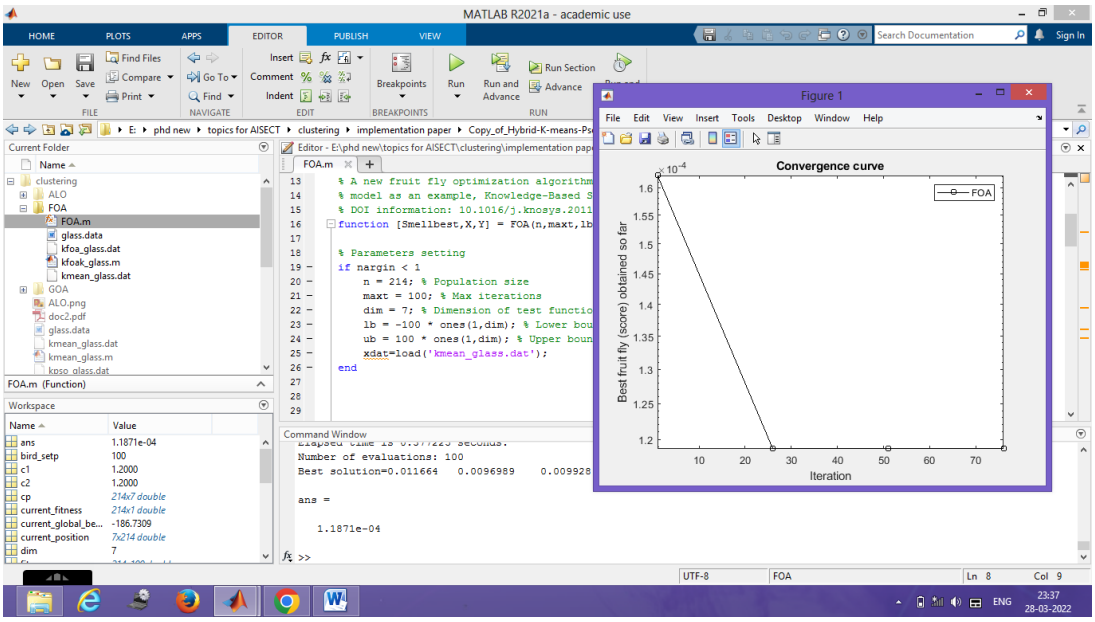

Figure 3.2 (a). Implementation in MATLAB

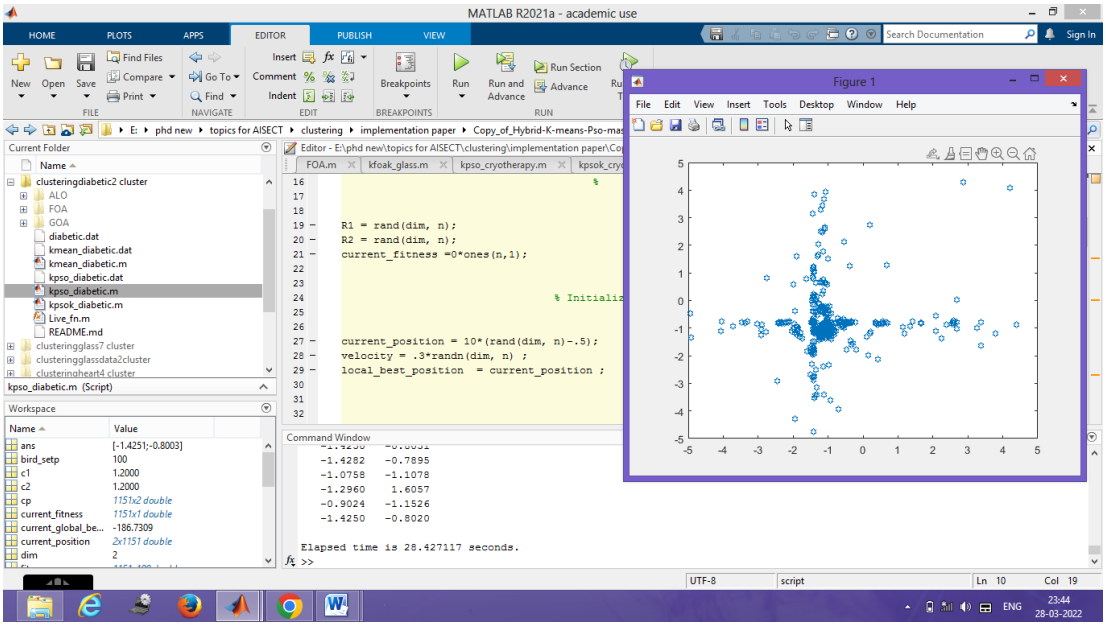

Figure 3.2 (b). Implementation in MATLAB

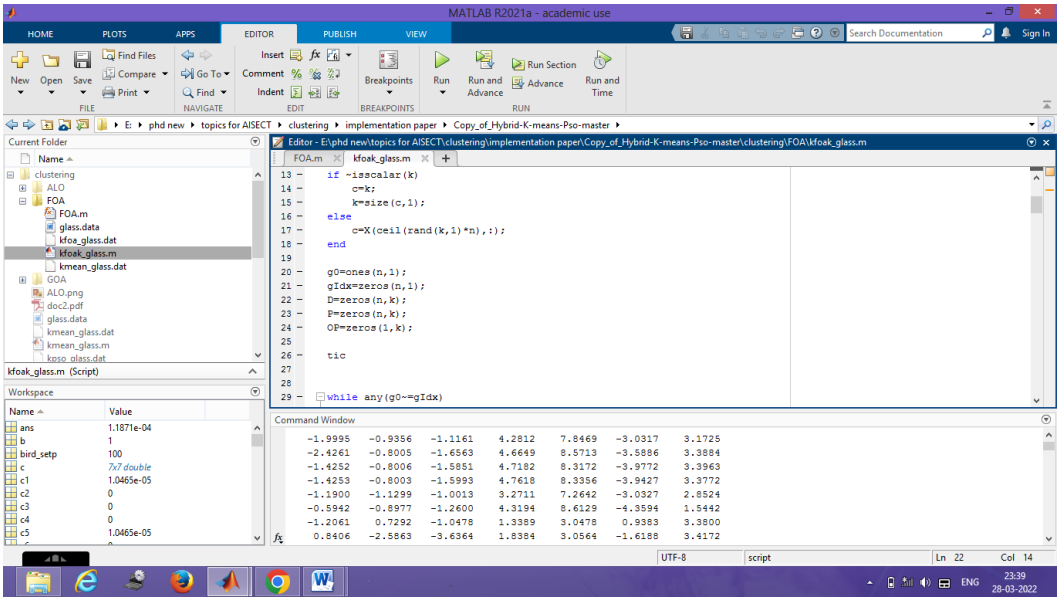

Figure 3.2 (c). Implementation in MATLAB

We performed the implementation of FFO and PSO on the multiple dimensional data like three dimensions ((Customer (C), Supplier (S), Part (P)), and four dimensions (Customer (C), Supplier (S), Part (P), Time (T)) evaluate the results.

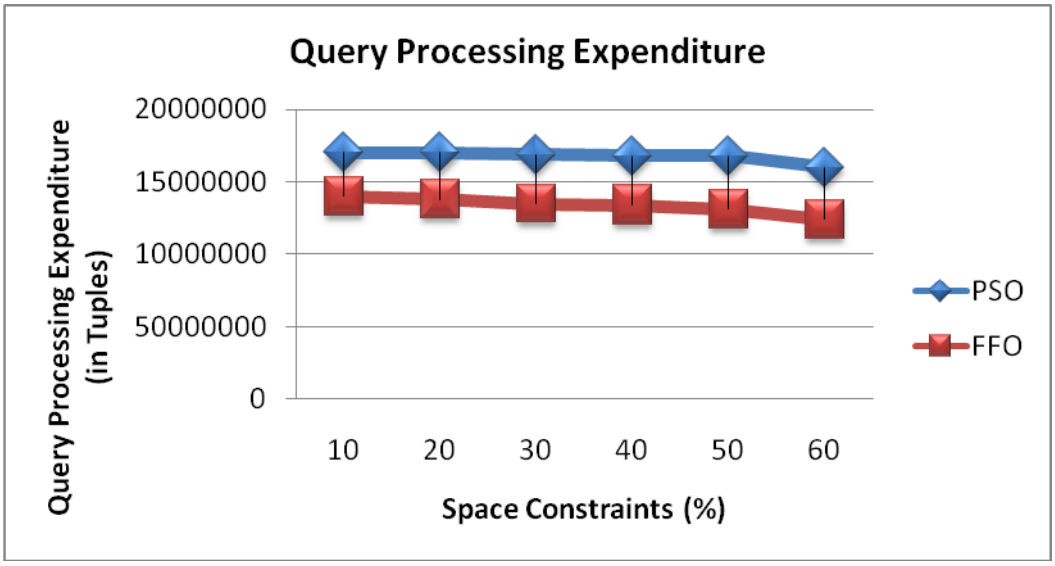

Figure 3.3. Query Processing Expenditure for PSO and FFO approaches (3 Dimensions)

Figure 3.3 illustrates that the FFO and PSO generate query processing expenditure (in tuples) as 124000000 and 160000000 in terms of space constrictions for selecting optimal data cubes. The FFO obtains superior results 26% than PSO with several space constraints for three dimensional data.

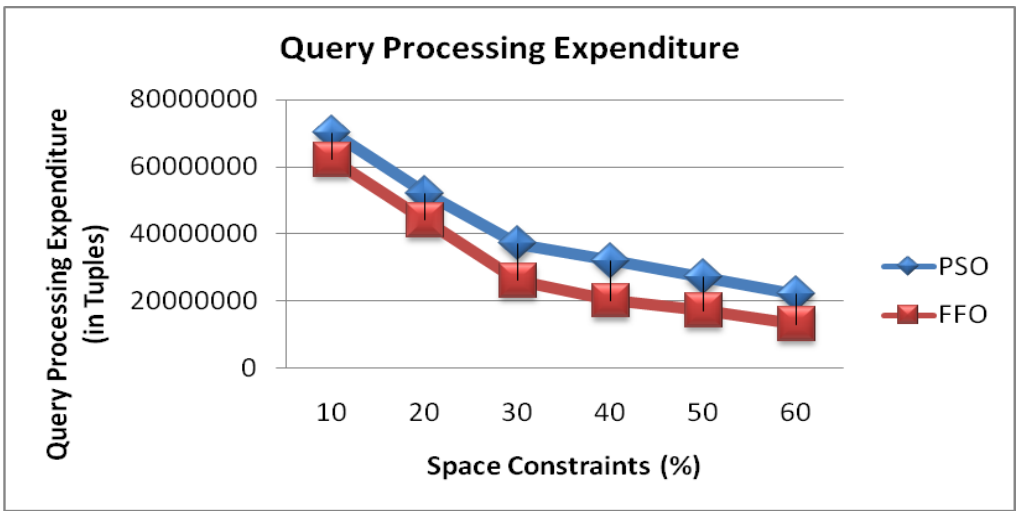

Figure 3.4. Query Processing Expenditure for PSO and FFO approaches (4 Dimensions)

Figure 3.4 illustrates that the FFO and PSO generate query processing expenditure (in tuples) as 130000000 and 220000000 in terms of space constrictions for selecting optimal data cubes. The FFO obtains superior results 41% than PSO with several space constraints for four dimensional data. Figure 3.3 & Figure 3.4 illustrates that the FFO generates better results for selecting optimal data cubes with minimum OLAP query processing expenditure as compared to PSO for all dimensional data.

We also evaluated the results of FFO and PSO in terms of identical and arbitrary frequencies (range 0 to 1).

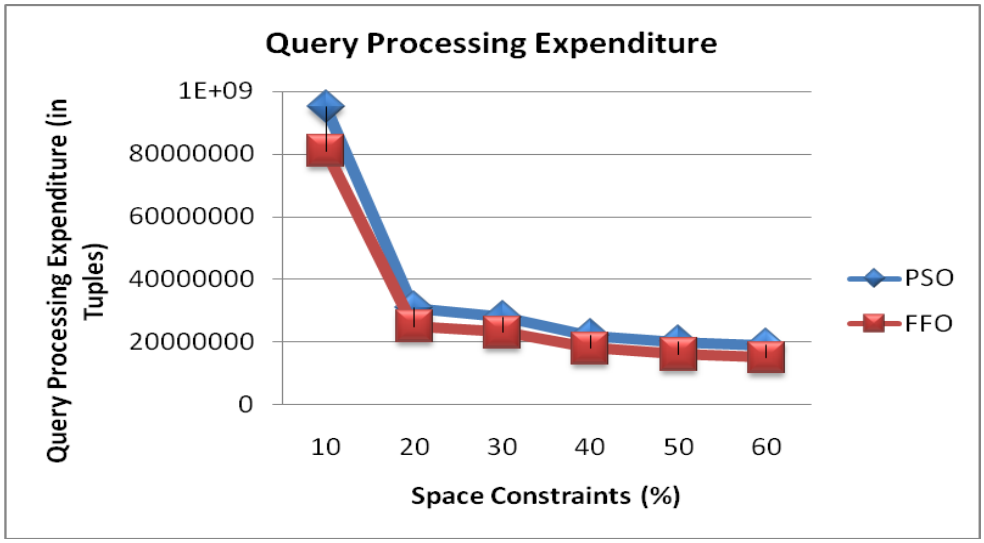

Figure 3.5. Query Processing Expenditure for PSO and FFO approaches (Identical frequencies)

Figure 3.5 illustrates that the FFO and PSO generate query processing expenditure (in tuples) as 150000000 and 190000000 in terms of space constrictions for selecting optimal data cubes. The FFO obtains superior results 22% than PSO with several space constraints for identical frequencies.

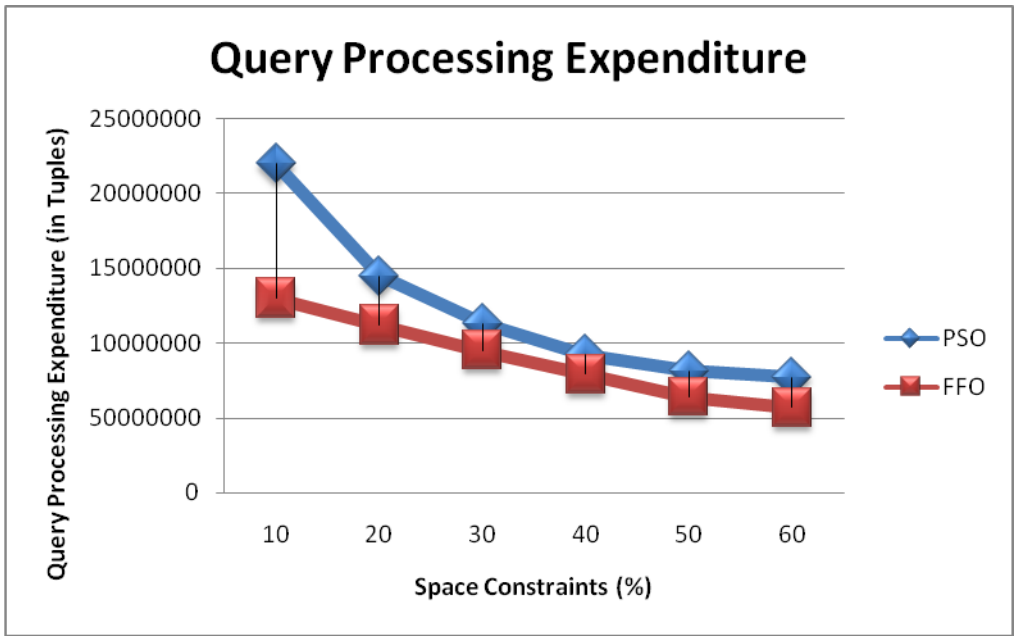

Figure 3.6. Query Processing Expenditure for PSO and FFO approaches (Arbitrary frequencies)

Figure 3.6 illustrates that the FFO and PSO generate query processing expenditure (in tuples) as 57000000 and 77000000 in terms of space constrictions for selecting optimal data cubes. The FFO obtains superior results 26% than PSO with several space constraints for arbitrary frequencies.

Figure 3.5 and Figure 3.6 illustrates that the FFO obtains more efficient performance for choosing optimal data cubes as compared to PSO in all frequencies with least OLAP query processing expenditure.

3.4. Summary and Discussion

In this work, FFO based optimal materialized cube selection is performed on lattice structure with multidimensional data over OLAP framework. The results are evaluated on multidimensional data in terms of frequency and number of dimensions. The analysis of performance of FFO illustrates the improved quality, efficiency of FFO to reduce the

OLAP query processing expenditure as compared to PSO. In the future, several optimization approaches will be implemented for optimized cube selection by considering the time complexity as a factor.

3.5. Limitation of FFO Approach

In this chapter, it has been demonstrated that FFO mechanism is oppressed to choose the optimal materialized cube on lattice structure. This mechanism has computed greater effectiveness according to query processing expenditure analyzed against PSO. On the other hand, there are tiny restrictions in the FFO that FFO purely enchanted in a limited unsurpassed prospective evaluation at the subsequently development stage due to small convergence accuracy. The FFO as fit might be unproductive to comprehend the utility most constructive as it moves away from the origin peak or in the negative boundary. These limits are isolated by developing another optimization mechanism for cube selection to further improve the effectiveness of FFO based optimal cube selection, which is prominent as Grey Wolf Optimization (GWO). The GWO is oppressed to choose the optimal data cubes and computed greater competence according to query processing expenditure.

CHAPTER-4
A Grey Wolf Optimization (GWO) based Cube Selection in OLAP Data Model

4.1. Introduction

The data cube assessments dependent on Online Analytical Processing (OLAP) trouble for numerous depositing splendours over broad information. In favour of appreciating question answering era pleasant with OLAP skeleton patrons and allowing complete industry organized notice compulsory, OLAP information is organized as a data cube model. The OLAP questions are answered in rapid and sturdy time by exploiting the cube embodiment for appraisals buyers. Until now this moreover insets insupportable charge, concerning to accumulation remembrance and time, yet as a data storage area had a typical length and extent which will be influential on stimulating procedure.

Thus, cube classification has visited to be refined fascinating to moderate question managing charge, preserving as a control the materializing breach. Numerous strategies and heuristics like divergence and voracious approaches have been exploited to suggest a vague solution. Here, a Grey Wolf Optimization (GWO) strategy is exploited in a lattice structure for finding the best data cube to decrease the question processing charge. The outputs describe the superior efficiency of GWO against GA, PSO [63] and ALO based on total dimensions and frequency.

In prior examination, data cubes are obtained from essential information through the vast quantity of information. The data cubes are obtained by utilizing optimization strategies like Genetic Algorithm (GA), Particle Swarm Optimization (PSO) [22, 23] and Ant Lion Optimization (ALO) in huge field to generate intellectual query dispensation in OLAP data model [24]. Here, a Grey Wolf Optimization (GWO) strategy is implemented in a lattice structure for discovering the best data cube and the results express the better quality efficiency of GWO against GA, PSO and ALO based on total dimensions and frequency to diminish the query dispensation expenses in numerous restriction places.

4.2. The Proposed Grey Wolf Optimization (GWO) based Cube Selection in OLAP Data Model

4.2.1. GWO Approach

The GWO is a bio inspired strategy; which exploits the grey wolves for guidance hierarchy and mode of hunting grouping among alpha (α), beta (β), delta (δ) and omega (ω) types and explores probing, offensive and encompassing quarry for optimization. The GWO is illustrated mathematically as below:

Social Hierarchy. The GWO obtains the values of α, β, δ and ω for Ist, IInd, IIIrd fittest and respite of solutions respectively.

Encircling Prey. The grey wolves are utilized for encircling prey to hunt evaluating through eq. (1) to eq. (4).

$$\vec{P} = \left|\vec{U}.\overrightarrow{X_L}(\tau) - \vec{X}(\tau)\right| \quad (1)$$

$$\vec{X}(\tau+1) = \overrightarrow{X_L}(\tau) - \vec{V}.\vec{P} \quad (2)$$

$$\vec{V} = 2.\vec{m}.\overrightarrow{a_1} - \vec{m} \quad (3)$$

$$\vec{U} = 2.\overrightarrow{a_2} \quad (4)$$

Where,

τ = current repetition

$\overrightarrow{X_L}$ = vector location of prey

$\vec{X}$ = vector location of grey wolf

$\vec{U}$ & $\overrightarrow{V}$ = performance indicator vector

$\overrightarrow{a_1}$ & $\overrightarrow{a_2}$ = arbitrary vector within [0, 1]

$\vec{m}$ = decrease the value from 2 to 0.

The grey wolf is altered its location by utilizing eq. (1) and eq. (2).

Hunting. The encircle and prey are positioned in space; which are discovered with the help of grey wolves. The hunting behavior is exploited to alter the locations for achieving the optimal location and to evaluate the optimal results for α, β, and δ through eq. (5) to eq. (7).

$$\overrightarrow{P_\alpha} = |\overrightarrow{U_1}.\overrightarrow{X_\alpha} - \vec{X}| \quad \& \quad \overrightarrow{P_\beta} = |\overrightarrow{U_2}.\overrightarrow{X_\beta} - \vec{X}| \quad \& \quad \overrightarrow{P_\delta} = |\overrightarrow{U_3}.\overrightarrow{X_\delta} - \vec{X}| \tag{5}$$

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{V_1}.(\overrightarrow{P_\alpha}) \quad \& \quad \overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{V_1}.(\overrightarrow{P_\alpha}) \quad \& \quad \overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{V_1}.(\overrightarrow{P_\alpha}) \tag{6}$$

$$\vec{X}(\tau + 1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3} \tag{7}$$

Attacking Prey (Exploitation). When grey wolves discontinue movement; subsequently they assault the prey through dropping the component of $\vec{m}$ from 2 to 0. When |V| < 1, it pressurizes the wolf crowd to assault the prey and |V| > 1, this pressurizes wolves to determine vast field as a choice of utilization.

4.2.2. Lattice Structure

The lattice structure modulates completely plausible data cubes at divergent levels of stockpiling through illustrating cubes on their reliance. Two cubes B_r and B_s are associated through a link for obtaining dependence relationship ($B_r \leq B_s$). This relationship explains that a reply to the OLAP query agreed by B_r, preserve furthermore be conferred through B_s. The dependence relationship of information has D dimensions with probable 2^D data cubes in lattice structure. For example, a 3 dimensional data “Electronics Sales” (Item (I), City (C), year (Y)) has 2^3 = 8 data cubes; which is represented in figure 4.1.

The least stockpiling shows through base cube (like ICY) and utmost stockpiling shows through the peak cube. The structure explains that a reply to the OLAP query specified through child cube (*,C,*) preserves furthermore through every parent cubes (*,C,Y), (I,C,*) or (I,C,Y) by closing data nearby little dimensions.

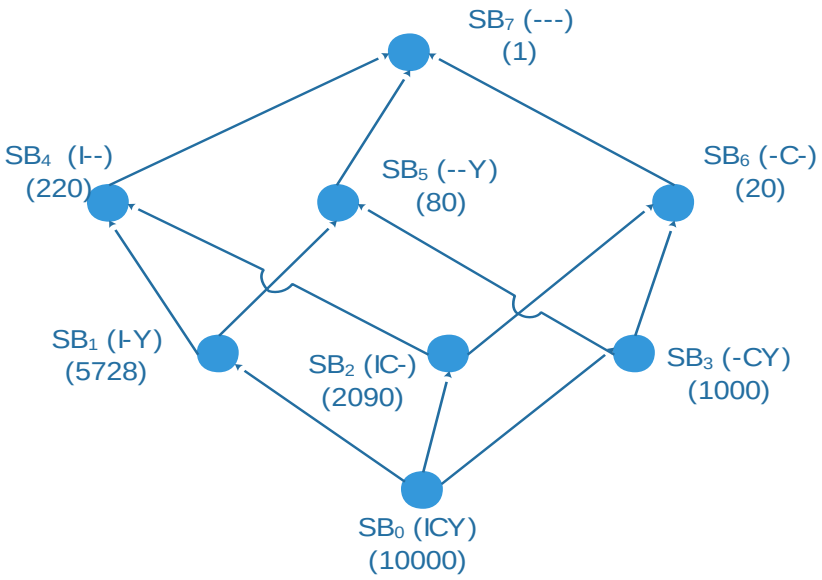

Figure 4.1: Lattice Structure

4.2.3. GWO approach for Cube Selection

The OLAP model is mainly utilized for decreasing expenses of query handling and maintenance with gratifying a condition such as materialized field. The Whole Query Spending (WQS) for an OLAP query is calculated through replying expenditure and occurrence of query utilizing eq. (8).

$$WQS = \sum_{e=1}^{D} Oq_e . S(q_e . B_M) \qquad (8)$$

Where,

Oq_e = Occurrence of q_e query

$S(q_e . B_M)$ = q_e query reply spending (expenses) with B_M (Materialized Cube) evaluating through eq. (9).

$$S(q_e.B_M) = Minimum(|SB_e|, Lant(q_e.B_M)) \quad (9)$$

Where,

$|SB_e|$ = total q_e query sub-cubes

$Lant(q_e.B_M)$ = Least antecedent (Lant) origin of q_e query with B_M

The Maintenance Spending (MS) of a materialized cube $B_r \in B_M$ is calculated depending on least antecedent (Lant) of B_r utilizing eq. (10).

$$MS(B_r, B_M) = |Lant(B_r, B_M)| \quad (10)$$

Later than Whole Maintenance Spending (WMS) is calculated through eq. (11).

$$WMS = O_{MS} \sum_{B \in B_M} MS(B, B_M) \quad (11)$$

Where,

O_{MS} = Occurrence of enclosure in maintaining relationships.

Therefore, Whole Spending Function (WSF) is evaluated by merging the WQS (Whole Query Spending) and WMS (Whole Maintenance Spending) through eq. (12).

$$Minimum \quad WSF = \sum_{e=1}^{D} Oq_e.S(q_e.B_M) + O_{MS} \cdot \sum_{B \in B_M} MS(B, B_M) \quad (12)$$

The GWO is applied on the function WSF to obtain optimal results (like data cubes). The primary values of spending can be calculated from platform relationship and value of B_M is null primarily. Later then, optimal results are calculated utilizing GWO. The WSF function is exploited as discrepancy standard, while it obtains the output fitness during subsequent numerous targets. The WSF function generates minimum value to provide smallest spending of query by means of utmost fitness.

4.3. Result Analysis

The MATLAB 2019a tool (windows 8, 8 GB RAM, Core i3 processor) is used to implement GWO technique and the results are evaluated depending on total dimensions

and frequency (Figure 4.2 (a) (b) (c)). The outcomes demonstrate the superior performance of GWO by means of a least query dispensation expense. At this time, numerous positions of space constraints are deliberated as 10% - 80% to examine the GWO technique.

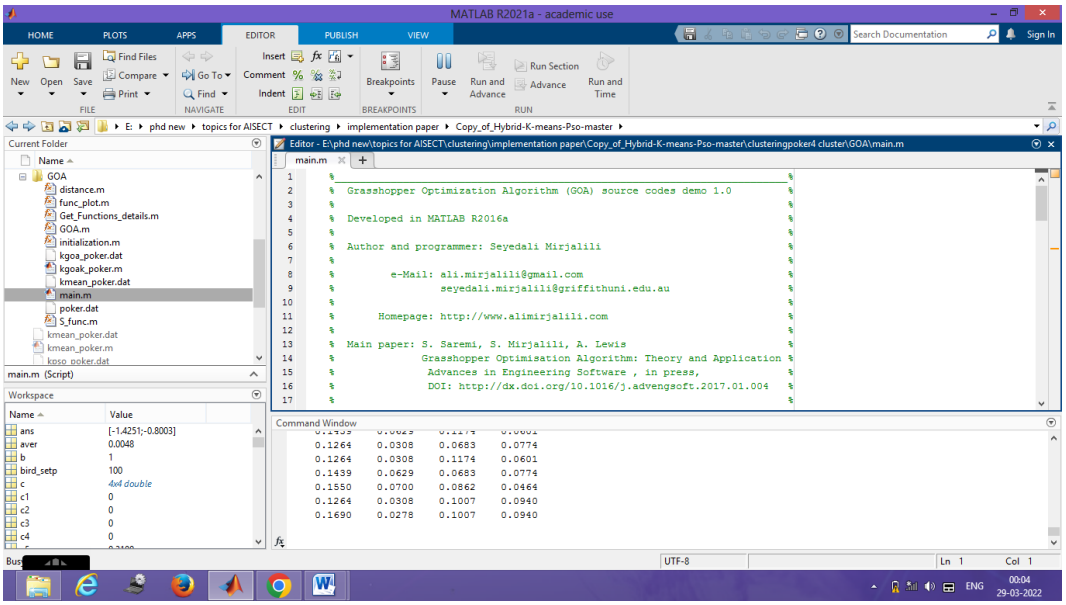

Figure 4.2 (a). Implementation in MATLAB

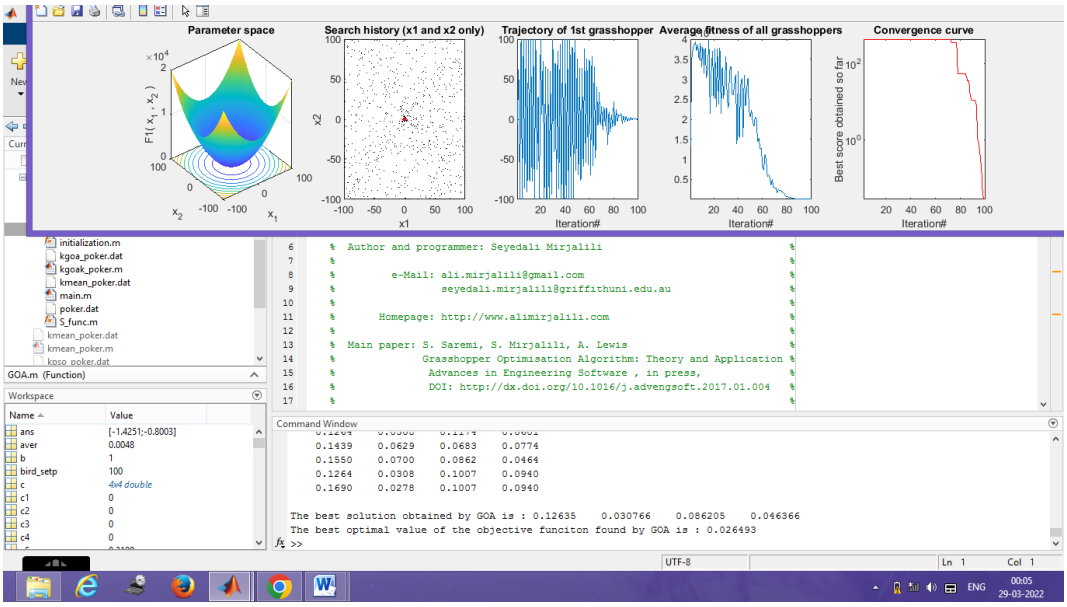

Figure 4.2 (b). Implementation in MATLAB

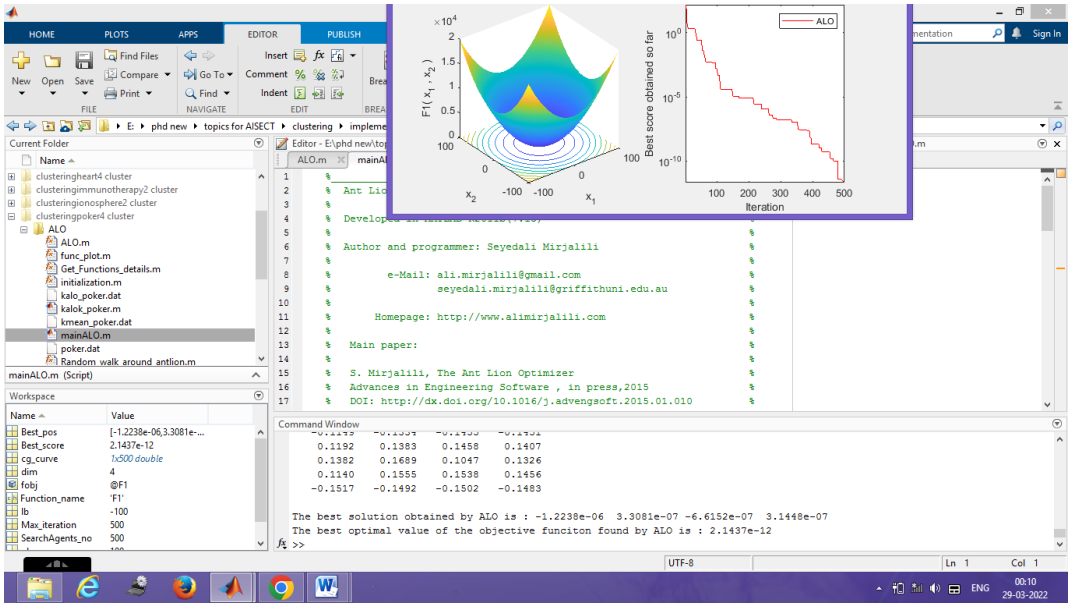

Figure 4.2 (c). Implementation in MATLAB

The GWO technique is applied on data with various dimensions such as 3 dimensions (Item (I), City (C), Year (Y)), 4 dimensions (Item (I), City (C), Year (Y), Sales-in-dollars (S)) and 5 dimensions (Item (I), City (C), Year (Y), Sales-in-dollars (S), Demand (D)) to calculate the outputs.

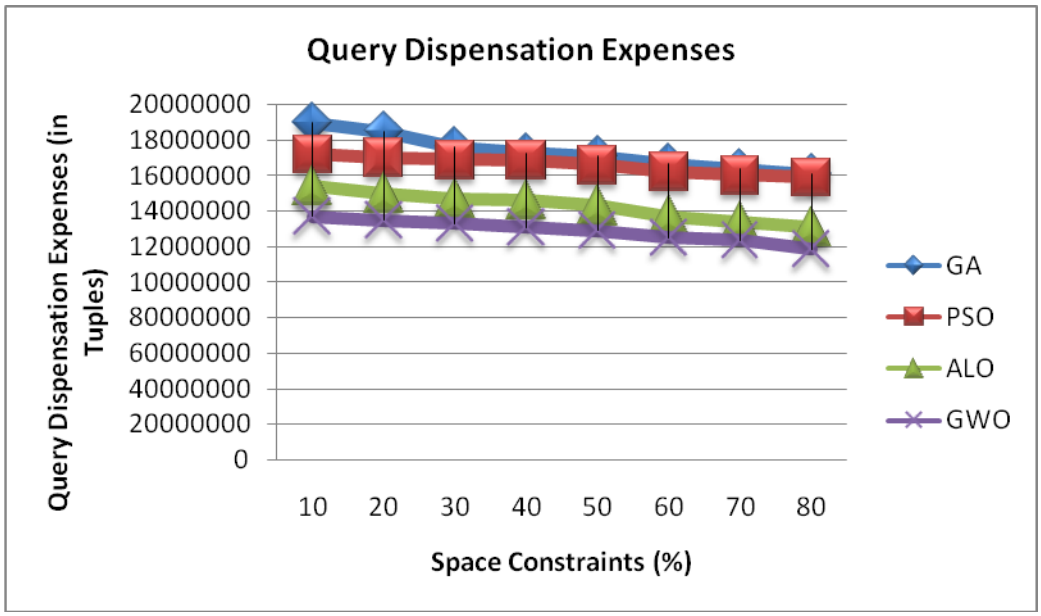

Figure 4.3: Query Dispensation Expenses on 3 Dimensions

Figure 4.3 demonstrates that the GWO, ALO, PSO and GA generate query dispensation expenses (in tuples) as 118700000, 131000000, 159000000 and 161000000 respectively

in terms of space constrictions for selecting optimal data cubes for three dimensional data. The PSO obtains superior results 5% than GA; The ALO obtains superior results 18% than PSO and 22% than GA; and GWO obtains superior results 10% than ALO and 26% than PSO and 29% than GA with several space constraints for three dimensional data.

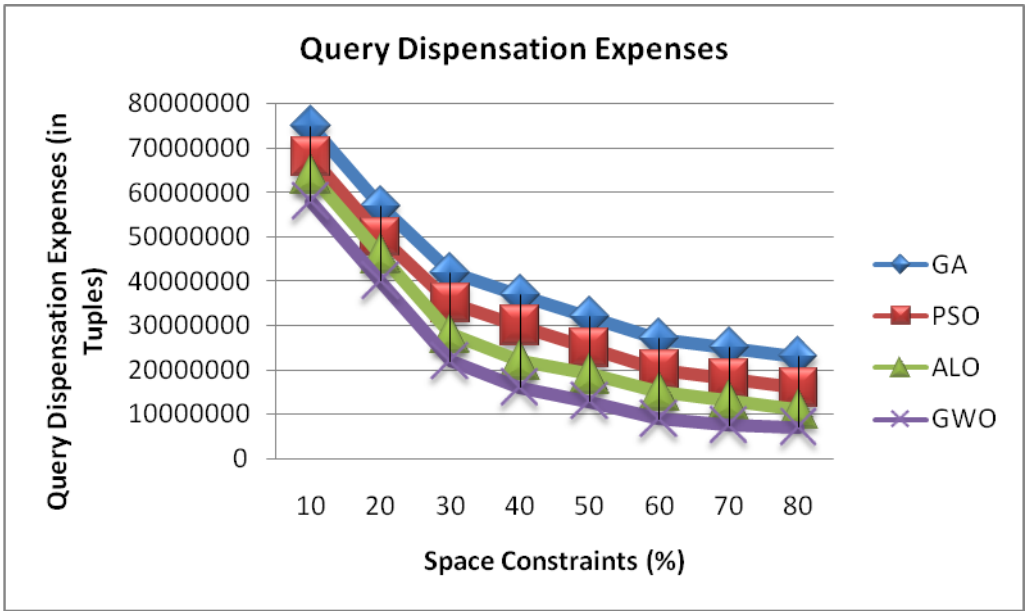

Figure 4.4: Query Dispensation Expenses on 4 Dimensions

Figure 4.4 demonstrates that the GWO, ALO, PSO and GA generate query dispensation expenses (in tuples) as 71000000, 110000000, 160000000 and 230000000 respectively in terms of space constrictions for selecting optimal data cubes for four dimensional data. The PSO obtains superior results 31% than GA; The ALO obtains superior results 32% than PSO and 53% than GA; and GWO obtains superior results 36% than ALO and 56% than PSO and 70% than GA with several space constraints for four dimensional data.

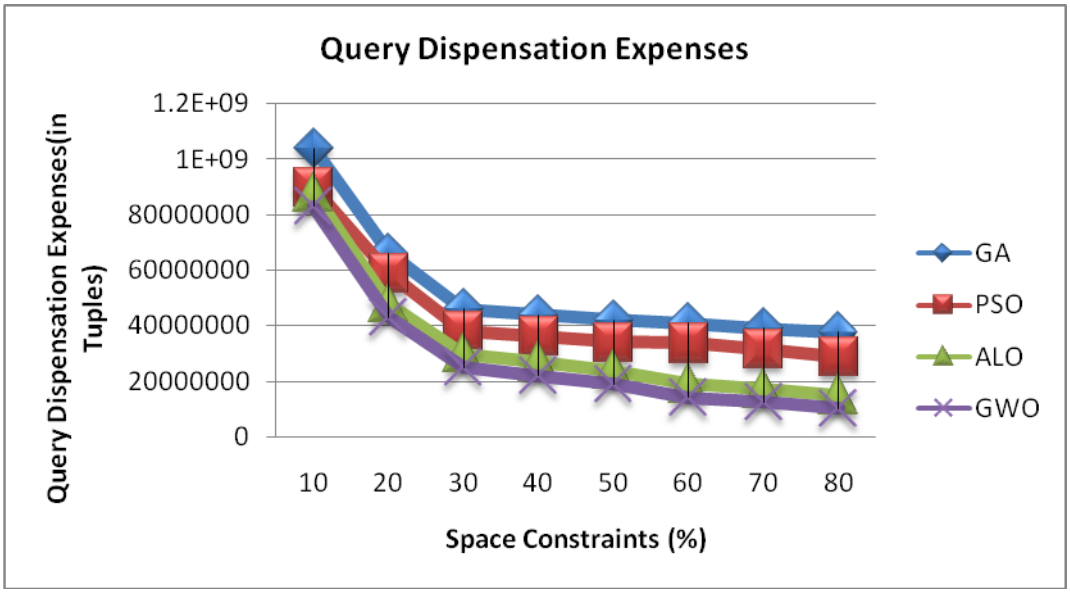

Figure 4.5: Query Dispensation Expenses on 5 Dimensions

Figure 4.5 demonstrates that the GWO, ALO, PSO and GA generate query dispensation expenses (in tuples) as 102154780, 150364120, 286352000 and 375620000 respectively in terms of space constrictions for selecting optimal data cubes for five dimensional data. The PSO obtains superior results 24% than GA; The ALO obtains superior results 48% than PSO and 60% than GA; and GWO obtains superior results 33% than ALO and 65% than PSO and 73% than GA with several space constraints for five dimensional data.

Figure 4.3 to figure 4.5 demonstrates that the GWO obtains superior outputs for picking up the best data cubes having least OLAP query dispensation expenses against prior methods like GA, PSO and ALO for whole dimensional data. Here, GWO obtains superior outputs 11% against ALO, 21% against PSO and 29% against GA for 3 dimensional data; 10% against ALO, 15% against PSO and 23% against GA for 4 dimensional data; 6% against ALO, 8% against PSO and 21% against GA for 5 dimensional data. The GWO has also calculated the outputs depending on uniform and capricious frequencies within (0, 1).

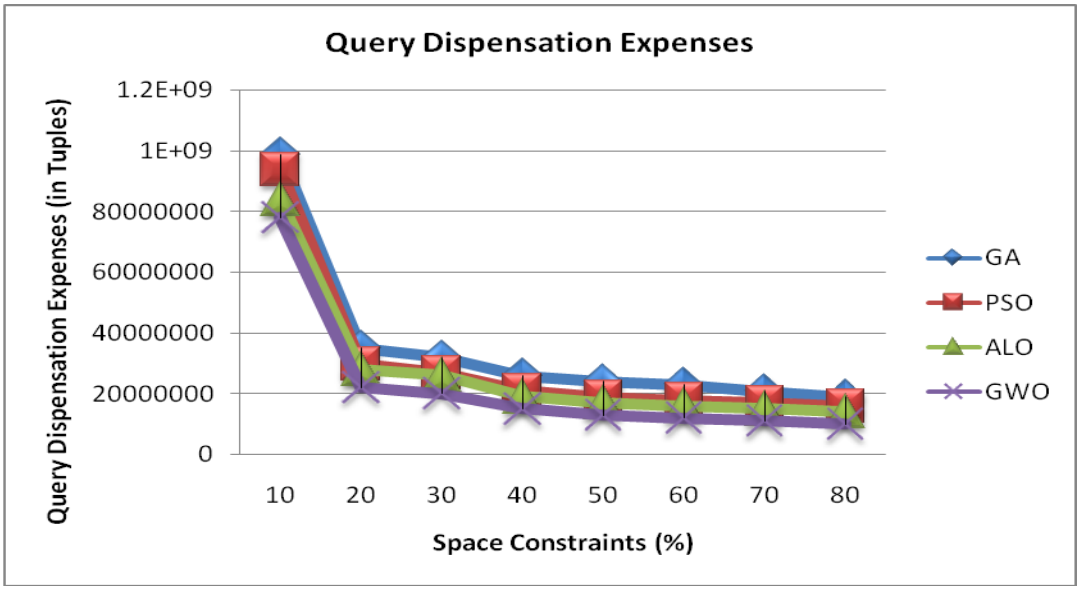

Figure 4.6: Query Dispensation Expenses for Uniform Frequencies

Figure 4.6 demonstrates that the GWO, ALO, PSO and GA generate query dispensation expenses (in tuples) as 100000000, 140000000, 160000000 and 190000000 respectively in terms of space constrictions for selecting optimal data cubes for uniform frequencies. The PSO obtains superior results 16% than GA; The ALO obtains superior results 13% than PSO and 27% than GA; and GWO obtains superior results 29% than ALO and 38% than PSO and 48% than GA with several space constraints for uniform frequencies.

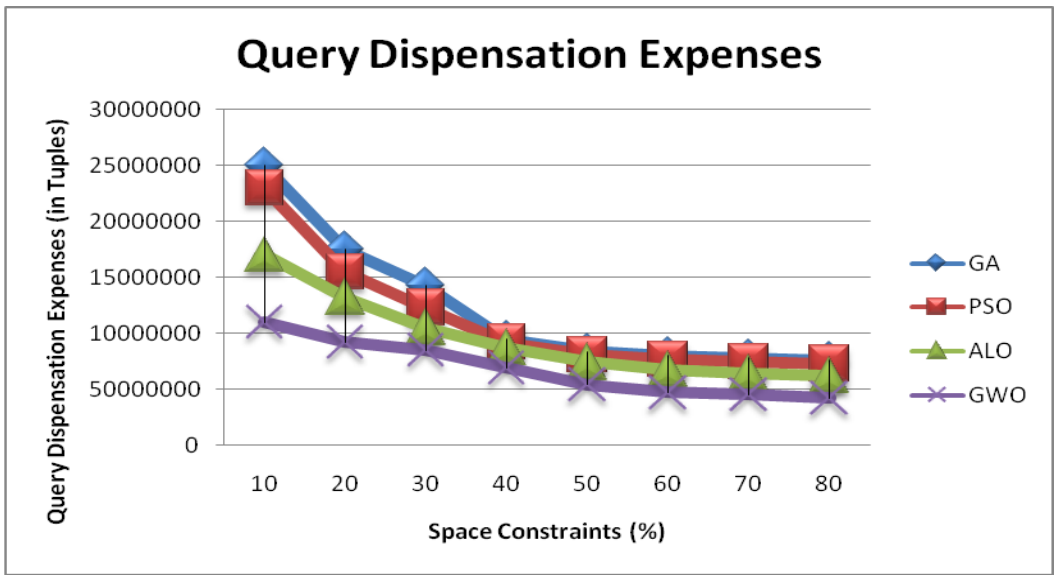

Figure 4.7: Query Dispensation Expenses for Capricious Frequencies

Figure 4.7 demonstrates that the GWO, ALO, PSO and GA generate query dispensation expenses (in tuples) as 42000000, 62000000, 73000000 and 76000000 respectively in terms of space constrictions for selecting optimal data cubes for capricious frequencies. The PSO obtains superior results 5% than GA; The ALO obtains superior results 16% than PSO and 19% than GA; and GWO obtains superior results 33% than ALO and 43% than PSO and 45% than GA with several space constraints for uniform frequencies.

Figure 4.6 and figure 4.7 demonstrates that the GWO obtains better quality outputs for choosing the best data cubes along with least OLAP query dispensation expenses against prior methods like GA, PSO and ALO for whole dimensional data. Here, GWO obtains better quality outcomes 8% against ALO, 17% against PSO and 22% against GA for uniform frequencies; 35% against ALO, 52% against PSO and 56% against GA for capricious frequencies.

4.4. Summary and Discussion

In this paper, GWO is introduced to choose the best materialized cube utilizing the OLAP multidimensional information model with lattice structure. The outcomes are calculated on data having various dimensions based on total dimensions and frequency. The GWO is examined over lattice structure to find out the optimal data cube for minimizing the query dispensation expenses. Various optimization strategies will be introduced to perform an optimal selection of data cubes with more dimensions and evaluating the time and space complexity as performance indicators in the future.

CHAPTER-5
Comparative Analysis of FFO and GWO Approaches

5.1. Introduction

For evaluating the performance of the FFO and GWO schemes describing in earlier chapter, eq. (12) $(Minimum\ WSF = \sum_{e=1}^{D} Oq_e.S(q_e.B_M) + O_{MS}.\sum_{B \in B_M} MS(B,B_M))$ (from heading 4.2.3 in chapter 4) is used to discover optimal data cubes for materialized OLAP models. Eq. (1) to eq. (13) is computed alike as earlier chapter (heading 4.2 in chapter 4). Subsequently FFO and GWO are performed (one by one) on function WSF to obtain the optimal materialized data cubes for OLAP model. Here analysis is computed on MATLAB 2019a environment by means of windows 8 and core i3 processor. The FFO and GWO results have been computed as compared to the earlier work such as ALO, PSO and GA depending on query dispensation expenses in terms of total dimensions and frequencies.

5.2. Comparative Analysis of FFO and GWO Approaches

The MATLAB 2019a tool (windows 8, 8 GB RAM, Core i3 processor) is used to implement FFO and GWO techniques and the results are evaluated depending on total dimensions and frequency. The outcomes demonstrate the superior performance of FFO and GWO by means of a least query dispensation expense. At this time, numerous positions of space constraints are deliberated as 10% - 80% to examine the FFO and GWO technique.

The FFO and GWO technique are applied on data with various dimensions such as 3 dimensions (Item (I), City (C), Year (Y)), 4 dimensions (Item (I), City (C), Year (Y), Sales-in-dollars (S)) and 5 dimensions (Item (I), City (C), Year (Y), Sales-in-dollars (S), Demand (D)) to calculate the outputs.

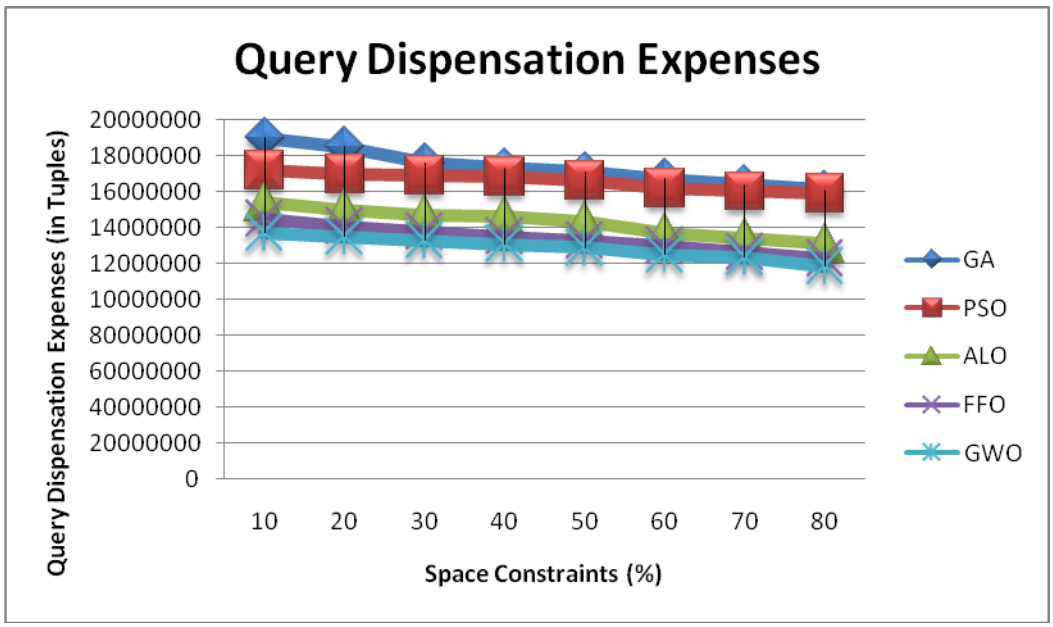

Figure 5.1: Query Dispensation Expenses on 3 Dimensions

Figure 5.1 demonstrates that the PSO obtains superior outputs 10% against GA; ALO obtains superior outputs 11% against PSO and 19% against GA; FFO obtains superior outputs 6% against ALO, 16% against PSO and 24% against GA; GWO obtains superior outputs 6% against FFO, 12% against ALO, 21% against PSO and 28% against GA for query dispensation expenses on 3 dimensional data.

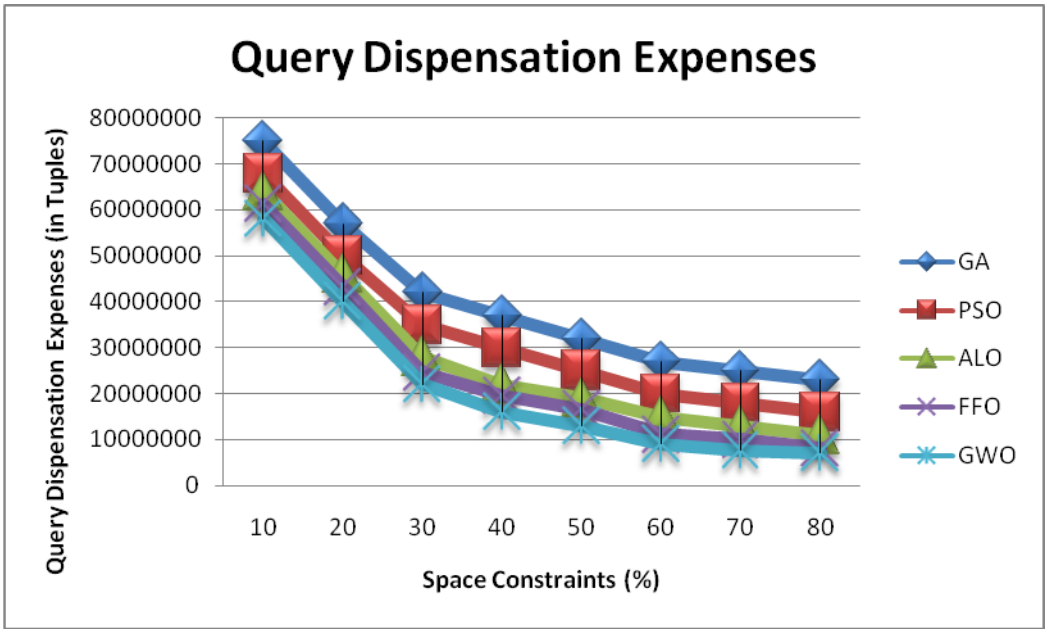

Figure 5.2: Query Dispensation Expenses on 4 Dimensions

Figure 5.2 demonstrates that PSO generates better quality outcomes 10% against GA; ALO generates better quality outcomes 6% against PSO and 15% against GA; FFO generates better quality outcomes 5% against ALO, 10% against PSO and 19% against GA; GWO generates better quality outcomes 6% against FFO, 10% against ALO, 15% against PSO and 23% against GA for query dispensation expenses on 4 dimensional data.

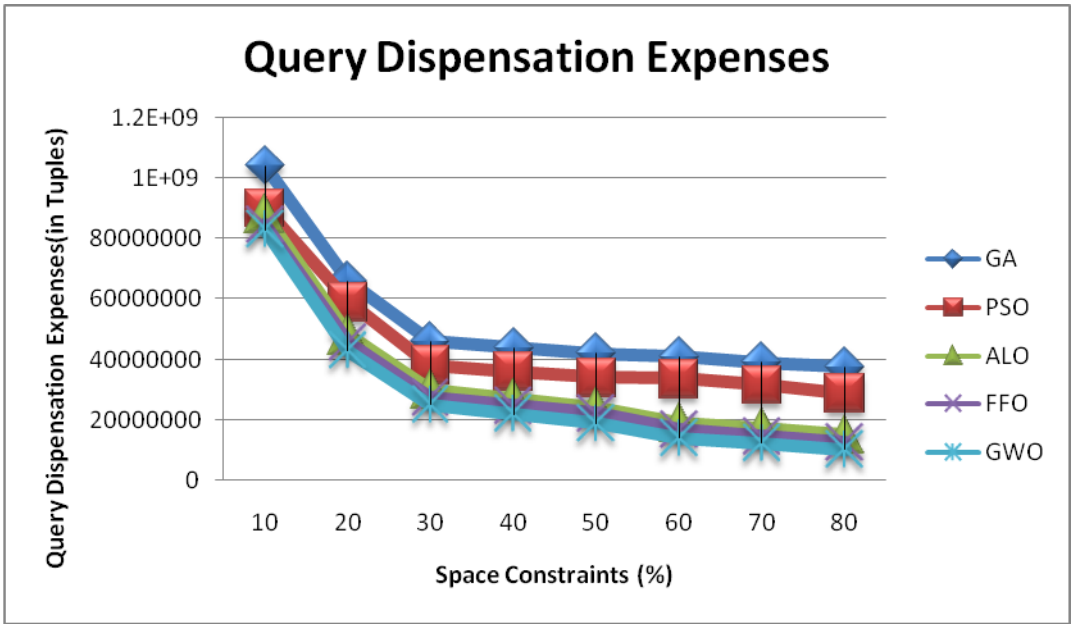

Figure 5.3: Query Dispensation Expenses on 5 Dimensions

Figure 5.3 demonstrates that PSO computes improved results 14% against GA; ALO computes improved results 3% against PSO and 16% against GA; FFO computes improved results 4% against ALO, 7% against PSO and 19% against GA; GWO computes improved results 2% against FFO, 6% against ALO, 8% against PSO and 21% against GA for query dispensation expenses on 5 dimensional data.

The GWO has also calculated the outputs depending on uniform and capricious frequencies within (0, 1).

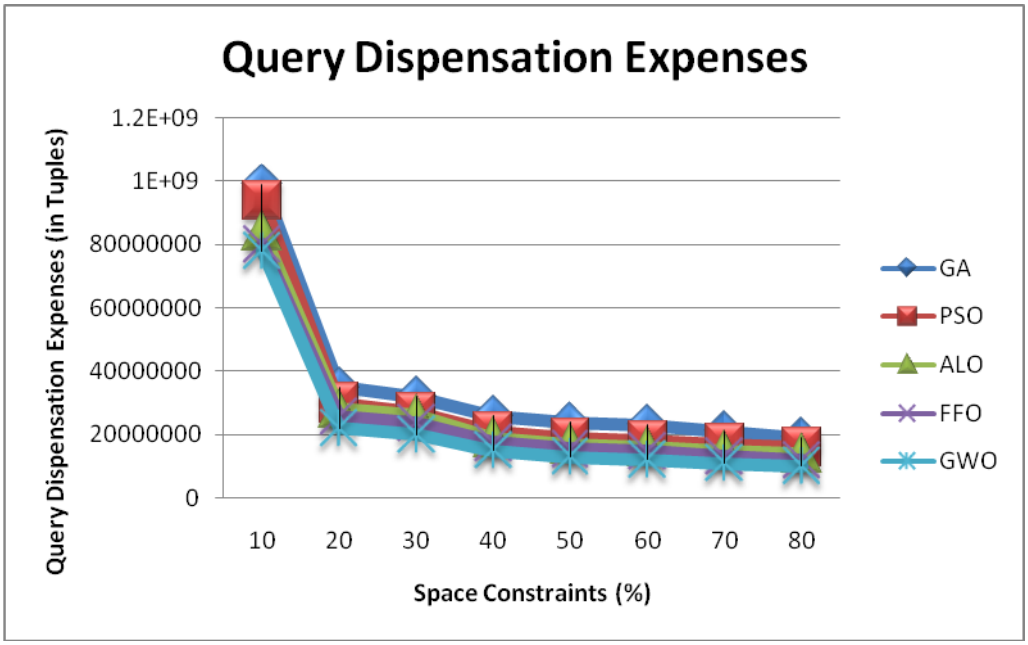

Figure 5.4: Query Dispensation Expenses for Uniform Frequencies

Figure 5.4 demonstrates that PSO evaluates advanced results 6% against GA; ALO evaluates advanced results 11% against PSO and 16% against GA; FFO evaluates advanced results 5% against ALO, 15% against PSO and 20% against GA; GWO evaluates advanced results 3% against FFO, 8% against ALO, 18% against PSO and 22% against GA for query dispensation expenses on uniform frequencies.

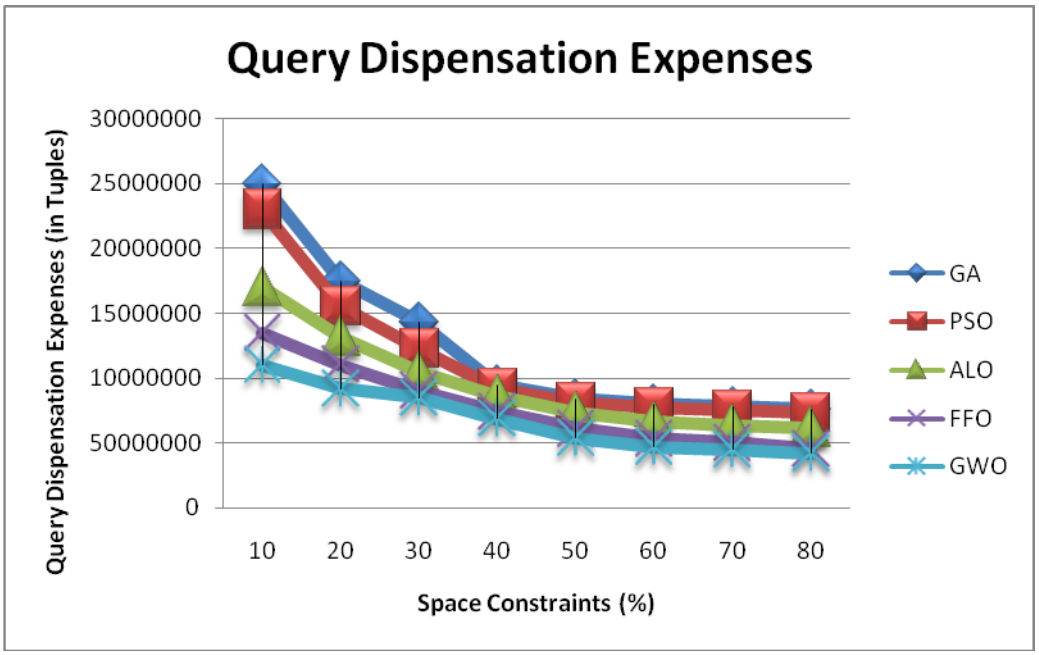

Figure 5.5: Query Dispensation Expenses for Capricious Frequencies

Figure 5.5 demonstrates that PSO calculates advanced outcomes 8% against GA; ALO calculates advanced outcomes 26% against PSO and 32% against GA; FFO calculates advanced outcomes 21% against ALO, 42% against PSO and 46% against GA; GWO calculates advanced outcomes 19% against FFO, 36% against ALO, 53% against PSO and 56% against GA for query dispensation expenses on capricious frequencies.

It summaries that first scheme FFO obtains advanced outputs as compared to ALO, PSO and GA depending on query dispensation expenses in terms of total dimensions and frequencies. The second scheme GWO obtains advanced outputs as compared to FFO, ALO, PSO and GA depending on query dispensation expenses in terms of total dimensions and frequencies. The figures (Figure 5.1 to Figure 5.5) demonstrate that the query dispensation expenses of all schemes are reduced by means of increasing the space constraints. Consequently, the expenses are reduced during the optimal solution searching process.

5.3. Summary and Discussion

Here, FFO based optimal materialized cube selection is performed on lattice structure with multidimensional data over OLAP framework. The results are evaluated on multidimensional data in terms of frequency and number of dimensions. The analysis of performance of FFO illustrates the improved quality, efficiency of FFO to reduce the OLAP query processing expenditure as compared to PSO.

Another scheme, GWO is introduced to choose the best materialized cube utilizing the OLAP multidimensional information model with lattice structure. The outcomes are calculated on data having various dimensions based on total dimensions and frequency. The GWO is examined over lattice structure to find out the optimal data cube for minimizing the query dispensation expenses as compared to FFO, ALO, PSO and GA in terms of space constraints. Various optimization strategies will be introduced to perform an optimal selection of data cubes with more dimensions and evaluating the time and space complexity as performance indicators in the future. In the future, several optimization approaches will be implemented for optimized cube selection by considering the time complexity as a factor.

In both the mechanism GWO and FFO, the optimal materialized cubes are selected in taking the account of query expenditure to improve the quality of OLAP model. Still, other strategy like clustering is also utilized to improve the OLAP model processing by dividing the big data into groups. So, in next chapter, a Dragon Fly Optimization based Clustering (DFOC) approach is developed to enhance the efficiency of data clustering by generating optimal clusters from multidimensional clinical data for OLAP.

CHAPTER-6
Improving the Performance of Multidimensional Clinical Data for OLAP using an Optimized Data Clustering Approach

6.1. Introduction

Medicine is a fresh way to utilize for curing, analyzing and detecting the diseases through data clustering with OLAP (Online Analytical Processing). The large amount of multidimensional clinical data is reduced the efficiency of OLAP query processing by enhancing the query accessing time. Hence, the performance of OLAP model is improved by using data clustering [2, 3, 4] in which huge data is divided into several groups (clusters) with cluster heads to achieve fast query processing in least time.

In this chapter, a Dragon Fly Optimization based Clustering (DFOC) approach is proposed to enhance the efficiency of data clustering by generating optimal clusters from multidimensional clinical data for OLAP. The results are evaluated on MATLAB 2019a tool and shown the better performance of DFOC against other clustering methods ACO, GA and K-Means [67, 79] in terms of intra-cluster distance, purity index, F-measure, and standard deviation.

Here, several researchers introduced data clustering techniques [17] for improving the efficiency of multidimensional data model [15, 16]. K-Means [19] is one of the widely useful clustering techniques for simple and easy development for huge amount of data. But, there is still some drawback in K-Means like highly dependable on initial cluster. So, here we utilized the optimization for data clustering on huge multidimensional data sets to obtain optimal results by removing the limitation of K-Means. The GA (Genetic Algorithm) and ACO (Ant Colony Optimization) are two most popular optimization approaches are used with data clustering to improve the quality of clustering. In this work, we implemented a DOFC (Dragon Fly Optimization based Clustering) approach on clinical multidimensional datasets to generate optimal clusters with cluster centroids and compared the results with ACO, GA and K-Means in terms of several parameters.

6.2. Dragon Fly Optimization based Clustering (DFOC) Approach

6.2.1. DFOC Approach

Dragon Fly Optimization (DFO) approach is a nature inspired methodology which is stirred by dragon fly`s stagnant and energetic behaviour on the basis of examination and utilization. DFO offers three crucial standard Severance (SR), Configuration (CF) and Consistency (CS) and two former significant convictions of brimming Foodstuff sources Appeal (FA) and Opponent Escaping (OE) represented in (1) to (5).

$$SR_p = -\sum_{q=1}^{N_n}(X - X_q) \quad (1)$$

$$CF_p = \frac{\sum_{q=1}^{N_n} V_q}{N_n} \quad (2)$$

$$CS_p = \frac{\sum_{q=1}^{N_n} X_q}{N_n} \quad (3)$$

$$FA_p = X^+ - X \quad (4)$$

$$OE_p = X^- - X \quad \mathbf{(5)}$$

Here, X =dragonfly individual location, X^+ =foodstuff location, X^- =opponent location, N_n =neighbours number, V_q & X_q = q^{th} individual`s velocity and location. The speed vector is evaluated by utilizing (6), then dragonfly`s location is updated through (7).

$$\nabla X_{t+1} = \left(sr \times SR_p + cf \times CF_p + cs \times CS_p + fa \times FA_p + oe \times OE_p\right) + wt \times \nabla X_t \quad (6)$$

$$X_{t+1} = X_t + \nabla X_{t+1} \quad (7)$$

Here, sr, cf, cs, fa, oe and wt are steady coefficient.

DFOC approach

START

Assign N data entities as cluster centroids randomly.

For each clusters

Initialize standards of dragonfly population (X_p) and speed vector (X_p) with p=1,2,3,.....N_n

While finish circumstance is not pleased

Calculate entire dragonfly`s intention standards

Update foodstuff and opponent source

Update sr, cf, cs, fa, oe and wt

Calculate SR, CF, CS, FA, and OE by (1) to (5)

Update neighbour`s area

If (minimum 1 neighbour locates in dragonfly area)

Update speed vector by (6)

Update location vector by (7)

Else

Update location vector by (7)

End If

Confirm and accurate next location of dragonfly based on capricious restrictions

End While

End For

STOP

In DFOC, the DFO is applied on multidimensional clinical datasets to obtain optimal clusters with cluster heads (centroids) with minimizing the intra-cluster distances among data elements. In DFO, every cluster is assigned as dragonfly and each data entities are assigned as explore agents. All dragon fly`s positions are updated according

to fitness standards with reducing the intra-cluster distances among data entities to find out the optimal clusters with centroids.

6.2.2. Multidimensional Clinical Datasets

The DFOC is applied on several multidimensional clinical datasets describing in table 1.

Table 6.1. Multidimensional Clinical Datasets

Sr. No.	Clinical Dataset			
	Dataset	***No. of instances***	***No. of dimensions***	***No. of Clusters***
1	Cancer	683	9	2
2	Cryotherapy	90	7	2
3	Liver Patient	583	10	2
4	Heart Patients	297	14	4

6.3. Result and Analysis

The DFOC is implemented on all four clinical data sets (table 1) on MATLAB 2019a tool (windows 8, 8 GB RAM, Core i3 processor) (Figure 6.1 (a) (b) (c) (d)). The results are obtained in terms of intra-cluster distance, purity index, F-measure, and standard deviation over 1000 repetitions.

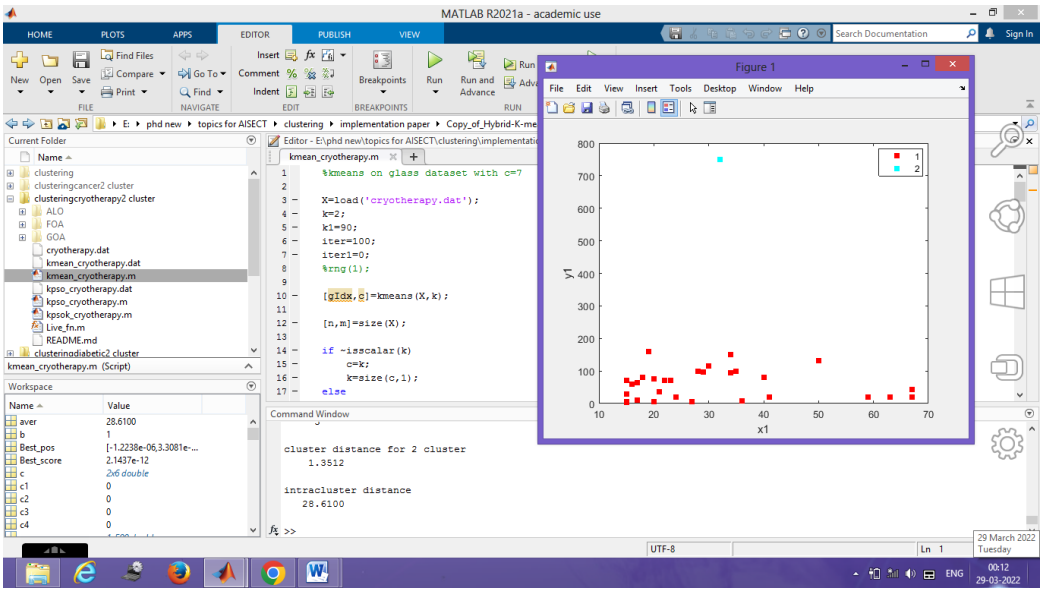

Figure 6.1 (a). Implementation in MATLAB

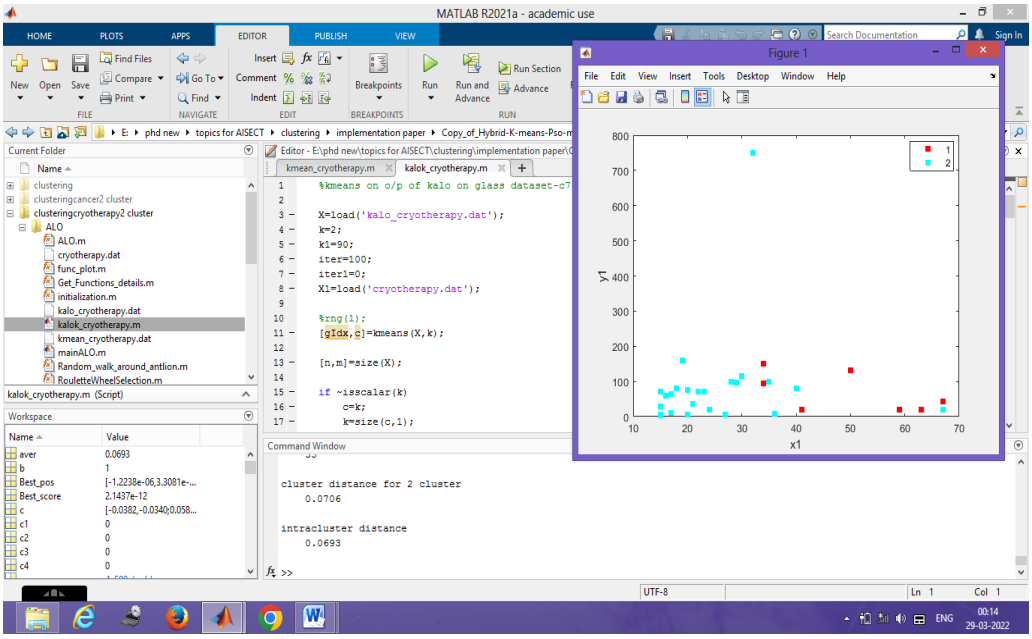

Figure 6.1 (b). Implementation in MATLAB

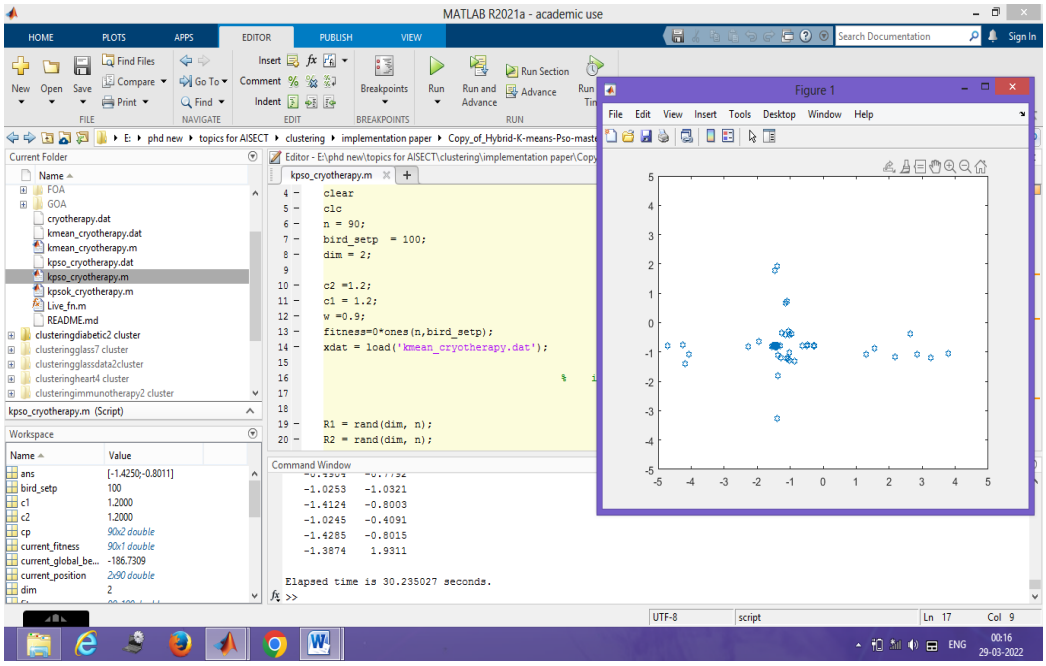

Figure 6.1 (c). Implementation in MATLAB

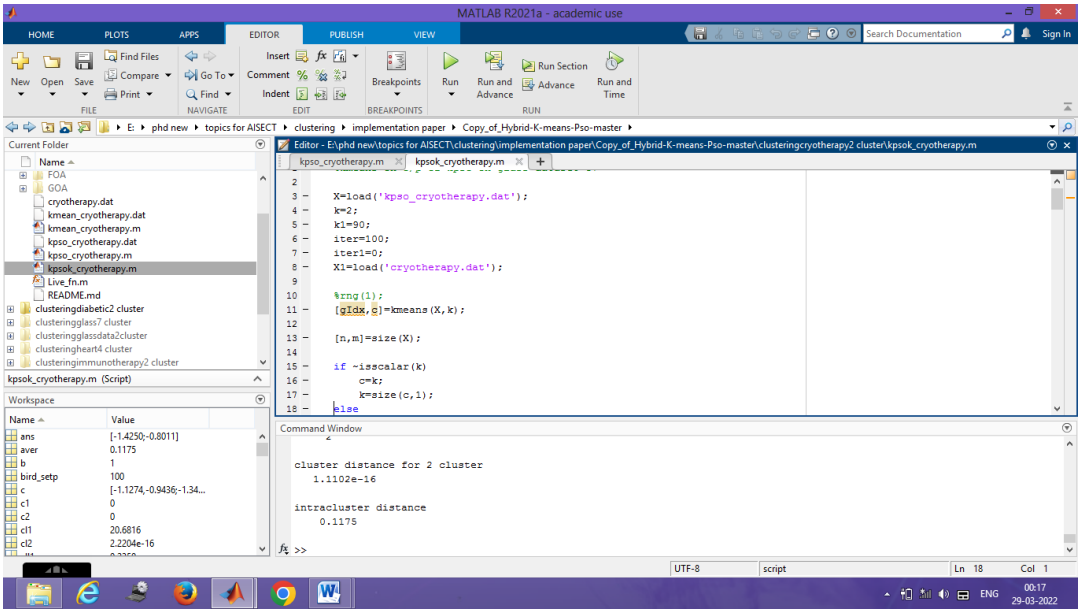

Figure 6.1 (d). Implementation in MATLAB

6.3.1. Intra-cluster distance

It is explained as the mean distance among data entities in identical cluster. It must have least value for optimized clustering.

6.3.2. Purity Index

It is illustrated the frequent clustering of data entities by using (8). It must have maximum value for optimized clustering.

$$P_I = \sum_{s=1}^{K} \frac{(|CR_r| \frac{\max(|CR_{rs}|)}{|CR_s|})}{|D_S|} \qquad (8)$$

Here, K = clusters number,

$|CR_r|$ and $|CR_s|$ = rth class and sth cluster length

$|D_S|$ = dataset length

$|CR_{rs}|$ = data entities of rth class locate to sth cluster.

6.3.3. F-Measure

It is obtained from precision (prec) and recall (rcl) for data reclamation by (9) to (12). It must have maximum value for optimized clustering.

$$prec(r,s) = \frac{|CR_{rs}|}{|CR_s|} \quad (9)$$

$$rcl(r,s) = \frac{|CR_{rs}|}{|CR_r|} \quad (10)$$

$$Fun(r,s) = \frac{2 \times prec(r,s) \times rcl(r,s)}{prec(r,s) + rcl(r,s)} \quad (11)$$

$$FM = \sum_{r=1}^{K} \frac{|CR_r|}{|D_S|} max\{Fun(r,s)\} \quad (12)$$

6.3.4. Standard Deviation

It is explained the data clustering strength about the mean standards using (13). It must have least value for optimal clustering.

$$S_D = \sqrt{\frac{\sum(de - \overline{de})}{|D_S|}} \quad (13)$$

Here, de = data entity in dataset,

$\left|\overline{de}\right|$ = mean of data entities in a dataset

Table 6.2. Results for Cancer Dataset

Approaches	Performance Parameters			
	Intra-cluster distance	***Purity index***	***Standard deviation***	***F-Measure***
K-Means	94.2641	0.86	0.5248	0.84
GA	0.3265	0.87	0.2153	0.85
ACO	0.08587	0.90	0.1042	0.87
DFOC	0.002514	0.95	0.024	0.92

Table 6.3. Results for Cryotherapy Dataset

Approaches	Performance Parameters			
	Intra-cluster distance	***Purity index***	***Standard deviation***	***F-Measure***
K-Means	19.3625	0.82	0.3521	0.76
GA	0.3142	0.90	0.1241	0.85
ACO	0.0541	0.91	0.0624	0.86
DFOC	0.00325	0.95	0.00786	0.90

Table 6.4. Results for Liver Patients Dataset

Approaches	Performance Parameters			
	Intra-cluster distance	***Purity index***	***Standard deviation***	***F-Measure***
K-Means	42.3214	0.87	0.4215	0.85
GA	0.4201	0.88	0.2641	0.86
ACO	0.0845	0.90	0.0758	0.87
DFOC	0.00464	0.91	0.00882	0.88

Table 6.5. Results for Heart Patients Dataset

Approaches	Performance Parameters			
	Intra-cluster distance	***Purity index***	***Standard deviation***	***F-Measure***
K-Means	12.3654	0.78	0.20365	0.74
GA	0.50241	0.81	0.07548	0.76
ACO	0.0365	0.84	0.02364	0.80
DFOC	0.00124	0.91	0.0074	0.95

Table 6.6. Results for Average Rank for all Datasets based on Intra-Cluster Distance

Approaches	Datasets				
	Cancer	***Cryotherapy***	***Liver Patients***	***Heart Patients***	***Average Rank***
K-Means	94.2641 (4)	19.3625 (4)	42.3214 (4)	12.3654 (4)	4
GA	0.3265 (3)	0.3142 (3)	0.4201 (3)	0.50241 (3)	3
ACO	0.08587 (2)	0.0541 (2)	0.0845 (2)	0.0365 (2)	2
DFOC	0.002514 (1)	0.00325 (1)	0.00464 (1)	0.00124 (1)	1

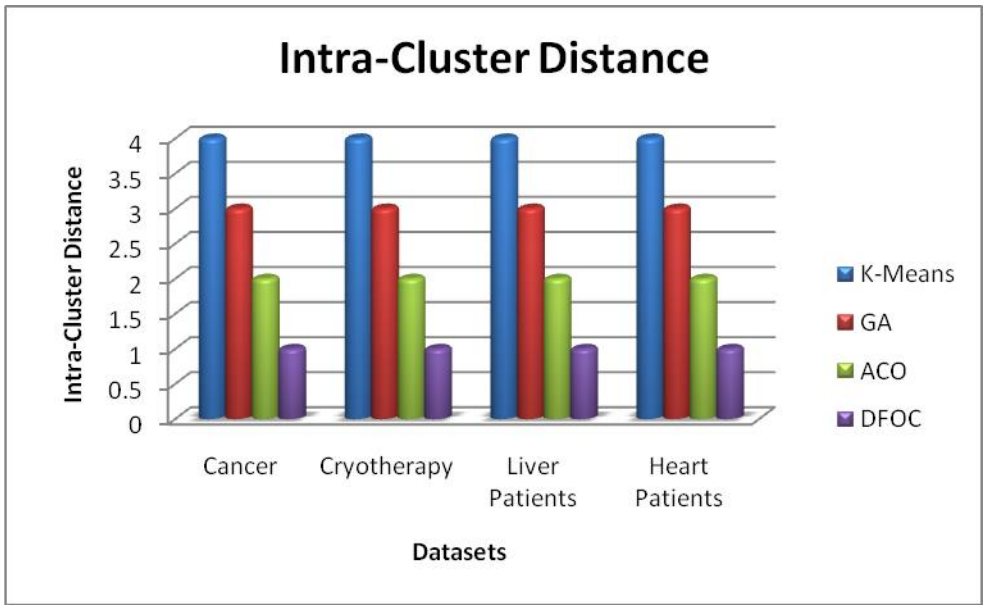

Figure 6.2. Average Rank for all datasets based on Intracluster Distance

The outputs in figure 6.2 illustrate that GA obtains superior results 25% than K-Means; ACO obtains superior results 34% than GA and 50% than K-Means; DFOC obtains superior results 50% than ACO and 67% than GA and 75% than K-Means in terms of intra-cluster distance for total four clinical datasets.

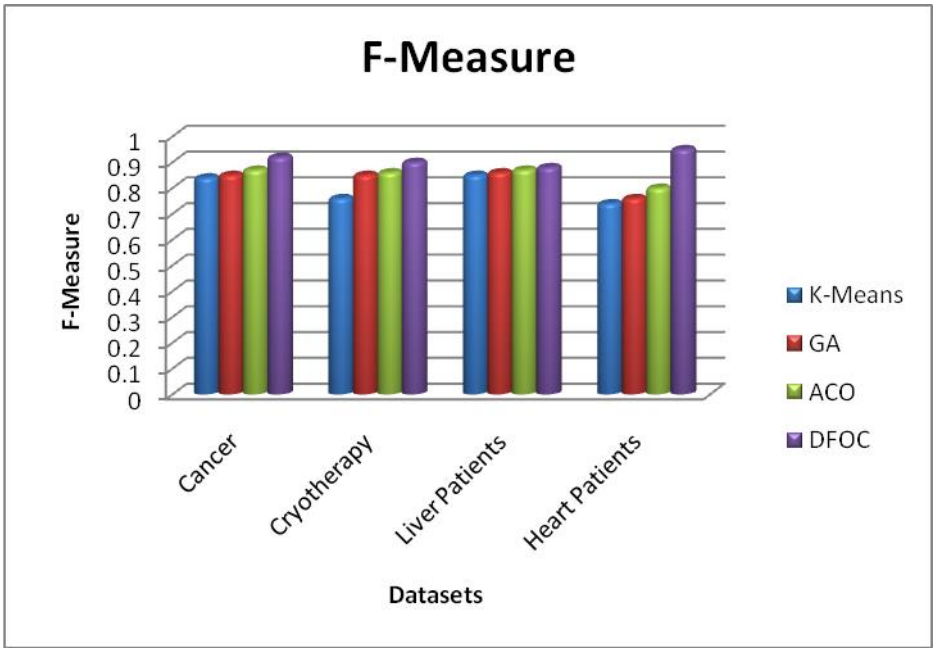

Figure 6.3. F-Measure for all datasets

The outputs in figure 6.3 illustrate that GA obtains superior results 5% than K-Means; ACO obtains superior results 8% than GA and 15% than K-Means; DFOC obtains superior results 19% than ACO and 25% than GA and 29% than K-Means in terms of F-Measure for total four clinical datasets.

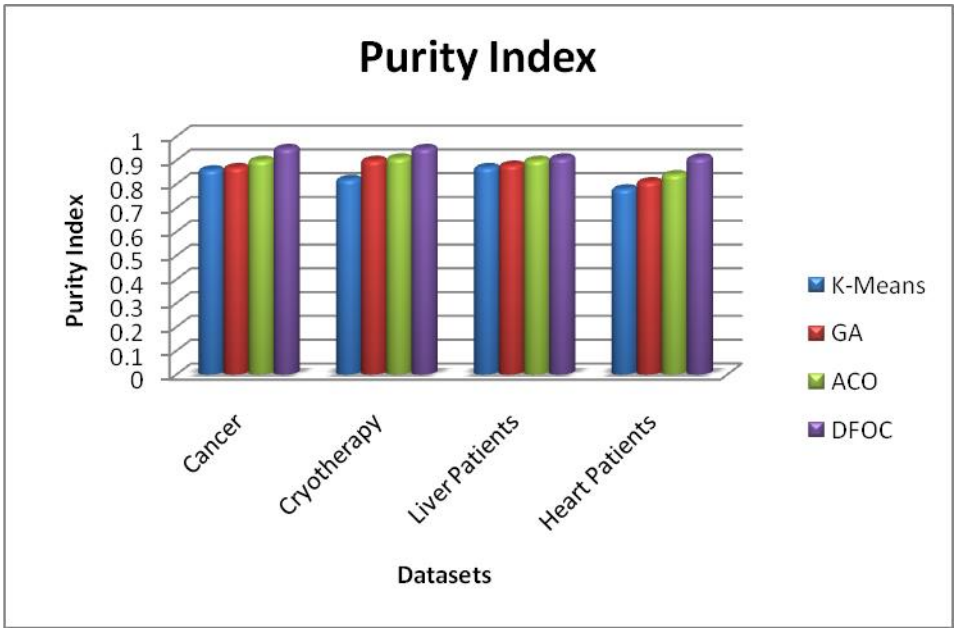

Figure 6.4. Purity Index for all dataets

The outputs in figure 6.4 illustrate that GA obtains superior results 6% than K-Means; ACO obtains superior results 10% than GA and 17% than K-Means; DFOC obtains superior results 21% than ACO and 27% than GA and 32% than K-Means in terms of purity index for total four clinical datasets.

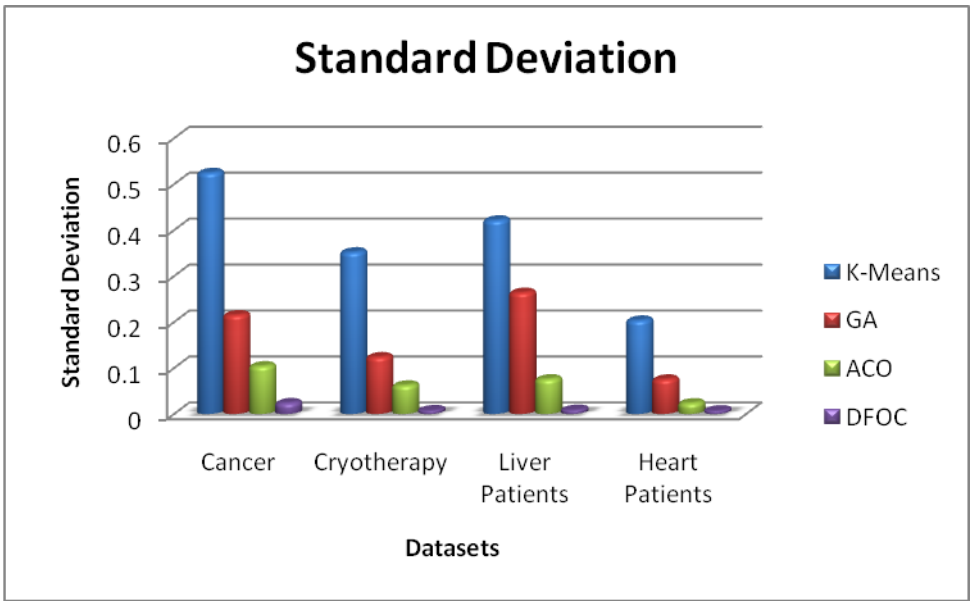

Figure 6.5. Standard Devition for all datasets

The outputs in figure 6.5 illustrate that GA obtains superior results 16% than K-Means; ACO obtains superior results 23% than GA and 38% than K-Means; DFOC obtains superior results 33% than ACO and 56% than GA and 72% than K-Means in terms of standard deviation for total four clinical datasets.

The results in table 6.2 to table 6.6 and figure 6.2 to figure 6.5 illustrates the better quality results of DFOC on all four multidimensional clinical datasets against K-Means, GA and ACO in terms of intra-cluster distance, F-measure, purity index and standard deviation. Due to better examination and utilization, DFOC improves the search space in global area for generating optimal cluster; hence DFOC generates enhanced outputs as compare to prior approaches.

6.4. Summary and Discussion

In this work, a Dragon Fly Optimization based Clustering (DFOC) approach is implemented to improve the performance of data clustering by obtaining optimized clusters from multidimensional clinical data for OLAP. The outcomes are examined on

MATLAB 2019a tool and illustrated the superior efficiency of DFOC as compared to prior approaches ACO, GA and K-Means in terms of intra-cluster distance, purity index, F-measure, and standard deviation.

6.5. Limitation of DFOC Approach

In this chapter, it has been demonstrated that DFOC mechanism is oppressed to choose the optimal clusters for clinical datasets. This mechanism has computed greater efficacy according to intra-cluster distance, purity index, F-Measure and standard deviation analyzed against ACO, GA and K-Means. On the other hand, there are tiny restrictions in the DFO that unique linear convergence restriction originate the development of exploration and exploitation unhinged, unsteady convergence velocity and simple to collapse into local optimum. These limits are isolated by developing another optimization mechanism for data clustering to further improve the effectiveness of DFOC based optimal data clustering, which is prominent as KMeans-Salp Swarm Optimization based Clustering (K-SSOC). The K-SSOC is oppressed to choose the optimal clusters for multidimensional datasets and computed greater competence according to various factors.

CHAPTER-7
Improve the Performance of Multidimensional Data for OLAP by using an Optimization Approach

7.1. Introduction

The performance of query processing over OLAP (Online Analytical Processing) model is decreased due to higher query access time for huge multidimensional data. Therefore, the clustering is introduced to improve the OLAP model efficiency by getting quick query processing because of dividing the large data into various clusters [21, 30]. The K-Means is a famous technique of clustering the data into groups to solve various real life issues. However, K-Means [26, 29] has some drawbacks like sensitivity to primary centroid assortment in cluster and local optimum convergence.

Hence, a KMeans-Salp Swarm Optimization based Clustering (K-SSOC) is implemented to improve the performance of K-Means by providing optimal clustering over huge OLAP multidimensional data. The outcomes are obtained on MATLAB 2019a environment based on the parameter purity index, standard deviation, F-measure, intra-cluster distance and running time complexity over 1000 iterations. The results illustrate the superior performance of K-SSOC against K-Means, ACO and PSO over total six multidimensional datasets based on parameters.

In the above literature, various researchers implemented clustering approaches over huge multidimensional data to enhance the performance of the OLAP model [18, 19]. The K-Means is an eminent clustering approach for performing division of huge data [20, 21] into groups in minimum time consumption. However, the K-Means has some limitations such as primary centroid selection sensitivity. These limitations are reduced or removed by using optimization approaches with clustering over numerous multidimensional datasets. The Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) [22, 23, 24] are two famous optimization algorithms, which are utilized to perform optimal clustering to enhance the clustering efficiency. In this paper, a K-SSOC (KMeans-Salp Swarm Optimization based Clustering) approach is

implemented to perform optimized clustering over multidimensional OLAP data and the outcomes are analyzed against K-Means, ACO and PSO on the basis of various performance parameters.

7.2. KMeans-Salp Swarm Optimization based Clustering (K-SSOC) Approach

7.2.1. K-SSOC Approach

K-Means is a famous and efficient clustering approach on the basis of definite concern of examination of variances. The total dataset is separated into C_n clusters by utilizing eq. 1.

$$E_D = \sum_{x=1}^{C_n} \sum_{y=1}^{S} {e_D}^2 \left(T_x - M_y\right) \quad (1)$$

Here, ${e_D}^2\left(T_x - M_y\right)$ = square of Euclidean distance from x^{th} cluster centroid to y^{th} data component, S = dataset length.

The K-Means assigns every data component to the cluster nearer to the cluster centroid (the total data components mean value in a cluster) and after that it generates the matches among data components and centroid. At last, the matches are utilized to decide reassignment of data components to cluster centroid and this process will carry on until a satisfied situation is established.

The Salp Swarm Optimization (SSO) is a nature enthused meta-heuristic technique for performing the salp`s aquatic and flying behaviour based optimization. The leader and follower are two groups of salp`s population in which the follower works under the supervision of leader. The d-dimensional exploration area based matrix P is used to hold the entire salp`s positions with a source of foodstuff where d is used for number of problem`s essentials. The leader`s position is changed by utilizing the source of foodstuff (eq. 2).

$$P_d^1 = \begin{cases} F_d + \sigma_1((UB_d - LB_d)\sigma_2 + LB_d & \sigma_3 \geq 0 \\ F_d - \sigma_1((UB_d - LB_d)\sigma_2 + LB_d & \sigma_3 < 0 \end{cases} \quad (2)$$

Here, P_d^1= leader`s position (1st salp), F_d = foodstuff source position, UB_d = upper bound, and LB_d = lower bound (everyone is evaluated in dth dimension). σ_1, σ_2 & σ_3 = arbitrary numbers $(\sigma_2, \sigma_3 \in [0,1])$ and σ_1 (exploitation-exploration equivalence) is calculated by utilizing eq. 3.

$$\sigma_1 = 2e^{-\left(\frac{4r}{R}\right)^2} \tag{3}$$

Here, r = current repetition and R = number of repetitions. The follower`s positions are changed by utilizing eq. 4.

$$P_d^a = \frac{1}{2} \propto t^2 + ut \tag{4}$$

Here, P_d^a = follower`s position (ath salp in dth dimension), u = initial velocity and t = time. The $\propto$ is calculated by utilizing eq. 5.

$$\propto = \frac{v_f}{u} \qquad where, \quad v_f = \frac{P_f - P_0}{t} \tag{5}$$

Here, P_f= follower`s final position, $P_0$= follower`s initial position, v_f= follower`s final velocity, variance among repetitions is 1 in use and u is put to 0. Hence, the eq. 4 is converted to eq. 6.

$$P_d^a = \frac{1}{2}(P_d^a + P_d^{a-1}) \qquad where, a \geq 2 \tag{6}$$

The salp swarm chain is generated by evaluating eq. 2 to eq. 6.

The performance of K-Means clustering is improved by combining the K-Means with SSO approach. Primarily, K-Means is applied to obtain clusters with centroids from datasets depend upon the least Euclidian distance. Later then, SSO is initiated to generate the optimal centroids for whole clusters. In SSO, every cluster is allocated as

salps and each data component is allocated as search agents. Hence, every salp is changing their position in terms of fitness standards to decrease the total of intra-cluster distances mean values. The salps are generated their optimal positions to obtain optimal centroids for every cluster.

The K-SSOC approach is described as follows with standard values.

Input: R = number of repetitions, P_s= salp`s population, and C_n = number of clusters.

Algorithm 1. The K-SSOC Approach.	**Number of Executions**
1. START	
2. Allocate C_n data components as cluster centroid erratically	
3. WHILE a satisfied situation is not established	(R+1)
4. FOR each component of data	$R*(P_s+1)$
5. Generate Euclidean distance of each component of data to the centroid	$R*P_s*C_n$
6. Assign the component of dataset to cluster with least distance	$R*C_n$
7. END FOR	
8. Obtain the total data components mean value in each cluster	$R*C_n$
9. Obtain the new centroids based on mean values	$R*C_n$
10. END WHILE	
11. Return C_n clusters	
12. For each cluster	
13. Allocate the standards of salps population (number of data component) $P_s(s = 1,2,3,4, \dots\dots\dots\dots\dots\dots\dots\dots N)$	
14. While (a satisfied situation is not established)	
15. Calculate the entire salp`s fitness (search agents)	C_n*P_s
16. S^b= the best salp	C_n*P_s
17. Update σ_1 (arbitrary number) (eq. 3)	$C_n*(R+1)$
18. FOR whole salps (P_s)	$R*C_n*(P_s+1)$

19. If (a = =1)
20. Update the leader salp`s position (eq. 2) $R*C_n*P_s$
21. Else
22. Update the follower salp`s position (eq.6) $R*C_n*(P_s+1)$
23. END FOR
24. Amend the salps in terms of P_d^a & P_d^1 of components C_n*R
25. END While
26. Return S^b C_n
27. End For
28. Assign S^b as the new centroid of cluster
29. **STOP**

7.2.2. Multidimensional Datasets

The K-SSOC is implemented on various multidimensional datasets representing in table 1.

Table 7.1. Multidimensional Datasets

Sr. No.	Dataset	No. of Instances	No. of Dimensions	No. of Clusters
1	Glass	214	10	7
2	Waveform	5000	40	3
3	Ionosphere	351	34	2
4	Heart Failure	300	12	2
5	Thyroid	300	5	3
6	Wine	178	13	3

7.3. Results and Discussion

The K-SSOC is applied to total six multidimensional datasets (table 1) on MATLAB 2019a environment (windows 8, 8 GB RAM, Core i3 processor) (Figure 7.1 (a) (b) (c) (d)). The outcomes are generated based on the parameter purity index, standard deviation, F-measure and intra-cluster distance over 1000 iterations.

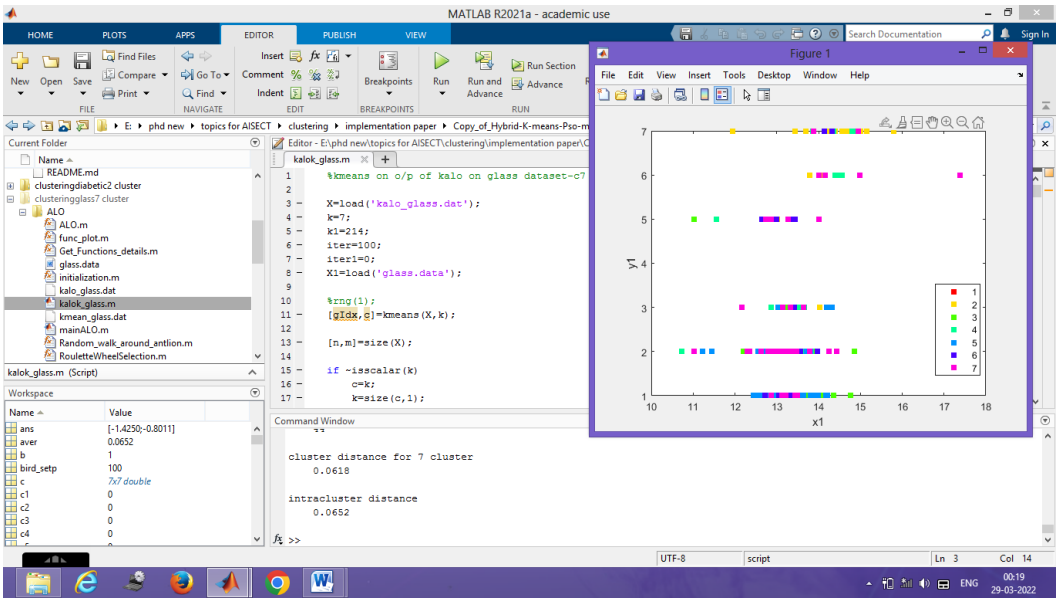

Figure 7.1 (a). Implementation in MATLAB

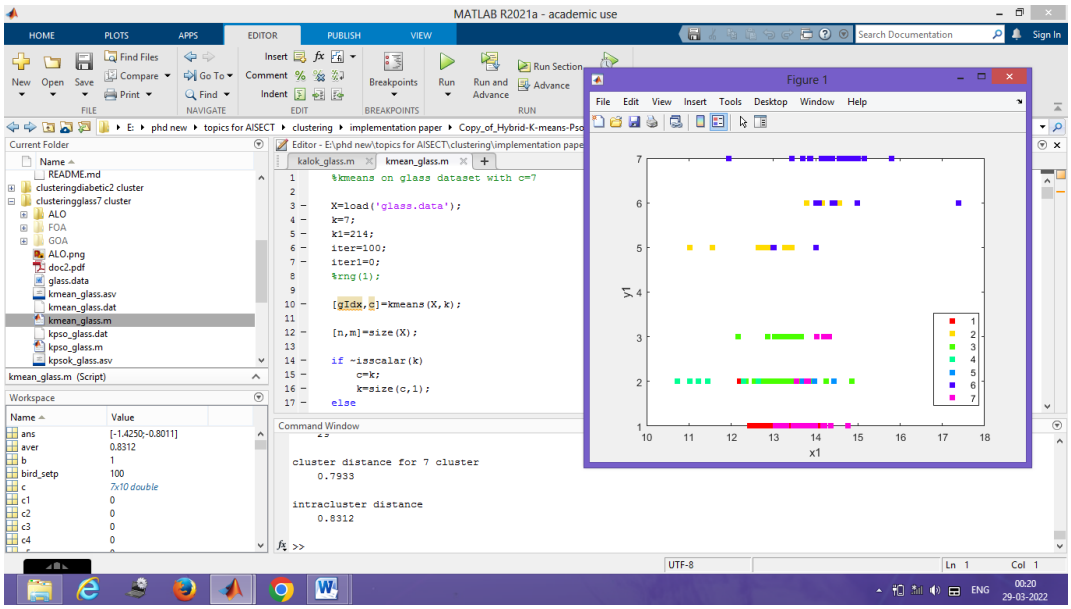

Figure 7.1 (b). Implementation in MATLAB

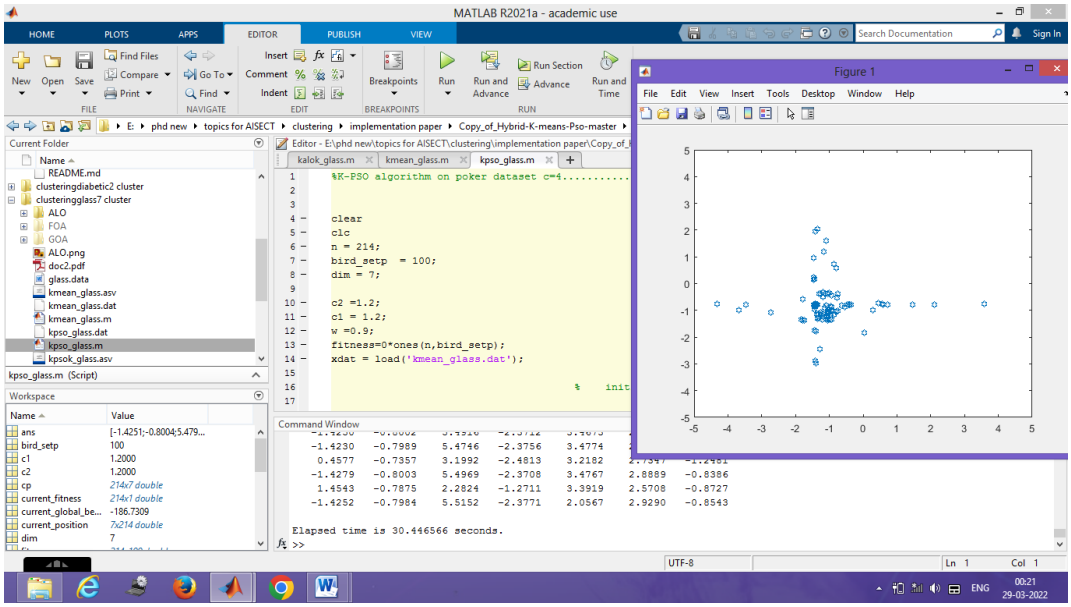

Figure 7.1 (c). Implementation in MATLAB

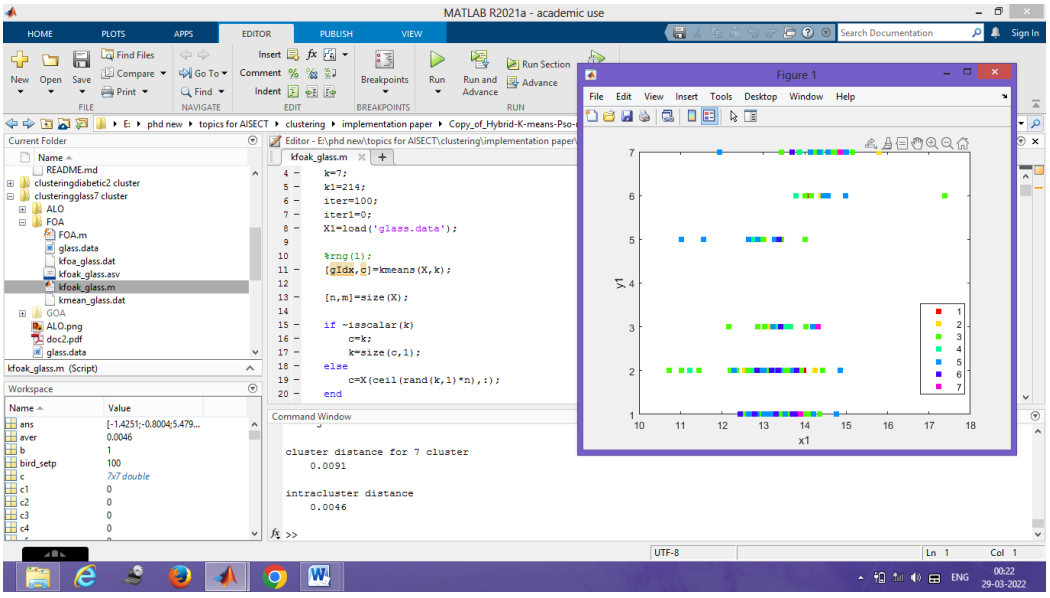

Figure 7.1 (d). Implementation in MATLAB

7.3.1. Intra-cluster Distance

Initially, the distances of each and every data components from all other data components are evaluated inside the same cluster. After that, the average value of the evaluated distances is obtained, which is known as intra-cluster distance. The minimum value of intra-cluster distance is best suitable for optimal clustering.

7.3.2. Purity Index

The data component`s recurrent clustering is indicated as purity index, which is calculated by utilizing eq. (7). The maximum value of purity index is best suitable for optimal clustering.

$$I^P = \sum_{\beta=1}^{C_n} \frac{\left(|CT_\alpha| \frac{max(|CT_{\alpha\beta}|)}{|CT_\beta|} \right)}{|S|} \quad (7)$$

Where, C_n = number of clusters,

$|CT_\beta|$ &$|CT_\alpha|$ = β^{th} cluster length and α^{th} class,

S = dataset length,

$|CT_{\alpha\beta}|$= data components of α^{th} class set to β^{th} cluster.

7.3.3. F-Measure

F-measure is used for data recovery by combining the recall (rl) and precision (prn), which is calculated by utilizing eq. 8 to 11. The maximum value of F-Measure is best suitable for optimal clustering.

$$prn(\alpha, \beta) = \frac{|CT_{\alpha\beta}|}{|CT_\beta|} \quad (8)$$

$$rl(\alpha, \beta) = \frac{|CT_{\alpha\beta}|}{|CT_\alpha|} \quad (9)$$

$$Function(\alpha, \beta) = \frac{2 \times prn(\alpha,\beta) \times rl(\alpha,\beta)}{prn(\alpha,\beta) + rl(\alpha,\beta)} \quad (10)$$

$$F_{Measure} = \sum_{\alpha=1}^{C_n} \frac{|CT_\alpha|}{|S|} max\{Function(\alpha, \beta)\} \quad (11)$$

7.3.4. Standard Deviation

The clustering power with reference to the mean values of dataset is known as standard deviation, which is obtained by utilizing eq. 12. The minimum value of standard deviation is best suitable for optimal clustering.

$$SD = \sqrt{\frac{\Sigma(dc - \overline{dc})}{|S|}} \qquad (12)$$

Here, dc = data component in dataset

$\overline{dc}$ = mean values of data components in dataset.

Table 7.2. Outcomes for Datasets

Datasets	Approaches	Performance Parameters			
		Intra-cluster distance	Purity Index	F-Measure	Standard Deviation
Glass	K-Means	0.6357	0.77	0.72	0.256874
	ACO	0.4357	0.80	0.75	0.165967
	PSO	0.2154	0.83	0.78	0.098647
	K-SSOC	0.00752	0.88	0.83	0.000853
Waveform	K-Means	0.85632	0.78	0.74	0.369875
	ACO	0.52461	0.82	0.78	0.165874
	PSO	0.09864	0.86	0.82	0.098647
	K-SSOC	0.00758	0.90	0.86	0.004527
Ionosphere	K-Means	0.75682	0.82	0.79	0.136892
	ACO	0.42586	0.87	0.84	0.086475
	PSO	0.08745	0.91	0.88	0.024517
	K-SSOC	0.00965	0.94	0.91	0.003684
Heart Failure	K-Means	10.6254	0.81	0.77	0.326547
	ACO	0.7842	0.84	0.80	0.123692
	PSO	0.5123	0.86	0.82	0.085342

	K-SSOC	0.00935	0.92	0.88	0.006584
Thyroid	K-Means	2.16824	0.84	0.80	0.126587
	ACO	0.82365	0.88	0.84	0.068574
	PSO	0.15142	0.90	0.86	0.012587
	K-SSOC	0.00868	0.94	0.90	0.003652
Wine	K-Means	1.36858	0.84	0.79	0.236541
	ACO	0.86241	0.87	0.82	0.102547
	PSO	0.23412	0.90	0.85	0.098648
	K-SSOC	0.00635	0.93	0.88	0.008756

Table 7.3. Average Ranking of all approaches for total Datasets in terms of total of intra-cluster distances mean values

Dataset	K-Means	ACO	PSO	K-SSOC
Glass	0.6357 (4)	0.4357 (3)	0.2154 (2)	0.00752 (1)
Waveform	0.85632 (4)	0.52461 (3)	0.09864 (2)	0.00758 (1)
Ionosphere	0.75682 (4)	0.42586 (3)	0.08745 (2)	0.00965 (1)
Heart Failure	10.6254 (4)	0.7842 (3)	0.5123 (2)	0.00935 (1)
Thyroid	2.16824 (4)	0.82365 (3)	0.15142 (2)	0.00868 (1)
Wine	1.36858 (4)	0.86241 (3)	0.23412 (2)	0.00635 (1)
Average Rank	4	3	2	1

The outputs in table 7.2 and table 7.3 illustrate that K-SSOC obtains superior results like greatest purity index and F-measure values and least standard deviation and intra-cluster distance against K-Means, ACO and PSO approaches for all datasets.

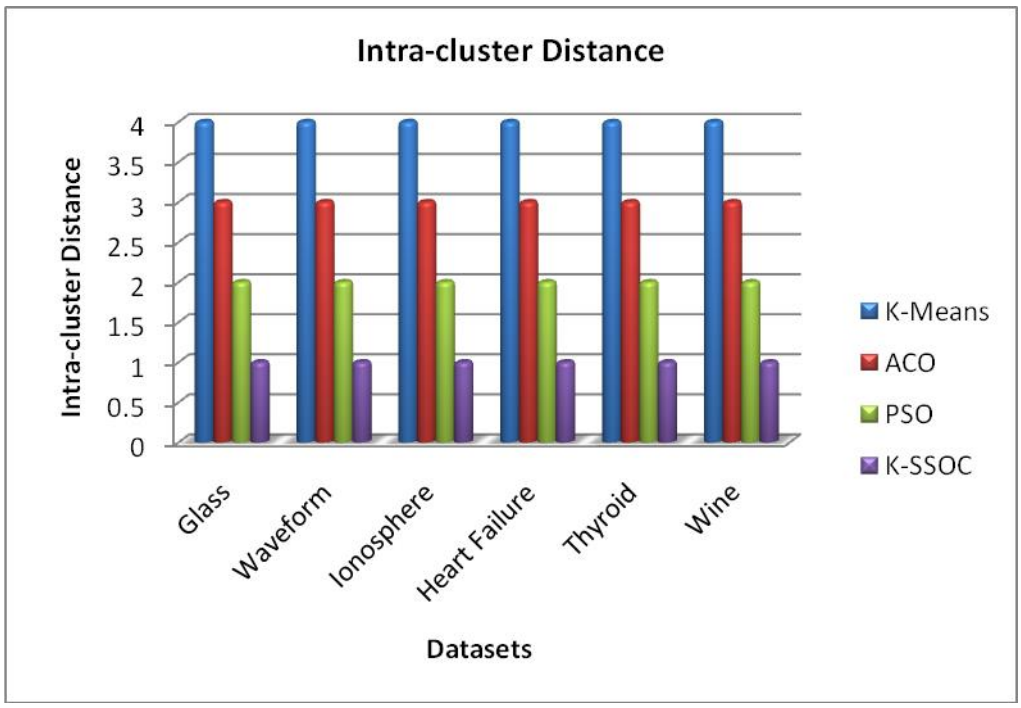

Figure 7.2. Average Rank for total six datasets in terms of Intra-cluster Distance

The outputs in figure 7.2 illustrate that ACO obtains superior results 25% than K-Means; PSO obtains superior results 34% than ACO and 50% than K-Means; K-SSOC obtains superior results 50% than PSO and 67% than ACO and 75% than K-Means in terms of intra-cluster distance for total six multidimensional datasets.

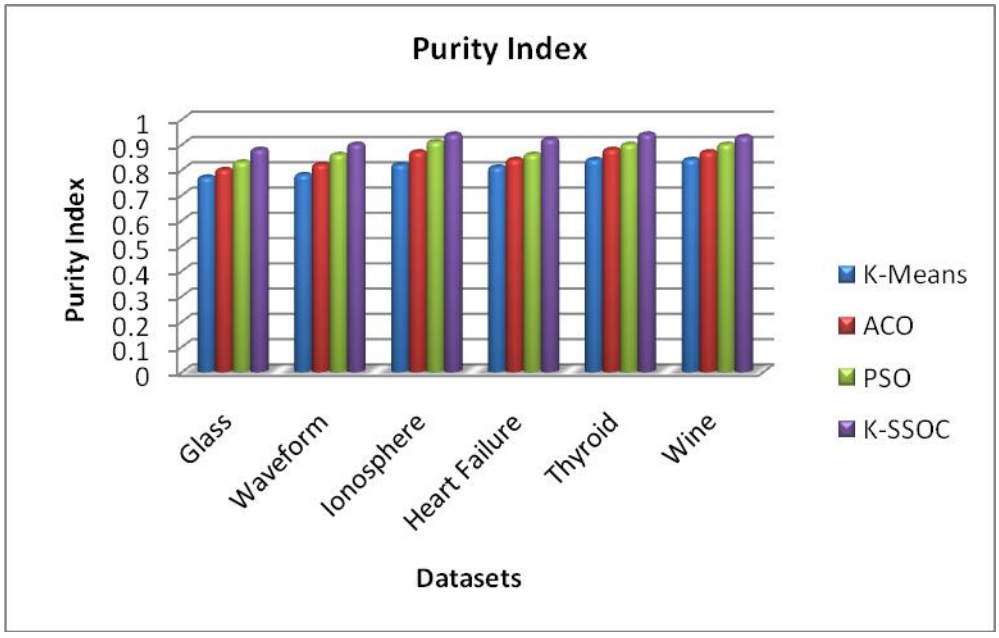

Figure 7.3. Purity Index for total six datasets

The outputs in figure 7.3 illustrate that ACO obtains superior results 8% than K-Means; PSO obtains superior results 16% than ACO and 27% than K-Means; K-SSOC obtains superior results 15% than PSO and 37% than ACO and 68% than K-Means in terms of purity index for total six multidimensional datasets.

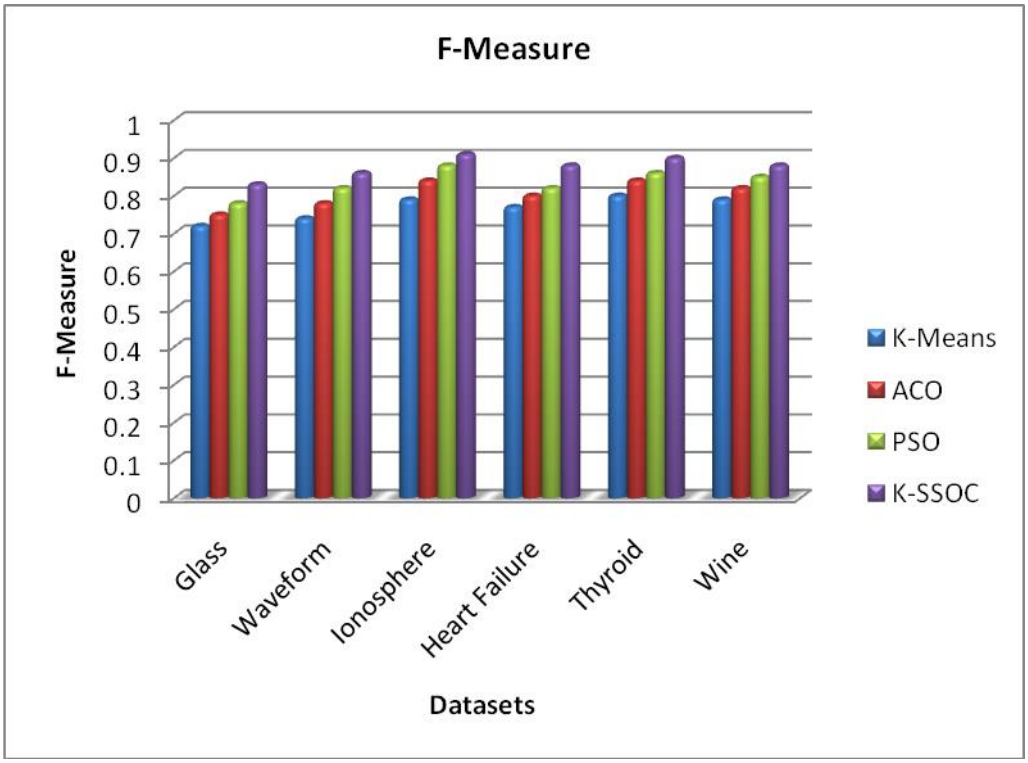

Figure 7.4. F-Measure for total six datasets

The outputs in figure 7.4 illustrate that ACO obtains superior results 11% than K-Means; PSO obtains superior results 18% than ACO and 31% than K-Means; K-SSOC obtains superior results 18% than PSO and 39% than ACO and 72% than K-Means in terms of F-Measure for total six multidimensional datasets.

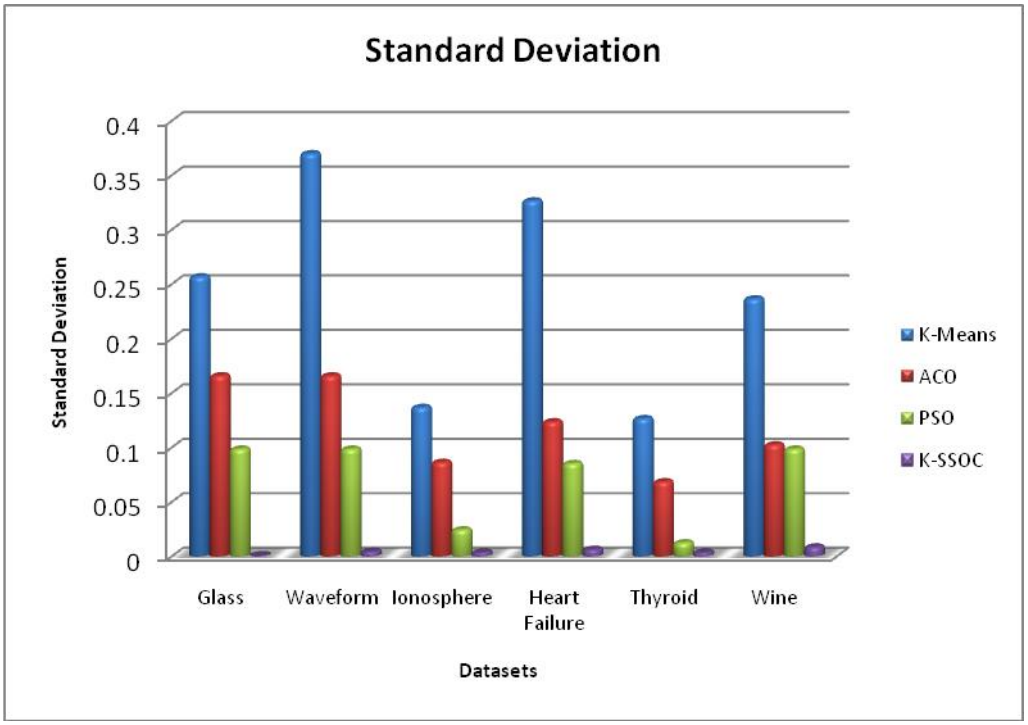

Figure 7.5. Standard Deviation for total six datasets

The outputs in figure 7.5 illustrate that ACO obtains superior results 18% than K-Means; PSO obtains superior results 27% than ACO and 43% than K-Means; K-SSOC obtains superior results 29% than PSO and 52% than ACO and 83% than K-Means in terms of standard deviation for total six multidimensional datasets.

The Figure 7.2 to Figure 7.5 represents the comparative results of K-Means, ACO, PSO and K-SSOC approaches for total six multidimensional datasets in terms of purity index, F-measure, standard deviation and intra-cluster distance. The K-SSOC improves the efficiency of K-Means by enhancing the exploitation and exploration strength of SSO approach to achieve optimum results.

7.3.5. Time Complexity

The performance of clustering approaches is also illustrated by utilizing running time complexity, which depends upon the input parameters. The input parameters are defined as: R = number of repetitions, P_s= salp`s population, and C_n = number of clusters, S = dataset, N^d = number of dimensions in dataset, N^i = number of instances.

The every step performing cost is set to be 1. Then the overall number of executions for K-SSOC is evaluated from algorithm 1 (section 7.2.1).

$$overall\ number\ of\ executions = (R+1) + R*(P_s+1) + R*P_s*C_n + R*C_n + R*C_n + R*C_n + C_n*P_s + C_n*P_s + C_n*(R+1) + R*C_n*(P_s+1) + R*C_n*P_s + R*C_n*(P_s+1) + C_n*R + C_n \quad (13)$$

$$overall\ number\ of\ executions = 4*R*P_s*C_n + 7*R*C_n + 2*C_n*P_s + R*P_s + 2*C_n + 2*R + 1 \quad (14)$$

The N^d multidimensional instances (N^i) of dataset are accessed by performing $N^d * N^i$ executions additionally. Hence, the eq. 15 is obtained by updating the eq. 14 to generate the execution cost for K-SSOC approach.

$$execution\ cost = 4*R*P_s*C_n*N^i*N^s + 7*R*C_n + 2*C_n*P_s*N^i*N^s + R*P_s + 2*C_n + 2*R + 1 \quad (15)$$

The whole input parameters are supposed to be equal, and then worst running time complexity of K-SSOC is evaluated by utilizing eq. 16.

$$execution\ cost = 4n^5 + 2n^4 + 8n^2 + 4n + 1 \quad (16)$$

The running time complexity of K-Means is $O(n^2)$, ACO is $O(n^5)$, PSO is $O(n^5)$ and K-SSOC is $O(n^5)$ in worst case. Hence, all K-Means, ACO, PSO and K-SSOC approaches are solvable in polynomial time.

7.4. Summary and Discussion

The clustering of multidimensional data is performed to enhance the OLAP model efficiency by attaining rapid query processing. Therefore, a KMeans-Salp Swarm Optimization based Clustering (K-SSOC) is implemented to overcome the K-Means limitations and improve the quality of clustering by generating optimal groups of data over the huge OLAP multidimensional dataset. The results are generated in MATLAB 2019a environment in terms of parameter purity index, standard deviation, F-measure, intra-cluster distance and running time complexity over 1000 iterations. The outcomes

represent the better quality efficiency of K-SSOC against K-Means, ACO and PSO over total six multidimensional datasets based on parameters. In future, the work will be performed over large size and unstructured datasets [87, 91] with reducing the time complexity.

CHAPTER-8
Comparative Analysis of DFOC and K-SSOC Approaches

8.1. Introduction

For computing the effectiveness of the DFOC and K-SSOC schemes explaining in former chapter, the experiment is performed to discover optimal clusters for ten multidimensional datasets. Here analysis is computed on MATLAB 2019a environment by means of 8 GB RAM, windows 8 and core i3 processor. The DFOC and K-SSOC results have been computed as compared to the earlier work such as K-Means, GA, ACO, and PSO depending on parameter purity index, standard deviation, F-measure, intra-cluster distance and running time complexity over 1000 iterations.

8.2. Multidimensional Datasets

The DFOC and K-SSOC are implemented on various multidimensional datasets representing in table 8.1.

Table 8.1. Multidimensional Datasets

Sr. No.	**Clinical Dataset**			***No. of Clusters***
	Dataset	***No. of instances***	***No. of dimensions***	
1	Cancer	683	9	2
2	Cryotherapy	90	7	2
3	Liver Patient	583	10	2
4	Heart Patients	297	14	4
5	Glass	214	10	7
6	Waveform	5000	40	3
7	Ionosphere	351	34	2

Sr. No.	Clinical Dataset			*No. of Clusters*
	Dataset	***No. of instances***	***No. of dimensions***	
8	Heart Failure	300	12	2
9	Thyroid	300	5	3
10	Wine	178	13	3

8.3. Comparative Analysis of DFOC and K-SSOC Approaches

The DFOC and K-SSOC are applied to total ten multidimensional datasets (table 8.1) on MATLAB 2019a environment. The outcomes are generated based on the parameter purity index, standard deviation, F-measure and intra-cluster distance over 1000 iterations.

Table 8.2. Average Ranking (Intra-cluster Distance) for Multidimensional Datasets

Dataset	K-Means	GA	ACO	PSO	DFOC	K-SSOC
Cancer	94.2641 (6)	0.3265 (5)	0.08587 (3)	0.1786 (4)	0.002514 (2)	0.001036 (1)
Cryotherapy	19.3625 (6)	0.3142 (5)	0.0541 (3)	0.2576 (4)	0.00325 (2)	0.00257 (1)
Liver Patient	42.3214 (6)	0.4201 (5)	0.0845 (3)	0.1257 (4)	0.00464 (2)	0.00136 (1)
Heart Patients	12.3654 (6)	0.50241 (5)	0.0365 (3)	0.3647 (4)	0.00124 (2)	0.00096 (1)
Glass	0.6357 (6)	0.3614 (4)	0.4357 (5)	0.2154 (3)	0.01085 (2)	0.00752 (1)
Waveform	0.85632	0.1574	0.52461	0.09864	0.02631	0.00758

	(6)	(4)	(5)	(3)	(2)	(1)
Ionosphere	0.75682 (6)	0.6215 (5)	0.42586 (4)	0.08745 (3)	0.01025 (2)	0.00965 (1)
Heart Failure	10.6254 (6)	0.5125 (4)	0.7842 (5)	0.5123 (3)	0.07456 (2)	0.00935 (1)
Thyroid	2.16824 (6)	0.6876 (4)	0.82365 (5)	0.15142 (3)	0.02657 (2)	0.00868 (1)
Wine	1.36858 (6)	0.5241 (4)	0.86241 (5)	0.23412 (3)	0.03654 (2)	0.00635 (1)
Average Ranking	6	4.5	4.1	3.4	2	1

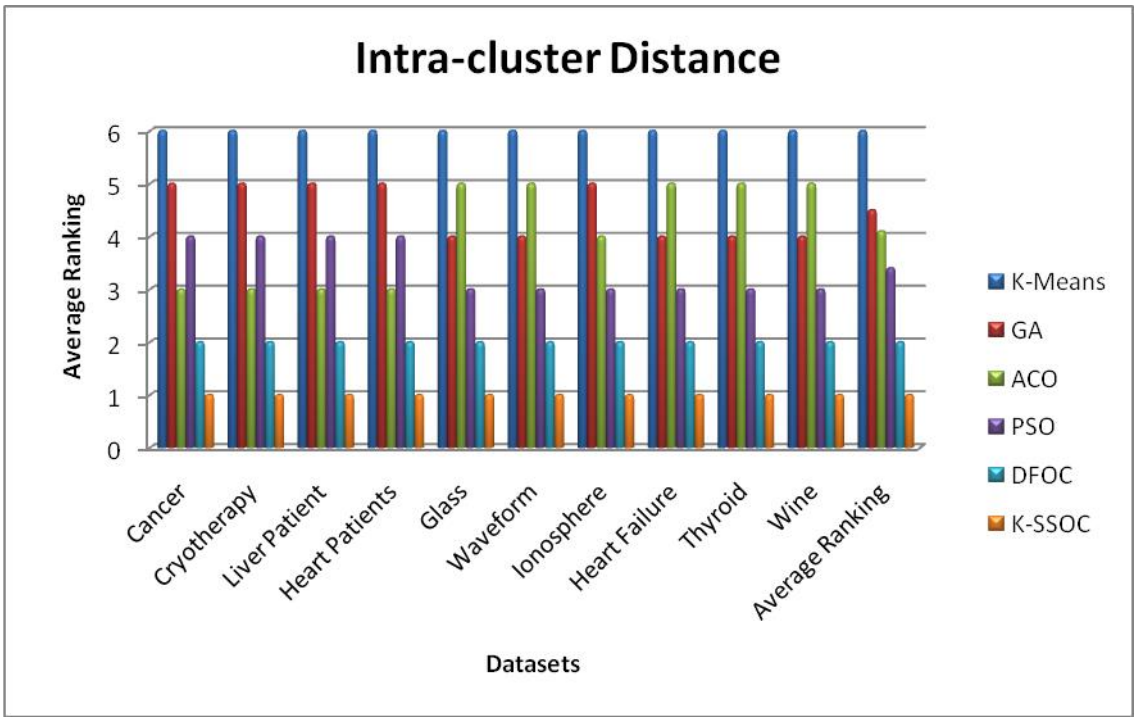

Figure 8.1. Average Rank for total ten datasets in terms of Intra-cluster Distance

The outputs in table 8.2 and figure 8.1 illustrate that GA obtains superior results 25% than K-Means; ACO obtains superior results 9% than GA and 32% than K-Means; PSO

obtains 17% than ACO and 25% than GA and 44% than K-Means; DFOC obtains superior results 42% than PSO and 52% than ACO and 56% than GA and 67% than K-Means; K-SSOC obtains superior results 50% than DFOC and 71% than PSO and 76% than ACO and 78% than GA and 84% than K-Means in terms of intra-cluster distance for total ten multidimensional datasets.

Table 8.3. F-Measure for Multidimensional Datasets

Dataset	K-Means	GA	ACO	PSO	DFOC	K-SSOC
Cancer	0.84	0.85	0.87	0.86	0.92	0.93
Cryotherapy	0.76	0.85	0.86	0.85	0.90	0.91
Liver Patient	0.85	0.86	0.87	0.86	0.88	0.90
Heart Patients	0.74	0.76	0.80	0.78	0.95	0.96
Glass	0.72	0.76	0.75	0.78	0.81	0.83
Waveform	0.74	0.79	0.78	0.82	0.85	0.86
Ionosphere	0.79	0.85	0.84	0.88	0.89	0.91
Heart Failure	0.77	0.81	0.80	0.82	0.86	0.88
Thyroid	0.80	0.82	0.84	0.86	0.88	0.90
Wine	0.79	0.83	0.82	0.85	0.87	0.88

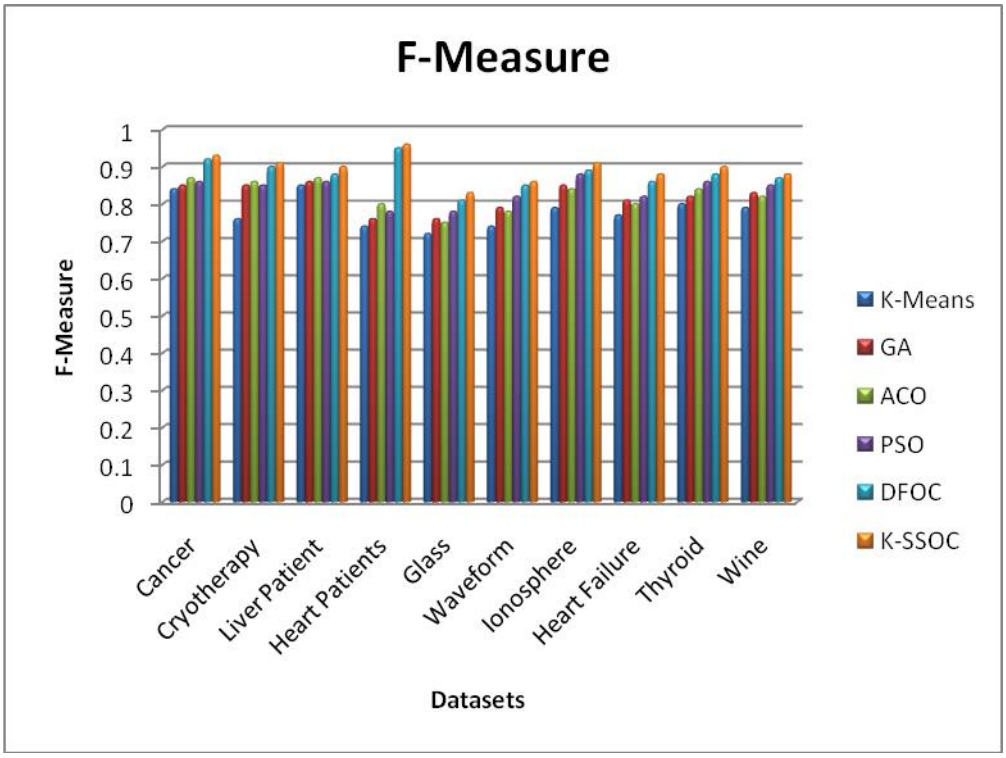

Figure 8.2. F-Measure for total ten Multidimensional datasets

The outputs in table 8.3 and figure 8.2 demonstrate that GA evaluates advanced outcomes 5% than K-Means; ACO evaluates advanced outcomes 4% than GA and 7% than K-Means; PSO evaluates advanced outcomes 4% than ACO and 6% than GA and 9% than K-Means; DFOC evaluates advanced outcomes 7% than PSO and 11% than ACO and 13% than GA and 17% than K-Means; K-SSOC evaluates advanced outcomes 3% than DFOC and 8% than PSO and 12% than ACO and 15% than GA and 19% than K-Means in terms of F-Measure for total ten multidimensional datasets.

Table 8.4. Purity Index for Multidimensional Datasets

Dataset	K-Means	GA	ACO	PSO	DFOC	K-SSOC
Cancer	0.86	0.87	0.90	0.88	0.95	0.96
Cryotherapy	0.82	0.90	0.91	0.90	0.95	0.96
Liver Patient	0.87	0.88	0.90	0.89	0.91	0.93
Heart	0.78	0.81	0.84	0.83	0.91	0.92

Patients						
Glass	0.77	0.81	0.80	0.83	0.86	0.88
Waveform	0.78	0.83	0.82	0.86	0.89	0.90
Ionosphere	0.82	0.88	0.87	0.91	0.92	0.94
Heart Failure	0.81	0.85	0.84	0.86	0.91	0.92
Thyroid	0.84	0.86	0.88	0.90	0.92	0.94
Wine	0.84	0.85	0.87	0.90	0.92	0.93

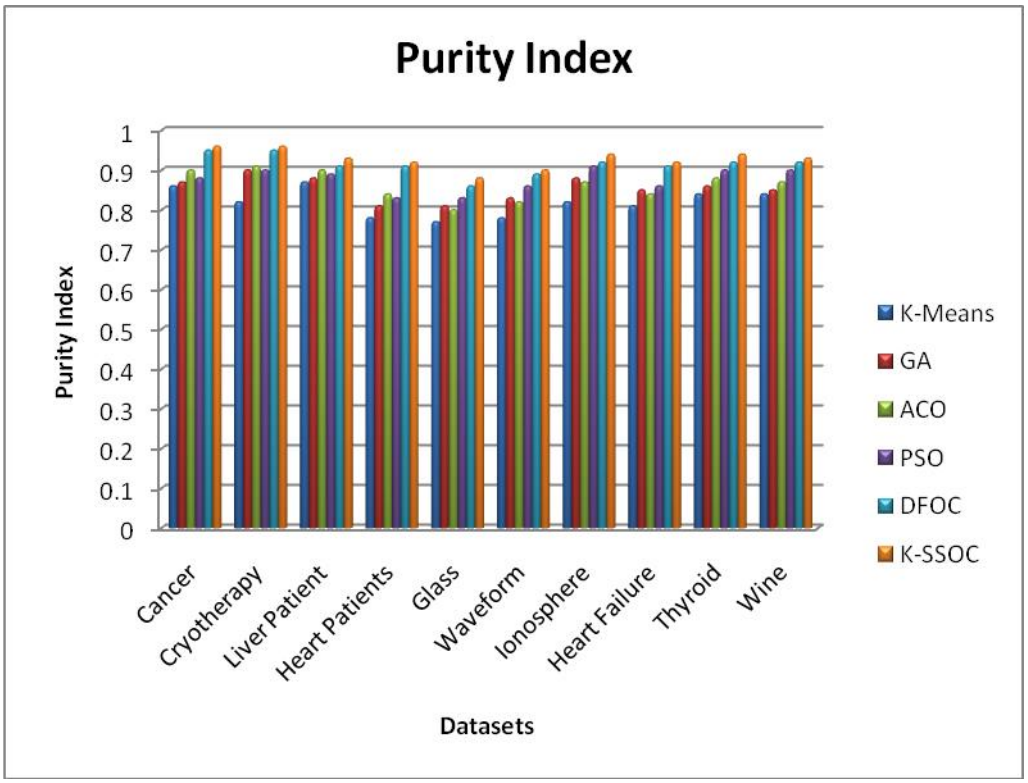

Figure 8.3. Purity Index for total ten Multidimensional datasets

The outputs in table 8.4 and figure 8.3 demonstrate that GA calculates improved outputs 4% than K-Means; ACO calculates improved outputs 3% than GA and 6% than K-Means; PSO calculates improved outputs 3% than ACO and 7% than GA and 9% than K-Means; DFOC calculates improved outputs 4% than PSO and 8% than ACO

and 10% than GA and 16% than K-Means; K-SSOC calculates improved outputs 5% than DFOC and 11% than PSO and 13% than ACO and 19% than GA and 21% than K-Means in terms of purity index for total ten multidimensional datasets.

Table 8.5. Standard Deviation for Multidimensional Datasets

Dataset	K-Means	GA	ACO	PSO	DFOC	K-SSOC
Cancer	0.5248	0.2153	0.1042	0.1587	0.024	0.011
Cryotherapy	0.3521	0.1241	0.0624	0.1245	0.00786	0.00157
Liver Patient	0.4215	0.2641	0.0758	0.1678	0.00882	0.00236
Heart Patients	0.20365	0.07548	0.02364	0.05671	0.0074	0.0023
Glass	0.256874	0.25716	0.165967	0.098647	0.0089	0.000853
Waveform	0.369875	0.10452	0.165874	0.098647	0.00697	0.004527
Ionosphere	0.136892	0.1287	0.086475	0.024517	0.00897	0.003684
Heart Failure	0.326547	0.03697	0.123692	0.085342	0.00826	0.006584
Thyroid	0.126587	0.17963	0.068574	0.012587	0.00741	0.003652
Wine	0.236541	0.2157	0.102547	0.098648	0.01582	0.008756

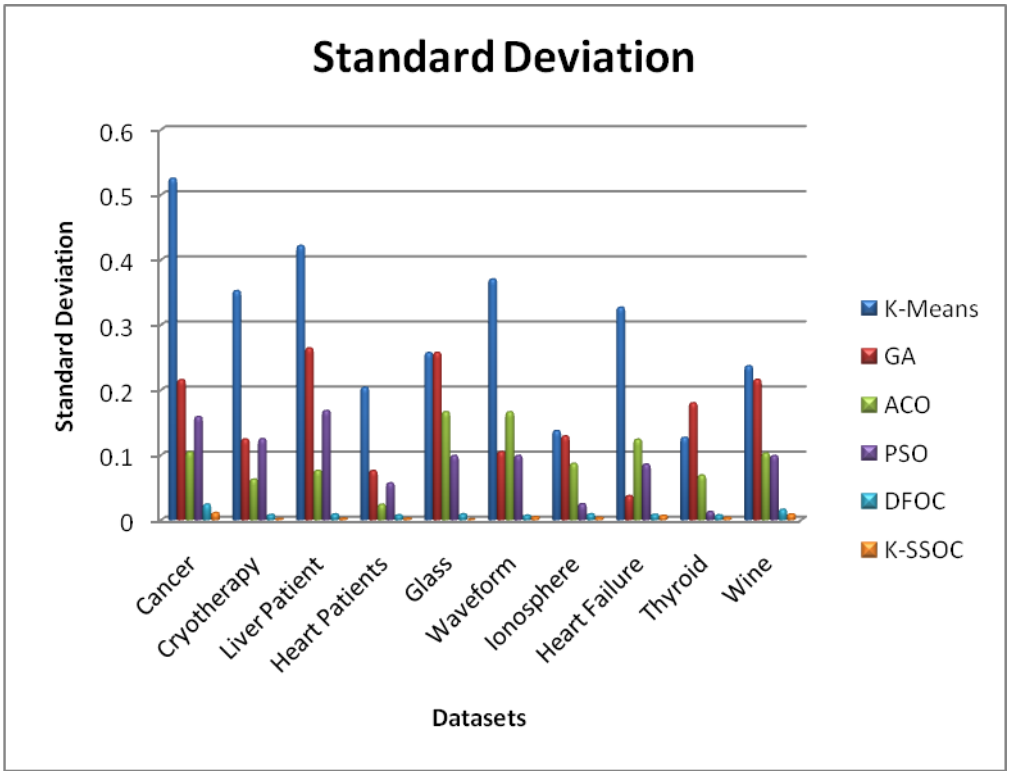

Figure 8.4. Standard Deviation for total ten Multidimensional datasets

The outputs in table 8.5 and figure 8.4 demonstrate that GA computes improved results 73% than K-Means; ACO computes improved results 54% than GA and 78% than K-Means; PSO computes improved results 43% than ACO and 61% than GA and 83% than K-Means; DFOC computes improved results 36% than PSO and 51% than ACO and 72% than GA and 86% than K-Means; K-SSOC computes improved results 28% than DFOC and 47% than PSO and 62% than ACO and 79% than GA and 91% than K-Means in terms of standard deviation for total ten multidimensional datasets.

The Figure 8.1 to Figure 8.4 and Table 8.2 to Table 8.5 represent the comparative results of K-Means, GA, ACO, PSO, DFOC and K-SSOC approaches for total ten multidimensional datasets in terms of purity index, F-measure, standard deviation and intra-cluster distance. The first scheme DFOC generates improved results than K-Means, GA, ACO, and PSO schemes for all ten multidimensional datasets. The second scheme K-SSOC improves the efficiency of K-Means by enhancing the exploitation and exploration strength of SSO approach to achieve optimum results. So, K-SSOC obtains advanced results than DFOC and prior schemes PSO, ACO, GA and K-Means for total ten multidimensional datasets.

8.4. Time Complexity

The performance of clustering approaches is also illustrated by utilizing running time complexity, which depends upon the input parameters. The running time complexity of K-Means is $O(n^2)$, GA is $O(n^5)$, ACO is $O(n^5)$, PSO is $O(n^5)$, DFOC is $O(n^4)$ and K-SSOC is $O(n^5)$ in worst case. Hence, all K-Means, GA, ACO, PSO, DFOC and K-SSOC approaches are solvable in polynomial time.

8.5. Summary and Discussion

In this chapter, a Dragon Fly Optimization based Clustering (DFOC) approach is implemented to improve the performance of data clustering by obtaining optimized clusters from multidimensional datasets for OLAP. The outcomes are examined on MATLAB 2019a tool and illustrated the superior efficiency of DFOC over ten multidimensional datasets as compared to prior approaches ACO, GA and K-Means in terms of intra-cluster distance, purity index, F-measure, and standard deviation.

The clustering of multidimensional data is performed to enhance the OLAP model efficiency by attaining rapid query processing. Therefore, next approach KMeans-Salp Swarm Optimization based Clustering (K-SSOC) is implemented to overcome the K-Means limitations and improve the quality of clustering by generating optimal groups of data over the huge OLAP multidimensional dataset. The results are generated in MATLAB 2019a environment in terms of parameter purity index, standard deviation, F-measure, intra-cluster distance and running time complexity over 1000 iterations. The outcomes represent the better quality efficiency of K-SSOC against K-Means, GA, ACO, PSO and DFOC over total ten multidimensional datasets based on parameters. In future, the work will be performed over large size and unstructured datasets with reducing the time complexity.

CHAPTER-9
CONCLUSION AND FUTURE DIRECTION

9.1. Conclusions

In this work, FFO based optimal materialized cube selection is performed on lattice structure with multidimensional data over OLAP framework. The results are evaluated on multidimensional data in terms of frequency and number of dimensions. The analysis of performance of FFO illustrates the improved quality, efficiency of FFO to reduce the OLAP query processing expenditure as compared to PSO. In the future, several optimization approaches will be implemented for optimized cube selection by considering the time complexity as a factor.

In this paper, GWO is introduced to choose the best materialized cube utilizing the OLAP multidimensional information model with lattice structure. The outcomes are calculated on data having various dimensions based on total dimensions and frequency. The GWO is examined over lattice structure to find out the optimal data cube for minimizing the query dispensation expenses. Various optimization strategies will be introduced to perform an optimal selection of data cubes with more dimensions and evaluating the time and space complexity as performance indicators in the future.

In this work, a Dragon Fly Optimization based Clustering (DFOC) approach is implemented to improve the performance of data clustering by obtaining optimized clusters from multidimensional clinical data for OLAP. The outcomes are examined on MATLAB 2019a tool and illustrated the superior efficiency of DFOC as compared to prior approaches ACO, GA and K-Means in terms of intra-cluster distance, purity index, F-measure, and standard deviation.

The clustering of multidimensional data is performed to enhance the OLAP model efficiency by attaining rapid query processing. Therefore, a KMeans-Salp Swarm Optimization based Clustering (K-SSOC) is implemented to overcome the K-Means limitations and improve the quality of clustering by generating optimal groups of data over the huge OLAP multidimensional dataset. The results are generated in MATLAB 2019a environment in terms of parameter purity index, standard deviation, F-measure,

intra-cluster distance and running time complexity over 1000 iterations. The outcomes represent the better quality efficiency of K-SSOC against K-Means, ACO and PSO over total six multidimensional datasets based on parameters.

9.2. Future Direction

The extension of this work can be further performed in numerous directions:

1. In future, the work can be performed over large size with reducing the time complexity.
2. The clustering can be performed on unstructured datasets in future.
3. The capability of K-SSOC and DFOC can be improved by utilizing some entropy and chaos theory concerns in future.
4. The data selection and clustering speed can be further improved.

REFERENCES

1. A. Abraham, S. Das and S. Roy, “Swarm Intelligence Algorithm for Data Clustering”, Norwegian University of Science and Technology, Trondheim, Norway, pp. 279-313, 2007.
2. A. Sabnis, “Hybrid Clustering with Application to Web Pages”, Master`s Projects, San Jose State University, pp. 1-44, 2009.
3. A. Rathee and J. K. Chhabra, “Improving Cohesion of a Software System by Performing Usage Pattern Based Clustering”, 6th International Conference on Smart Computing and Communications (ICSCC), Procedia Computer Science, Elsevier, Kurukshetra, India, pp. 740-746, 7-8 December 2017.
4. A. K. Jain, R. Chitta and R. Jin, “Clustering Big Data”, Department of Computer Science, Michigan State University, pp. 1-44, 2012.
5. A. K, and M. K Nair, “FGPSO - A Novel Algorithm for Multi Objective Data Clustering”, WSEAS Transactions on Computers, Vol. 17, pp-1-9, 2018.
6. A. K. D., U. T. A. and S. C., “A Comparative Study on K-Means And Genetic Algorithm For Data Clustering”, International Journal of Engineering Research and Development, Vol. 12, Issue 11, pp. 1-9, 2016.
7. A. Ekbal, S. Saha, D. Moll´a and K. Ravikumar, “Multi-Objective Optimization for Clustering of Medical Publications”, In Proceedings of Australasian Language Technology Association Workshop, pp-53-61, 2013.
8. A. Gosain and Heena, “Materialized Cube Selection using Particle Swarm Optimization algorithm”, 7th International Conference on Communication, Computing and Virtualization, Elsevier, pp. 2-7, 2016.
9. A. Vaisman and E. Zimanyi, “Mobility Data Warehouses. International Journal of Geo-Information”, MDPI, 8 (170), pp. 1-22, 2019.
10. A. Papacharalampopoulos, C. Giannoulis, P. Stavropoulos, and D. Mourtzis, “A Digital Twin for Automated Root-Cause Search of Production Alarms Based on KPIs Aggregated from IoT”, Applied Science, MDPI, 10 (2377), pp. 1-16, 2020.

11. A. Tsois, N. Karayannidis and T. Sellis, “MAC: Conceptual Data Modeling for OLAP”, International Workshop on Design and Management of Data Warehouses, pp. 1-12, 2001.
12. A. S. Maniatis, “OLAP Presentation Modeling with UML and XML”, pp. 1-10, 2005.
13. A. Tripathy, L. Mishra, P. K. Patra; “A multi dimensional design framework for querying spatial data using concept lattice”; Proc. of the 2nd International Advance Computing Conference (IACC); pp. 394 – 399, 2010.
14. B. Noh, J. Son, H. Park, and S. Chang, “In-Depth Analysis of Energy Efficiency Related Factors in Commercial Buildings Using Data Cube and Association Rule Mining”, Sustainability, MDPI, 9 (2119), pp. 1-20, 2017.
15. B. W. Jo, R. M. A. Khan, Y. S. Lee, J. H. Jo and N. Saleem, “A Fiber Bragg Grating-Bsed Condition Monitoring and Early Damage detection System for the Structural Safety of Underground Coal Mines Using the Internet of Things”, Journal of Sensors, Hindawi, pp. 1-16, 2018.
16. C. Ciferri, R. Ciferri, L. Gomez, M. Schneider, A. Vaisman, and E. Zimanyi, “Cube Algebra: A Generic User-Centric Model and Query Language for OLAP Cubes”, International Journal of Data Warehousing and Mining, pp. 1-23, 2012.
17. C. J. Lee, C. C. Hsu and D. R. Chen, “A Hierarchical Document Clustering Approach with Frequent Itemsets”, International Journal of Engineering and Technology, Vol., No. 2, pp. 1-5, 2017.
18. C. F. Tsai, H. C. Wu and C. W. Tsai, “A new Data Clustering Approach fo Data Mining in Large Databases”, proceedings of the International Symposium on Parallel Architecture, Algorithms and Networks (ISPAN), Computer Society, IEEE, pp. 1-6, 2002.
19. C. Sreedhar, N. Kasiviswanath and P. C. Reddy, “Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop”, Journal of Big Data, Vol. 27, pp-1-19, 2017.
20. C. Gong, H. Chen, W. He, and Z. Zhang, “Improved multi-objective clustering algorithm using particle swarm optimization”, PLOS ONE, pp-1-19, 2017. (https://doi.org/10.1371/journal.pone.0188815)

21. C. T, K. N, G. S, and P. A, “Challenges in Big Data Clustering for Data Analytics”, IJARMATE, pp. 230-232, 2016.
22. D. R. D. Almeida, C. D. S. Baptista, F. G. D. Andrade, and A. Soares, “A Survey on Big Data for Trajectory Analytics”, International Journal of Geo-Information, MDPI, 9 (88), pp. 1-24, 2020.
23. D. Camilovic, D. B. Vujaklija and N. Gospic, “A Call Detail Records Data Mart: Data Modeling and OLAP Analysis”, ComSIS, Vol. 6, No. 2, pp. 87-110, 2009.
24. D. Bulos and S. Forsman, “OLAP Database Design”, Symmetry Corporation, pp. 1-19, 2019.
25. E. Emmanuel, A. Obiageli and V. Osinachi, “Design and Implementation of Multidimensional Students Result Analytical Processing for Tertiary Institutions”, International Journal of Engineering and Computer Science, Vol. 8, Issue 8, pp. 24814-24828, 2019.
26. E. Naaz, D. Sharma, D. Sirisha, and V. M, “Enhanced K-Means Clustering Approach for Health Care Analysis Using Clinical Documents”, International Journal of Pharmaceutical and Clinical Research (IJPCR), Vol. 8, No. 1, pp-60-64, 2016.
27. G. Agapito, C. Zucco, and M. Cannataro, “COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data”, International Journal of Environment Research and Public Health, MDPI, 17 (5596), pp. 1-22, 2020.
28. G. A. Kiran, M. Puri and S. S. Suresh, “PSO-Enabled Privacy Preservation of Data Clustering”, Indian Journal of Science and Technology, Vol. 10, No. 11, pp-1-10, 2017. (DOI: 10.17485/ijst/2017/v10i11/89318)
29. G. Ravali, P. S. Moulika, A. A. Parna, J. Abhinaya and T. Anuradha, “Analysing YouTube Data Using K-Means Clustering”, International Journal of Emerging Trends of Technology in Computer Science (IJETICS), Vol. 6, Issue 2, pp. 62-66, 2017.
30. G. Bathla, H. Aggarwal and R. Rani, “A Novel Approach for Clustering Big Data based on MapReduce”, International Journal of Electrical and Computer Engineering (IJECE), Vol. 8, No. 3, pp. 1711-1719, 2018.

31. G. Miao, J. Tatemura, W. P. Hsiung, A. Sawires and L. E. Moser, “Extracting Data Records from the Web Using Tag Path Clustering”, International World Wide Web Conference Committee (IW3C2), Madrid, Spain, pp. 981-991, 2009.
32. H. R., and C. A.; “Abstract Interpretation of Database Query Languages”, Journal of Computer Languages, Systems and Structures Vol. 38 Issue 2, pp. 123-157, 2012.
33. H. Jinghua, Y. Mei, L. Xiaowei, and S. Xinna; “The Design and Implementation of MDSS Based on Data Warehouse”; Proc. of the 1st International Conference on Computing, Control and Industrial Engineering (CCIE); pp.42-45,2010.
34. H. Bangui, M. Ge, and B. Buhnova, “Exploring Big Data Clustering Algorithms for Internet of Things Applications”, In Proceedings of the 3rd International Conference on Internet of Things, Big Data and Security (IoTBDS), pp-269-276, 2018.
35. H. R. Tzhoosh, “Opposition Based Learning: A New Scheme for Machine Intelligence”, International Conference on Computational Intelligence for Modelling, Control and Automation, IEEE, pp-1-8, 2005.
36. I. L. Cruz, R. Berlanga, and M. J. Aramburu, “Modelling Analytical Streams for Social Business Intelligence”, Informatics, MDPI, 5 (53), pp. 1-17, 2018.
37. J. G. Aher and V. A. Metre, “PSO based Multidimensional Data Clustering: A Survey”, International Journal of Computer Applications, Vol. 87, No. 16, pp. 41-48, 2014.
38. J. G. Aher and V. A. Metre, “Clustering Multidimensional Data with PSO based Algorithm”, Soft Computing and Artificial Intelligence, pp. 1-6, 2014.
39. J. N. S. Rubi and P. R. L. Gondim, “IoMT Platform for Pervasive Healthcare Data Aggregation, Processing, and Sharing Based on OneM2M and OpenEHR”, Sensors, MDPI, 19 (4283), pp. 1-25, 2019.
40. J. L. S. Cervantes, M. Radzimski, C. A. R. Enriquez, G. A. Hernandez, L. R. Mazahua, C. S. Ramirez, and A. R. Gonzalez, “SREQP: A Solar Radiation Extraction and Query Platform for the Production and Consumption of Linked Data from Weather Stations Sensors”, Journal of Sensors, Hindawi, pp. 1-19, 2016.

41. J. Loureiro and O. Belo, “A Discrete Particle Swarm Algorithm for OLAP Data Cube Selection”, 8th International Conference on Enterprise Information Systems-DISI, pp. 46-53, 2006.
42. J. Nasiri and F. M. Khiyabani, “A whale optimization algorithm (WOA) approach for clustering”, Cogent Mathematics & Statistics, Vol. 5, pp-1-13, 2018. (https://doi.org/10.1080/25742558.2018.1483565)
43. J. P. Kilborn, D. L. Jones, E. B. Peebles and D. F. Naar, “Resemblance profiles as Clustering decision criteria: Estimating statistical power, error and correspondence for a hypothesis test for multivariate structure”, Ecology and Evolution, Wiley Publication, pp. 2039-2057, 2017.
44. J. Gupta and A. Mahajan, “BPSO Optimized K means Clustering Approach for Medical data Analysis”, International Journal of Scientific Research Engineering & Technology (IJSRET), Vol. 4, Issue 8, pp.822-825, 2015.
45. J. Yi, Y. Zhang, X. Zhao and J. Wan, “A Novel Text Clustering Approach Using Deep Learning Vocabulary Network”, Mathematical Problems in Engineering, Hindawi, pp. 1-14, 2017.
46. K. Singhal and R. Grover, “A Detailed Approach for Data Mining and Clustering of Unstructured Data using R”, International Journal of Allied Practice, Research and Review (IJAPRR), Vol. 5, Issue 1, pp. 28-45, 2018.
47. K. K. Mohan, K. P. S. Reddy, K. G. Sri, A. P. Deva and M. Sundarababu, “Efficient Big Data Processing in Haddop MapReduce”, Special Issue on 5th National Conference on Recent Trends in Information Technology, P.V.P. Siddhartha Institute of Technology Kanuru, Vijayawada, India, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 6, Issue 3, pp. 1-5, 2016.
48. L. Shen, S. Liu, S. Chen, and X. Wang, “The Application Research of OLAP in Police Intelligence Decision System” International Workshop on Information and Electronics Engineering (IWIEE), Elsevier, 29, pp. 1-6, 2012.
49. L. Celardo, D. F. Iezzi and M. Vichi, “Multi mode partitioning for text clustering to reduce dimensionality and noises”, 13th Journees internationals d`Analyse statistique des Donnees Textuelles (JADT), pp.1-11, 2016.
50. M. Eltabakh, “OLAP & Data Mining”, Spring, WPI, pp. 1-38, 2012.

51. M. R. Murty, A. Naik, J. V. R. Murthy, P. V. G. D. P. Reddy, S. C. Satapathy and K. Parvathi, “Automatic Clustering Using Teaching Learning Based Optimization”, Applied Mathematics, Vol. 5, pp. 1202-1211, 2014. (http://dx.doi.org/10.4236/am.2014.58111)
52. M. Lashkari and M. H. Moattar, “Improved COA with Chaotic Initialization and Intelligent Migrattion for Data Clustering”, Journal of AI and Data Mining (JAIDM), Vol. 5, No. 2, pp. 293-305, 2017.
53. M. R. Thakare, S. W. Mohod, A. N. Thakare, “Clustering of Big Data Using Different Data Mining Techniques”, International Research Journal of Engineering and Technology (IRJET), Vol. 3, Issue 1, pp. 1088-1094, 2016.
54. M. Kaur and N. Kaur, “Text Clustering using PBO algorithm for Analysis and Optimization”, International Journal of Current Engineering and Technology, Vol. 4, No. 6, pp. 3876-3878, 2014.
55. M. Lashkari and A. Rostami, “Extended PSO Algorithm For Improvement Problems K-Means Clustering Algorithm”, International Journal of Managing Information Technology (IJMIT), Vol.6, No.3, pp-1-13, 2014.
56. M. Hosseini, M. Sadeghzade and R. Nourmandipour, “An efficient approach based on different evolution algorithm for data clustering”, Decision Science Letters, pp. 319-324, 2014.
57. M. Chen, S. A. Ludwig and K. Li, “Clustering in Big Data”, Big Data Management and Processing, pp. 333-347, 2107.
58. M. A. Nemmich, F. Debbat, and M. Slimane, “A Data Clustering Approach Using Bees Algorithm with a Memory Scheme”, Springer Nature Switzerland AG, pp-261–270, 2019. (https://doi.org/10.1007/978-3-319-98352-3_28)
59. M. Ghesmoune, M. Lebbah and H. Azzag, “State of the art on clustering data”, Big Data Analytics, Vol. 1, pp. 1-27, 2016.
60. V. A. Metre and P. B. Deshmukh, “Scope of Research on Particle Swarm Optimization Based Data Clustering”, International Journal of Computer Science Trends and Technology (IJCST), Vol. 6, Issue 6, pp-87-93, 2018.
61. N. Niraula, “Web Log data Analysis: Converting Unstructured Web Log Data into Structured Data Using Apache Pig”, Saint Cloud state University, pp. 1-54, 2014.

62. N. Haghtalab, “Clustering in the Presence of Noise”, University of Waterloo, Ontario, Canada, pp. 1-68, 2013.
63. N. G. Blas and O. L. Tolic, “Clustering using Particle Swarm Optimization”, International Journal Information theories and Applications, Vol.23, No. 1, pp. 24-33, 2016.
64. N. Jukic, B. Jukic, and Malliaris, “Online Analytical Processing (OLAP) for Decision Support”, pp. 1-25, 2008.
65. N. Stefanovic, “Proactive Supply Chain Performance Management with Predictive Analytics”, the Scientific World Journal, Hindawi, pp. 1-18, 2014.
66. O. Kurasova, V. Marcinkevicius, V. Medvedev, A. Rapecka, and P. Stefanovic, “Strategies for Big Data Clustering”, 26th International Conference on Tools Artificial Intelligence (IEEE) pp. 740- 747, 2014.
67. P. Maniriho and A. Effendi, ”Examining the Performance of K-Means Clustering Algorithm” International Journal of Research in Engineering, Science and Management(IJRESM) Vol.1, Issue-3, pp. 1-5, 2018.
68. P. and P. Khanchi, “GSA Optimized Agglomerative Clustering” International Journal of Institutional & Industrial Research, vol.1, Issue.3, pp. 6-11, 2016.
69. P. Rigollet and D. Shah, “Data Science: Data to Insights” LECTURE TRANSCRIPTS MIT Professional Education, pp.1-340.
70. P. Lama, “CLUSTERING SYSTEM BASED ON TEXT MINING USING THE K- MEANS ALGORITHM” pp. 1-47 2013.
71. P. Nerurkar, A. Shirke, M. Chandane, and S. Bhiru, “A Novel Heuristic for Evolutionary Clustering “, 6th International Conference on Smart Computing and Communications, ICSCC, Kurukshetra, India, pp-780-789, 2017.
72. P. Raut, and N. Khochare, “Web Document Clustering System Using Fuzzy Logic and Feature extraction” International Research Journal of Engineering and Technology (IRJET), Vol.3, Issue.6, pp. 2057-2061, 2016.
73. P. Nerurkar, A. Shirke, M. Chandane, and S. Bhirud, “Empiricial Analysis of Data Clustering Algorithms” 6th International Conference on Smart Computing and Communications (ICSCC), pp.770-779, 2017.
74. P. Vaijayanthi, X. S. Yang, N. A. M and R. Murugadoss, “High Dimensional Data Clustering Using Cuckoo Search Optimization Algorithm”, International

Journal of Advanced Computer Engineering and Communication Technology (IJACECT), Vol. 3, Issue 3, pp-1-5, 2014.

75. P. Westerlund, “Business Intelligence: Multidimensional Data Analysis”, ECTS Credits, pp. 1-58, 2008.
76. R. P. Dagde, and S. Dongre, “A Review on Clustering Analysis based on Optimization Algorithm for Data mining” International Journal of Computer Science and Network(IJCSN), Vol.6, Issue.1, pp.36-41, 2017.
77. R. K. Saidala, and N. Devarakonda, “Multi-Swarm Whale Optimization Algorithm for Data Clustering Problems using Multiple Cooperative Strategies”, I.J. Intelligent Systems and Applications, MECS, Vol. 8, pp-36-53, 2018.
78. S. P. Navia, E. Toreini, M. Mehrnejad and S. K. Shekofteh, “Analysis of The Usage of Chaotic Theory in Data Clustering using Particle Swarm Optimization”, Indian J. Sci. Res., Vol. 4, No. 3, pp- 335-353, 2014.
79. S. Rana, S. Jasola, and R. Kumar, “A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm”, International Journal of Engineering, Science and Technology, Vol. 2, No. 6, pp-167-176, 2010.
80. S. Bandyopadhyay, U. Maulik, and M. K. Pakhira, “Clustering Using Simulated Annealing With Probabilistic Redistribution” International Journal of Pattern Recognition and Artificial Intelligence, Vol.15, No. 2, pp. 269-285, 2001.
81. S. K. Majhi, and S. Biswal, “Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer”, Karbala International Journal of Modern Science, pp-347-360, 2018.
82. S. Gajawada and D. Toshniwal, “SPPS: Supervised Projected Clustering Method Based on Particle Swarm Optimization”, International Journal of Machine Learning and Computing, Vol. 2, No. 3, pp-1-6, 2012.
83. S. Madhusudhanan, S. Jganathan, and J. L. S, “Incremental Learning for Classification of Unstructured Data Using Extreme Learning Machine”, Algorithms, MDPI, pp. 1-19, 2018.
84. S. A. Fahad and M. M. Alam, “A Modified K-Means Algorithm for Big Data Clustering”, IJCSET, Vol. 6, Issue 4, pp. 129-132, 2016.

85. S. Karol, and V. Mangat, “Evaluation of text document clustering approach based on particle swarm optimization”, Cent. Eur. J. Comp. Sci., Vol. 3, No. 2, pp-69-90, 2013.
86. S. V, V. V, L. R and I. V, “Unstructured Data Analysis on Big Data using Map Reduce”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC), Procedia Computer Science, Elsevier, Vol. 50, pp. 456-465, 2015.
87. S. Jaganathan, D. V. Prasad, and S. Madhusudanan, “uCLUST-a new algorithm for clustering unstructured data”, ARPN Journal of Engineering and Applied Sciences, Vol. 10, No. 5, pp-1-11, 2015.
88. S. Venkatraman, “A Proposed Business Intelligent Framework for Recommender Systems”, Informatics, MDPI, 4 (40), pp. 1-12, 2017.
89. S. Ullah, M. D. Awan and M. S. H. Khiyal, “Big Data in Cloud Computing: A Resource Management Perspective”, Scientific Programming, Hindawi, pp. 1-18, 2018.
90. S. Z. Bin, J. Y. Xia; “Research on semantics of concept hierarchy based on Formal Concept Analysis”; Proc. of the 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD); pp.964-967, 2012.
91. T. Singh and J. S. Prasad, “Efficient of Unstructured Data: A defiance Perspective”, International Journal of Pure and Applied Mathematics, Vol. 117, No. 17, pp. 199-205, 2017.
92. T. Samal, A. Baral and H. S. Behera, “High Dimensional Data Clustering Using Hybridized Teaching-Learning-Based Optimization”, J. Comp. & Math. Sci., Vol.4, No. 3, pp-167-177, 2013.
93. U. S. Patki and P. G. Khot, “A Literature Review on Text Document Clustering Algorithms used in Text Mining”, Journal of Engineering Computers & Applied Sciences (JECAS), Vol. 6, No. 10, pp. 16-20, 2017.
94. V. Shanu, and S. Vydehi, “Optimal and Fast Health Data Clustering Using Hybrid Meta Heuristic Algorithm”, International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE), Vol. 5, Issue 7, pp-13339-13347, 2017. (DOI: 10.15680/IJIRCCE.2017. 0507069)
95. W. M. S. Yafooz, “Model of Textual Data Linking and Clustering in Relational Databases”, Research Journal of Information Technology (RJIT), pp-1-12, 2016.

96. W. L. Chang, J. Kanesan, A. J. Kulkarni, and H. Ramiah, “Data clustering using seed disperser ant algorithm”, Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 25, pp-4522 -4532, 2017.
97. W. Q. Qwaider, “Apply On-Line Analytical Processing (OLAP) With Data Mining For Clinical Decision Support”, International Journal of Managing Information Technology (IJMIT), 4 (1), pp. 1-13, 2012.
98. W. Fuertes, F. Reyes, P. Valladares, F. Tapia, T. Toulkeridis, and E. Perez, “An Integral Model to Provide Reactive and Proactive Services in an Academic CSIRT Based on Business Intelligence”, Systems, MDPI, 5 (52), pp. 1-20, 2017.
99. Z. Dong, H. Jia, and M. Liu, “An Adaptive Multiobjective Genetic Algorithm with Fuzzy K-Means for Automatic Data Clustering”, Hindawi Mathematical Problems in Engineering, pp-1-13, 2108. (https://doi.org/10.1155/2018/6123874)
100. Z. Chuan, “Clustering and Entity Resolution for Semi Structured Data”, Washington State University, pp. 1-208, 2011.
101. Z. Zainol, A. M. Azahari, S. Wani, S. Marzukhi, P. N. E. Nohuddin and O. Zakaria, “Visualizing Military Explicit Knowledge using Document Clustering Techniques”, International Journal of Academic Research in Business & Social Sciences, pp. 1-17, 2018.
102. Z. Marx, I. Dagan, J. M. Buhmann and E. Shamir, “Coupled Clustering: a Method for Detecting Structural Correspondence”, pp. 1-29, 2002.